DATA ANALYTICS 2015

The Fourth International Conference on Data Analytics

ISBN: 978-1-61208-423-7

July 19 - 24, 2015

Nice, France

DATA ANALYTICS 2015 Editors

Thomas Klemas, AIRS, Swansea University, USA

Steve Chan, Swansea University & Hawaii Pacific University, USA

DATA ANALYTICS 2015

Forward

The Fourth International Conference on Data Analytics (DATA ANALYTICS 2015), held between July 19-24, 2015 in Nice, France, continued a series of events on fundamentals in supporting data analytics, special mechanisms and features of applying principles of data analytics, application-oriented analytics, and target-area analytics.

Processing of terabytes to petabytes of data, or incorporating non-structural data and multi-structured data sources and types require advanced analytics and data science mechanisms for both raw and partially-processed information. Despite considerable advancements on high performance, large storage, and high computation power, there are challenges in identifying, clustering, classifying, and interpreting of a large spectrum of information.

The conference had the following tracks:

- Resiliency and Sustainability through Analytics
- Application-oriented analytics
- Mechanisms and Features
- Target analytics
- Big Data

Similar to previous editions, this event attracted excellent contributions and active participation from all over the world. We were very pleased to receive top quality contributions.

We take here the opportunity to warmly thank all the members of the DATA ANALYTICS 2015 technical program committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and effort to contribute to DATA ANALYTICS 2015. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the DATA ANALYTICS 2015 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that DATA ANALYTICS 2015 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of data analytics. We also hope that Nice, France, provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

DATA ANALYTICS 2015 Chairs

DATA ANALYTICS Advisory Chairs

Fritz Laux, Reutlingen University, Germany
Lina Yao, The University of Adelaide, Australia
Eiko Yoneki, University of Cambridge, UK
Takuya Yoshihiro, Wakayama University, Japan
Felix Heine, University of Applied Sciences & Arts Hannover, Germany
Dominik Slezak, University of Warsaw & Infobright Inc., Poland
Panos M. Pardalos, University of Florida, USA
Michele Melchiori, Università degli Studi di Brescia, Italy
Rudolf Berrendorf, Bonn-Rhein-Sieg University, Germany
Sandjai Bhulai, VU University Amsterdam, The Netherlands
Kerstin Lemke-Rust, Hochschule Bonn-Rhein-Sieg, Germany
Sergio Ilarri, University of Zaragoza, Spain
Les Sztandera, Philadelphia University, USA
Prabhat Mahanti, University of New Brunswick, Canada
Dominique Laurent, University of Cergy Pontoise, France
Ryan G. Benton, University of Louisiana at Lafayette, USA
Erik Buchmann, Karlsruhe Institute of Technology, Germany
Andrew Rau-Chaplin, Dalhousie University, Canada
Takuya Yoshihiro, Wakayama University, Japan

DATA ANALYTICS Industry/Research Liaison Chairs

Qiming Chen, HP Labs - Palo Alto, USA
Alain Crolotte, Teradata Corporation - El Segundo, USA
Tiziana Guzzo, National Research Council/Institute of Research on Population and Social Policies - Rome, Italy
Shlomo Geva, Queensland University of Technology - Brisbane, Australia
Farhana Kabir, Intel, USA
Serge Mankovski, CA Technologies, Spain
Sumit Negi, IBM Research, India
Vedran Sabol, Know-Center - Graz, Austria
George Tambouratzis, Institute for Language and Speech Processing - Athena Research Centre, Greece
Yanchang Zhao, RDataMining.com, Australia
Masahiro Terabe, Mitsubishi Research Institute, Inc., Japan
Marina Santini, Santa Anna IT Research Institute AB, Sweden
Mario Zechner, Know-Center, Austria

DATA ANALYTICS Publicity Chairs

Johannes Leveling, Dublin City University, Ireland
Tim Weninger, University of Illinois in Urbana-Champaign, USA
Roberto Zicari, Johann Wolfgang Goethe - University of Frankfurt, Germany
Shandian Zhe, Purdue University, USA
Michael Schaidnagel, Reutlingen University, Germany

DATA ANALYTICS Special Area Chairs

Resiliency and Sustainability Through Analytics
Thomas Klemas, AIRS, Swansea University, USA
Steve Chan, Swansea University & Hawaii Pacific University, USA

DATA ANALYTICS 2015

Committee

DATA ANALYTICS Advisory Chairs

Fritz Laux, Reutlingen University, Germany
Lina Yao, The University of Adelaide, Australia
Eiko Yoneki, University of Cambridge, UK
Takuya Yoshihiro, Wakayama University, Japan
Felix Heine, University of Applied Sciences & Arts Hannover, Germany
Dominik Slezak, University of Warsaw & Infobright Inc., Poland
Panos M. Pardalos, University of Florida, USA
Michele Melchiori, Università degli Studi di Brescia, Italy
Rudolf Berrendorf, Bonn-Rhein-Sieg University, Germany
Sandjai Bhulai, VU University Amsterdam, The Netherlands
Kerstin Lemke-Rust, Hochschule Bonn-Rhein-Sieg, Germany
Sergio Ilarri, University of Zaragoza, Spain
Les Sztandera, Philadelphia University, USA
Prabhat Mahanti, University of New Brunswick, Canada
Dominique Laurent, University of Cergy Pontoise, France
Ryan G. Benton, University of Louisiana at Lafayette, USA
Erik Buchmann, Karlsruhe Institute of Technology, Germany
Andrew Rau-Chaplin, Dalhousie University, Canada
Takuya Yoshihiro, Wakayama University, Japan

DATA ANALYTICS Industry/Research Liaison Chairs

Qiming Chen, HP Labs - Palo Alto, USA
Alain Crolotte, Teradata Corporation - El Segundo, USA
Tiziana Guzzo, National Research Council/Institute of Research on Population and Social Policies - Rome, Italy
Shlomo Geva, Queensland University of Technology - Brisbane, Australia
Farhana Kabir, Intel, USA
Serge Mankovski, CA Technologies, Spain
Sumit Negi, IBM Research, India
Vedran Sabol, Know-Center - Graz, Austria
George Tambouratzis, Institute for Language and Speech Processing - Athena Research Centre, Greece
Yanchang Zhao, RDataMining.com, Australia
Masahiro Terabe, Mitsubishi Research Institute, Inc., Japan

Marina Santini, Santa Anna IT Research Institute AB, Sweden
Mario Zechner, Know-Center, Austria

DATA ANALYTICS Publicity Chairs

Johannes Leveling, Dublin City University, Ireland
Tim Weninger, University of Illinois in Urbana-Champaign, USA
Roberto Zicari, Johann Wolfgang Goethe - University of Frankfurt, Germany
Shandian Zhe, Purdue University, USA
Michael Schaidnagel, Reutlingen University, Germany

DATA ANALYTICS Special Area Chairs

Resiliency and Sustainability Through Analytics

Thomas Klemas, AIRS, Swansea University, USA
Steve Chan, Swansea University & Hawaii Pacific University, USA

DATA ANALYTICS 2015 Technical Program Committee

Mohd Helmy Abd Wahab, Universiti Tun Hussein Onn Malaysia, Malaysia
Sayed Abdel-Wahab, Sadat Academy for Management Sciences, Egypt
Rajeev Agrawal, North Carolina A&T State University - Greensboro, USA
Pranay Anchuri, Rensselaer Polytechnic Institute, USA
Fabrizio Angiulli, University of Calabria, Italy
Annalisa Appice, Università degli Studi di Bari Aldo Moro, Italy
Giuliano Armano, University of Cagliari, Italy
Ryan G. Benton, University of Louisiana at Lafayette, USA
Sandjai Bhulai, VU University Amsterdam, The Netherlands
Erik Buchmann, Karlsruhe Institute of Technology, Germany
Luca Cagliero, Politecnico di Torino, Italy
Huiping Cao, New Mexico State University, USA
Omar Andres Carmona Cortes, Federal Institute of Maranhao (IFMA), Brazil
Michelangelo Ceci, University of Bari, Italy
Federica Cena, Università degli Studi di Torino, Italy
Steve Chan, Swansea University & Hawaii Pacific University, USA
Lijun Chang, University of New South Wales, Australia
Qiming Chen, HP Labs - Palo Alto, USA
Been-Chian Chien, National University of Tainan, Taiwan
Silvia Chiusano, Politecnico di Torino, Italy
Alain Crolotte, Teradata Corporation - El Segundo, USA
Tran Khanh Dang, National University of Ho Chi Minh City, Vietnam
Mrinal Kanti Das, Aalto University, Finland
Ernesto William De Luca, University of Applied Sciences Potsdam, Gernmany

Shicong Meng, Georgia Institute of Technology, USA
George Michailidis, University of Michigan, USA
Victor Muntés Mulero, CA Technologies, Spain
Sumit Negi, IBM Research, India
Feiping Nie, University of Texas at Arlington, USA
Oliver Niggemann, Institut für Industrielle Informationstechnik, Germany
Sadegh Nobari, Singapore Management University, Singapore
Panos M. Pardalos, University of Florida, USA
Dhaval Patel, Indian Institute of Technology-Roorkee, India
Jan Platoš, VSB-Technical University of Ostrava, Czech Republic
Ivan Radev, South Carolina State University, USA
Zbigniew W. Ras, University of North Carolina - Charlotte, USA & Warsaw University of Technology, Poland
Jan Rauch, University of Economics - Prague, Czech Republic
Yenumula B. Reddy, Grambling State University, USA
Manjeet Rege, University of St. Thomas, USA
Vedran Sabol, Know-Center - Graz, Austria
Abdel-Badeeh M. Salem, Ain Shams University Abbasia, Egypt
Marina Santini, SICS East Swedish ICT AB, Sweden
Ivana Šemanjski, University of Zagreb, Croatia / University of Gent, Belgium
Hayri Sever, Hacettepe University, Turkey
Micheal Sheng, Adelaide University, Australia
Shuichi Shinmura, Seikei University, Japan
Fabrício A.B. Silva, FIOCRUZ, Brazil
Josep Silva Galiana, Universidad Politécnica de Valencia, Spain
Dan Simovici, University of Massachusetts - Boston, USA
Dominik Slezak, University of Warsaw & Infobright Inc., Poland
Paolo Soda, Università Campus Bio-Medico di Roma, Italy
Qinbao Song, Xi'an Jiaotong University, China
Theodora Souliou, National Technical University of Athens, Greece
Srivathsan Srinivas, Cognizant, USA
Vadim Strijov, Computing Center of the Russian Academy of Sciences, Russia
Les Sztandera, Philadelphia University, USA
George Tambouratzis, Institute for Language and Speech Processing - Athena Research Centre, Greece
Tatiana Tambouratzis, University of Piraeus, Greece
Mingjie Tang, Purdue University, U.S.A.
Maguelonne Teisseire, Irstea - UMR TETIS (Earth Observation and Geoinformation for Environment and Land Management research Unit) - Montpellier, France
Masahiro Terabe, Mitsubishi Research Institute, Inc., Japan
Ankur Teredesai, University of Washington - Tacoma, USA
A. Min Tjoa, TU-Vienna, Austria
Li-Shiang Tsay, North Carolina A & T State University, U.S.A.
Chrisa Tsinaraki, EU Joint Research Center - Ispra, Italy

Xabier Ugarte-Pedrero, Universidad de Deusto - Bilbao, Spain
Michael Vassilakopoulos, University of Thessaly, Greece
Maria Velez-Rojas, CA Technologies, Spain
Zeev Volkovich, ORT Braude College Karmiel, Israel
Stefanos Vrochidis, Information Technologies Institute - Centre for Research and Technology Hellas, Greece
Jason Wang, New Jersey Institute of Technology, U.S.A.
Guan Wang, LinkedIn Corporation, USA
Leon S.L. Wang, National University of Kaohsiung, Taiwan
Tim Weninger, University of Illinois in Urbana-Champaign, USA
Wolfram Wöß, Johannes Kepler University Linz - Institute for Application Oriented Knowledge Processing, Austria
Yuehua Wu, York University, Canada
Guandong Xu, Victoria University - Melbourne, Australia
Divakar Yadav, Jaype Institute of Information Technology, Noida, India
Divakar Singh Yadav, South Asian University - New Delhi, India
Lina Yao, The University of Adelaide, Australia
Eiko Yoneki, University of Cambridge, UK
Takuya Yoshihiro, Wakayama University, Japan
Aidong Zhang, State University of New York at Buffalo, USA
Xiaoming Zhang, Beihang University, China
Yanchang Zhao, RDataMining.com, Australia
Yichuan Zhao, Georgia State University, USA
Shandian Zhe, Purdue University, USA
Roberto Zicari, Johann Wolfgang Goethe - University of Frankfurt, Germany
Albert Zomaya, The University of Sydney, Australia

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Technology Roadmap for Hawaii Resiliency

Resiliency and Sustainability through Advanced Analytics

Thomas J. Klemas and Steve Chan
Sensemaking Fellowship, Swansea University Network Science Research Center
Swansea, Wales
email: tklemas@alum.mit.edu, stevechan@post.harvard.edu

Abstract—**The state of Hawaii faces numerous challenges that threaten its survival. Tsunamis, weather-induced mud-slides, and global climate change impacts are just a few of the threats that could cripple Hawaii. The state is increasingly vulnerable because of aging infrastructure and the fact that its economy is highly dependent on tourism and construction, military, or government projects. To improve Hawaii's resilience, this effort proposes steps that will enable analytics based decision support to combat the major threats. In this paper, we describe a technology roadmap that will guide Hawaii to pioneer a sound resiliency approach, facilitate implementation of a resiliency plan, develop the required resiliency technologies, deploy the resiliency technology and improve Hawaii's infrastructure, and foster growth of a technology-based resiliency industry that will sustain Hawaii's resilience. As a successful pathfinder for resiliency, Hawaii will be positioned to lead the way for many other cities, states, nations, and even regions of the world that face similar threats.** ***Keywords—resiliency; sustainability; technology roadmap; analytics; decision support; high performance computing; hyper-local weather forecasting.***

I. INTRODUCTION

In industry, the concept of formulating a technology roadmap in order to envision the future, articulating a desired end-state, and developing a plan of action to reach that end-state is valuable to align the key players and motivate them to mobilize their resources towards a common goal. The technology roadmap provides a valuable mechanism for communication between all parties involved and enables key decisions, such as standards, and other required preparations, to take place in advance of arriving at the end-state goals. Since none of the key players in this Hawaiian resiliency initiative possess sufficient resources to execute the plan individually, a Public Private Partnership Initiative (P3I) is the best mechanism to achieve this goal. In this paper, we will show how we adapted the technology roadmap concept, borrowed from industry, formed a public private partnership aimed to achieve Hawaiian resiliency, and developed the roadmap details.

The central element of the technology roadmap is the vision or desired end-state. As hinted in the previous paragraph, the end-state for this roadmap targets improvements in Hawaii resiliency. However, the future envisioned for Hawaii is even greater, proposing that Federal, State, and local authorities team with academic and industrial partners to construct and implement a plan that will position Hawaii to be a pathfinder and world leader in developing resiliency technologies. The motivation for this ambitious end-state is to counter serious threats that pose grave danger to Hawaii's very survival.

Hawaii faces several major, imminent threats. The first threat is an impending economic downturn that will be compounded by dependence upon a tourism-heavy economy [1]. The second threat is the weather and environment [9] [10] [11]. These threats are deeply compounded by Hawaii's isolated location and aging infrastructure. The third threat is the emerging cyber-threat that is shared by the entire globe and could cripple Hawaii's economy even if steps are taken to avert the predicted economic downturn and could be wielded by an adversarial actor(s) to intentionally attack elements of Hawaii's economy or difficult-to-defend critical/strategic infrastructure, or to hamper Hawaii's emergency response mechanisms.

It is important to note that these threats are linked. Achieving resiliency requires a stable economy, and a stable economy requires resiliency. Conversely, the coupling magnifies the potential impact of any of these threats and poses severe challenges to Hawaii's resiliency. In a sense, the combined threats facing Hawaii represent the “perfect storm” that is looming in the not-too-distant future, darkening the horizon. To make matters worse, there seems to be universal agreement that Hawaii is under-prepared to meet these challenges should they should solidify and that should any of the threats materialize, it would exact a terrible toll.

To fully understand the Hawaiian resiliency vision and technology roadmap end-state, it is necessary to describe each major threat in detail. Section II will endeavor to do so. Section III will present the technical solutions required to counter the threats facing Hawaii and repair its vulnerabilities. This section will present a technology roadmap that implements the required solutions as part of a comprehensive approach to develop a new resiliency industry that will sustain and extend Hawaii into the future. Section IV will present the status and progress towards achieving Hawaiian resilience, Section V will describe the technology roadmap, and Section VI will detail future steps that remain to be executed. Finally, the summary section will highlight the main points of the

 ISBN: 978-1-61208-423-7

document and present reasons why the successful implementation of this technology roadmap is of strategic importance to the U.S. and the rest of the world.

II. DETAILED CHALLENGES

A. *Environmental Challenges*

The Hawaiian islands are the most isolated chain of islands in the world, located in the middle of the Pacific Ocean and quite exposed to the impacts of mother nature. As such, Hawaii faces tremendous challenges due to hostile weather and environmental phenomena like hurricanes, tsunamis, storms and related flooding, volcanic eruptions and lava flows, mud slides, global climate change, and more.

The danger posed by a tsunami is well-known and widely feared. On March 11, 2011, the Tohoku earthquake triggered a tsunami that inundated the Fukushima Nuclear Power Plant on the coast of Japan and resulted in the Fukushima Daiichi nuclear disaster when 3 of the 6 nuclear reactors melted down. While most of the direct damage of a major tsunami cannot be averted, certain measures to prepare can indeed reduce some of the secondary impacts, if there is sufficient warning. Often a major storm or tsunami will take down the power grid. In fact, the power grid can be shut down intentionally in anticipation of an impending tsunami. Pumping stations for water distribution and waste water sewage can be prepared with extra fuel for backup generators to allow water pumping to continue when it is most necessary, particularly with the proviso that there is sufficient warning to do so. With regards to the tsunami warning buoys placed in the Pacific to warn Hawaii of impending tsunami, the National Research Council of the National Academies have called for a replacement strategy. Furthermore, the antiquated technology on the buoys are more of a liability to Hawaii than as the intended functionality of a safety mechanism, for the buoys have no robust protection mechanism to defend against hackers attempting buoy spoofing as well as other adverse actors that could trigger a warning and induce decision-makers to shut down the power grid in Hawaii.

Hawaii struggles to produce sufficient, affordable, and stable power. Outages occur due to demand combined with infrastructural issues, vegetation overgrowth, and other phenomenon affect the islands. Such interruptions incur economic penalties, but also can have deadly consequences when emergency equipment ceases to run, pumping stations fail, and similar shutdowns occur. Not only from the result of severe weather events, on a daily level, Hawaii faces a continuous occurrence of adverse micro-weather affects that may have causal impacts. While Hawaii has enthusiastically adopted solar technologies to leverage its tropical sunshine to produce energy, unanticipated micro-changes in cloud cover can profoundly affect matters. As a result, the switching points between solar and fossil-fuel based electricity production can indeed be better optimized so as to cope with these micro-weather changes.

As a longer-term threat, global climate change threatens to elevate sea level and the consequences could be dire for Hawaii. In Hawaii, a rising sea level could lead to increased soil salinity levels in coastal areas along the perimeter of all the islands. This change in soil chemistry could potentially force insects, such as termites to higher grounds, thereby impacting telephone poles and similar infrastructures that were previously not as vulnerable.

B. *Economic Challenges*

Economic challenges are not usually direct threats to existence, but in the case of Hawaii, economic issues elevate Hawaii's vulnerability and compound the impact of the environmental threats. Upgrades to aging infrastructure so as to counter the threats require a significant amount of funds. Sustainment of the upgraded infrastructure will require long-term economic stability. Along this vein, longer-term economic stability depends upon a broad-based economy supported by more than just tourism and the construction, military, and government projects that have traditionally boosted Hawaii's economy. Thus, in order to achieve sustained resiliency, Hawaii needs to foster new technology-based drivers for its economic growth that will allow it to compete favorably in the increasingly global marketplace even when its traditional sources of funds are less available.

Due to its geographical location, Hawaii is the most isolated island state in the world. As such, it has always faced steep economic challenges. For one, any enterprise in Hawaii, whether individual, commercial, or government, raw materials and energy cost significantly more. Solar energy may offset some of the additional cost of energy, but as of yet (despite its enormous potential), it does not make up the difference. Electricity can cost 3-4 times the price it does on the mainland. Furthermore, climate change and global warming impacts will place more drag on the economy by forcing Hawaii to respond to the changing conditions to well protect its infrastructure. By depending upon tourism and construction, military, and government projects to drive the economy for many years, Hawaii has not kept pace with the increasingly global and competitive marketplace. Thus, Hawaii's economy is highly vulnerable and dependent on numerous factors beyond its control.

Many of the aforementioned disadvantages described above already disfavor industry investment in Hawaii and lead new companies to take root elsewhere than in Hawaii. However, there is yet another far more significant factor that threatens to drive away future business opportunities. Hawaii lags in certain critical/strategic infrastructure to support the burgeoning Internet bandwidth need that is crucial to the modern global marketplace, and its current trans-oceanic cables faces physical degradation by traditional end-of-life factors. Streaming video can be problematic because bandwidth is so limited, which means that tele-meeting technologies, such as video teleconferencing (VTC), educational opportunities for remote learning, and many other high-bandwidth data streaming applications may be impacted.

Model predictions indicate that Hawaii will face an extended economic downturn in less than 5 years, if no measures are taken. In order to minimize impacts of the downturn and, conversely, actually drive the economy, it is imperative to incorporate new growth factors that will broaden Hawaii's economy beyond its current pillars of tourism, construction, military, and government. In particular, to compete in the increasingly competitive and global market place, it is vital that Hawaii rejuvenate its economic engine by both attracting and starting technology innovation. One means

 ISBN: 978-1-61208-423-7

to accomplish this crucial revitalization of the economy is for Hawaii to boldly address its resiliency problems by developing innovative and comprehensive solutions, fostering a novel resiliency industry, and committing education and training resources to train others. In doing so, Hawaii would become a world leader in resiliency, and serve as a resource from which others can learn. Due to its unique property as being the most isolated island chain in the world, Hawaii is, in essence, a bounded problem set and is an ideal "living lab" for industry.

III. APPROACH TO OVERCOME CHALLENGES

Hawaii does have several resources and advantages that it can leverage to overcome the challenges threatening its future. For one, Hawaii is a relatively closed system that makes it easier to attack the resiliency problem from a comprehensive perspective. Second of all, Hawaii is still on solid economic footing, so there are economic resources to utilize in improving its resilience posture. Additionally, Maui still is the home of the Maui High Performance Computing Center (MHPCC), which greatly enhances Hawaii's ability to build up the sophisticated analytic-based decision support aids that are required to implement a smart-grid, smart-buoy defense system, and —ultimately — smart-cities.

To counter the dangers outlined in the previous section, a number of inter-related improvement efforts must be initiated. First, the tsunami-warning buoy system must be upgraded. According to the 2010 Report "An Assessment of the U.S. Tsunami Program and the Nation's Preparedness Efforts," by the National Research Council of the National Academies, the current system, known as the Deep-ocean Assessment and Reporting of Tsunamis (DART), which services the U.S. (and Hawaii) as well as 50 other countries, is unreliable. For example, "of the 39 stations deployed in 2008 only an estimated 60 percent were operational by 2009." New sensors and security must be incorporated into the buoys to truly improve the tsunami defense, improve Hawaii's defensive posture, and reduce security risks. These buoys sensors will collect data that will have to be transmitted to a central location for analysis and decision support. One proposed solution is to utilize Unmanned Aerial Vehicle (UAV) communications to create a network that ultimately connects the buoys to the MHPCC, perhaps in conjunction with satellites. Second, the broadband initiative is critical to replace existing broadband before it reaches saturation and/or physically degrades. Interruption in Internet capabilities would expose Hawaii's security and economy to grave risks. Furthermore, due to the time required for such an upgrade, the time to act on this upgrade is quite limited. Third, Hawaii's grid and other infrastructure must be instrumented appropriately to enable smart-grid and smart-city capabilities. Finally, many of the new sensors must be connected, ultimately, to the Maui High Performance Computing Center, so as to enable analytic engines to derive insights from the raw data, and edge analytic systems must be deployed in those situations where such connection is not feasible.

To process the live data streams, the MHPCC will be equipped with advanced software systems that comprise a crucial component of a modeling and analytics infrastructure that is aiming to achieve lasting resilience for the State of Hawaii. These computational capabilities will be both enabled and sustained by increased bandwidth resulting from new data channels provided by the Hawaii Broadband Initiative. To support resilience, this data analytics architecture will support connections to a variety of data streams, including data from sensors from a future smart-grid utility infrastructure, the future replacement sensor buoys for the aging storm-warning/tsunami-detection buoys, processing satellite imagery from the existing Earth Observation System, sensor data from existing and future weather balloons, and potentially many other sources.

To enhance Hawaiian resilience, MHPCC high performance computing systems will host software algorithms that are designed to accept real-time data streams, analyze the data within a suitable historical context by leveraging available meta data, automatically detect patterns from which insights may be derived, infer relationships from interconnections between data elements, and provide advanced decision-making tools that will help local, state, and federal leaders to protect Hawaii's electrical grid, provide hyper-local weather prediction in addition to alarm for storms, tsunamis, mud slides, and other adverse environmental events. These same analytic tools can, ultimately, be harnessed to empower a growing technology based economy.

Initially, the most critical of the myriad of challenges to tackle seems to revolve around the weather challenges. This sub-challenge benefits from the fact that the technical approach, computing approach, and sensors already exist. The IBM Deep Thunder system offers technology that has been developed to provide local, high-resolution weather predictions customized to weather-sensitive specific operations. For example, it could be used to predict situations ranging from flooding and/or damaged power lines to anticipating cloud cover over the Hawaiian island of Maui.

IV. STATUS AND PROGRESS OF RESILIENCE INITIATIVE

There has been significant progress related to the resiliency initiative. As a Public Private Partnership Initiative (P3I), the resiliency initiative is a combined effort among the State of Hawaii, U.S. Pacific Command (PACOM), Swansea University's Network Science Research Center, IBM Smarter Cities and Safer Planet, Mehta Tech, Synerscope, and others. On July 7, 2014, State of Hawaii Senate Bill 2742 (Act 229) — co-sponsored by then Senator David Ige (now Governor for the State of Hawaii) — was signed into law by then Governor Neil Abercrombie. The 28th Legislature of the State of Hawaii is currently working on legislation related to the Pacific-Asia Institute for Resiliency and Sustainability (AIRS) mission, including explorations for a new trans-oceanic broadband connection. In addition, PACOM co-sponsored a Resiliency Symposium in Honolulu, Hawaii, which featured presentations by the Sensemaking-PACOM Fellowship.

V. TECHNOLOGY ROADMAP

Outlining the process to develop a technology roadmap is itself an important element of achieving success. In this

 ISBN: 978-1-61208-423-7

paper, we adapt the steps defined in [3] from the industry context to the Hawaii Resiliency problem space. First, it is important to ensure that developing a technology roadmap will actually yield benefits to the parties involved. After all, we have already identified strong candidate solutions to the problems that threaten Hawaii's future. Do we really need a resiliency technology roadmap? In this case, the motivation for developing a technology roadmap is two-fold. From a resiliency perspective, while potential solutions have been identified to many of the threats described earlier, some of the technologies have not been fully developed to achieve the sustained resiliency that Hawaii aims to achieve. For example, cyber-technology seems to lack the required level of maturity. Therefore, to achieve Hawaii's resiliency vision, it will be essential to develop a new resiliency technology base. As such, the technology roadmap will be of great value to outline the steps to achieve this. Additionally, the logic is similar from a sustainment point of view that in order to maintain the fledgling resiliency technology industry, which will provide the necessary latent stability, it will actually be advantageous to grow much of the resiliency technology locally. This will effectuate a sustained econometric framework that will serve as a sustained driver to Hawaii's economy and will help insulate Hawaii's isolated islands from the uncertainties inherent in the historical principal economic dependence on tourism, construction, government, and military projects.

Once the benefit of a technology roadmap is established, it is important to select champions that have the know-how, resources, and leadership to bring about the benefits envisioned by the roadmap. In order to achieve Hawaiian Resiliency, the optimal choice is a Public-Private Partnership Initiative, involving the state of Hawaii, local leaders, PACOM, MHPCC, AIRS, the Sensemaking Fellowship, several academic institutions, and several industry partners. None of these parties alone have the resources to effectuate the desired outcome, but together in partnership, many of the foundational pieces that will be required by the resiliency technology roadmap are already being put in place. To enact this resiliency vision for Hawaii, critical resource elements are required, which include political capital, funding, computational resources, subject matter expertise, academic support, and a strong network of relationships with potential candidate technology partners, some of which might help realize the vision. Many of these elements are provided by the P3I members, some of whom are referenced in the February 7, 2014 and October 31, 2014 Hawaii Department of Defense press releases regarding their participation with PACOM and others on methods to improve energy efficiency and grid operations. The process that will be used to develop the technology roadmap has already started. This may seem counter-intuitive, but the vision of the technology roadmap described above represents a sustained resilience approach that is clearly a super set of the solutions developed to mitigate the immediate threats posed by an aging tsunami-warning system, aging infrastructure, looming global climate change effects, and broadband end-of-life. Thus, essentially, the resilience initiative has evolved to achieve a sustainable solution that involves growth of an entire industry.

The process has included a hybrid expert and local workforce-based approach. PACOM co-hosted a Hawaii resilience symposium organized in a workshop setting that highlighted a fairly comprehensive resilience approach. The Senate Majority Leader discussed legislative progress related to various elements of the resiliency initiative. Subject matter experts from the Sensemaking Fellowship and AIRS discussed aspects of the resilience approach with audience members which included officials, legislators, and department members from the state of Hawaii, local officials, representatives of the MHPCC, academic leaders, and other interested parties. Subsequently, the Sensemaking Fellows met individually with representatives of many of these groups to explain technical components, highlight requirements of the overall approach, and solidify the crucial elements of the technology roadmap.

As discussed before, this technology roadmap has been developed to achieve sustained resilience, evolving far beyond simple mitigation of several of the immediate threats, so as to help position Hawaii as a leader at the forefront of resilience technologies and to demonstrate this leadership by example, such that the rest of the world will look to Hawaii for resilience solutions. The components of the roadmap were adapted from [3] and [4] to fit within the P3I approach. The technical roadmap in [3] included 5 major components: “Goals”, “Milestones”, “Gaps and Barriers”, “Action Items”, and “Priorities and Timelines.” The elements of the technology roadmap presented in [4] were greater in number and more specific and focused on long-term benefits from primarily an industry point of view. The technology roadmap presented here differs from [3] and [4] because it has both a primary goal of solving the Hawaiian resilience challenges and a secondary goal of positioning Hawaii — for sustained resilience — as the leader in resilience by fostering a local technology industry and building local training programs, both academic and otherwise, related to resilience technologies.

The first section of the technology roadmap is an analysis of the specific challenges facing Hawaii, a summary of existing technologies available to counter those challenges, aspects of the threats for which no solution exists that will require research and innovation to overcome, and an evaluation of resilience technologies and industry trends. The threats facing Hawaii were discussed in section II and solutions to overcome those challenges were presented in section III. These aforementioned threats do not fully describe or address the hurdles that exist for Hawaii to attract technology companies; there are fairly generous tax credits for companies to locate in Hawaii, but there are severe disadvantages as well. Businesses seeking to locate within Hawaii face difficulties ranging from its geographical remoteness to its relatively high energy costs (more expensive than the mainland for electricity) high cost of living, high cost for businesses, challenge of recruiting, and a limited venue for

 ISBN: 978-1-61208-423-7

venture capital. In light of these obstacles for an individual company, the goal of growing an entire new technology industry in Hawaii represents a steep hill to climb.

In the next step, it is important to determine resilience technology and expertise focus areas in which Hawaii already has an edge or that a potential budding Hawaiian resilience industry could excel and also directly support Hawaii's resilience needs.

In this case, the Hawaii island of Maui already hosts one of 5 DoD super computing sites, the Maui High Performance Computing Center (MHPCC). The MHPCC supercomputers already host numerous tools for computational modeling, many of which can be of great value for design of new systems, enabling rapid prototyping and significant savings. New network science and analytic tools can be added [6] atop hyper-local forecasting capabilities, such as Deep Thunder. Deep Thunder is a research project by IBM, being offered in this instance by the IBM Center for Resiliency and Sustainability, which aims to improve short-term local weather forecasting through the use of high-performance computing. It is part of IBM's Deep Computing initiative that also produced the Deep Blue chess computer. Deep Thunder is intended to provide local, high-resolution weather predictions customized to weather-sensitive specific operations. For example, it could be used to predict situations ranging from the wind velocity at an Olympic diving platform to flooding and/or damaged power lines. Additionally, the Sensemaking Fellowship provides expertise in numerous science and technical areas that will directly support these efforts. For example, members of the Sensemaking Fellowship already have experience running complex computational electromagnetic modeling simulations [5] on high performance computer systems similar to the MHPCC, conducting information theory and network science [2] research highly relevant to the analytic systems that will be required for the resiliency initiative, and currently study many of the topics that are directly relevant to the resilience initiative. Another participant, Hawaii Pacific University intends to start a Resilience Masters program for future Sensemaking Fellows in conjunction with Swansea University. These kinds of shifts resulting from initiating a resiliency industry could stimulate significant growth and improvement in the quality of the overall high technology workforce and related educational institutions.

The threats facing Hawaii and solutions to those threats were described earlier, but it is also important to describe the principal obstacles that will hamper both the Resilience initiative and a potential new resilience technology industry in Hawaii. Primary of these, are inertia effects, such as political inertia, the fact that currently Hawaii has no technology industry to speak of, disadvantages posed by the high cost of materials, energy and operating budget, a lack of top quality technical talent currently existing on the islands, and high costs of the its isolation. For those that live in Hawaii, paying significantly more for gas, milk, bread, housing, and many other daily needs are only part of the penalty for living in paradise. The aforementioned financial penalty is compounded by other related impacts of Hawaii's isolation. Technology companies are reluctant to translocate themselves to the isolated environment, which has a ripple effect. The lack of technology industry translates to a technical workforce with fewer graduate degrees in technical topics. The result of this shortage equates to a decreased number of highly skilled engineers, scientists, and mathematicians to sustain a high quality education system related to any of these topics. This problem is self-reinforcing because families of highly skilled technologists are less likely to choose to raise children in a weaker academic setting.

It is important to identify crucial steps of the plan that imperil the entire initiative if these elements fail. For the Hawaii resilience initiative, showstopper failure steps include passage of pertinent legislation, critical tsunami-warning buoy replacements/upgrades and related UAV support to transmit data to analysis centers, the smart grid upgrade, deployment of the hyper-local forecasting analytic engine, execution of the full scope of the broadband initiative, and assimilation of critical cybersecurity measures. For any plan, it is helpful to employ success metrics to evaluate progress. The initial performance metrics for the resilience initiative are mostly discrete in nature. These include, successful passage of the legislative elements, validation tests for the smart grid, validation tests for the hyper-local weather forecasting, demonstrations and validation tests for the tsunami-warning buoy system, a tangible and sizeable increase in the number of technology companies initiating business in Hawaii to support the resiliency initiative as well as related training companies, a significant increase in the number of students completing the resiliency academic programs, significant level of commerical leasing of the new broadband fibers, and measures that assess the cybersecurity posture of Hawaii. So, it is crucial that this stage of the initiative produce the measures described above.

To achieve sustained resilience, it is important to develop a plan to build and foster the industry, maximize commercialization of resulting technologies, advertise the resulting new technologies, training programs, and expertise, and to acquire or otherwise build and grow the budding industry. This is already the approach that is underway. The state, AIRS, and PACOM have used numerous open forums to advertise the initiative broadly and provide training, including multiple Resiliency and Sustainment Symposiums, conferences, workshops, and meetings. Table 1 lists numerous of these activities and events.

Table 1: Sustainment and Resiliency Outreach Activities

Date Range	Venue
06/22/14 to 06/25/14	Naval Postgraduate School's Cyber Endeavour/Cyber X-Games in Monterey, CA
09/09/14 to 09/10/14	IBM i2 Summit in Washington DC
09/29/14 to 10/03/14	TAG Summit in San Diego, CA
07/23/14 to 07/27/14	Aspen Security Forum in Aspen, CO
10/14/14 to 10/16/14	The 4th National Conference on Building Resilience through Public-Private Partnerships in Washington DC

 ISBN: 978-1-61208-423-7

Of note, the TAG's attendance in Cyber Endeavour/Cyber X-Games at the Naval Postgraduate School resulted in a mutually beneficial exchange with the CIP practitioners from the mainland. Parties from the State of Hawaii and the mainland acknowledged ongoing vulnerabilities, such as the San Jose attack on critical infrastructure as well as the Chicago Aurora Radar Center fire, which devolved operations to other airports. Unanticipated issues were discussed, such as those delineated in the recent publication, "Milk or Wine: Are Critical Infrastructure Protection Architectures Improving with Age?" Additionally, the TAG Summit utilized a Hawaii-centric approach, and the outlined engineering pathway is consistent with the spirit of DOD High Performance Computing Modernization Program (HPCMP) and the aforementioned EOs and PPDs.

Finally, this section of the strategic roadmap describes the Public Private Partnership Initiative that has been assembled for the resilience initiative and outlines roles for the various participants. Excerpts of this can be found on the Hawaii DOD websites [7] [8].

These documents outline the Public-Private Partnership for the Resiliency and Sustainment Initiative.

VI. CONCLUSION

This paper has presented a technology roadmap that outlines the path to achieve Resilience and Sustainability for Hawaii. This technology roadmap is a key element of the resiliency and sustainment initiative, which will lead development of these industries.

ACKNOWLEDGEMENT

The authors would like to thank the Cyber Futures Center, an initiative of the Sensemaking-U.S. Pacific Command Fellowship, IBM Center for Resiliency and Sustainability, and the Dr. Steve Chan Center for Sensemaking — one of the centers of the Asia-Pacific Institute for Resilience and Sustainability (AIRS), which is jointly anchored at Swansea University's Network Science Research Center and Hawaii Pacific University — for the opportunity to study the challenges facing Hawaii and to contribute towards the Public Private Partnership Initiatives aimed at developing solutions to overcome those challenges.

References

[1] State of Hawaii Department of Business, Economic Development, & Tourism, "Research and Economic Analysis: Outlook for the Economy", dbedt.hawaii.gov/economic/qser/outlook-economy/, 2015.

[2] D. Rachwald and T. Klemas, "Evolutionary Clustering Analysis of Multiple Edge Set Networks used for Modeling Ivory Coast Mobile Phone Data and Sensemaking" Data Analytics 2014, The Third International Conference on Data Analytics, pp. 100-104, August 2014.

[3] Industry Science Resources, "Technology Planning for Business Competitiveness, A Guide to Developing Technology Roadmaps", Emerging Industries Occasional Paper 13, Aug 2001.

[4] International Energy Agency, "Energy Technology Roadmaps, A guide to Development and Implementation", 2014 Edition.

[5] Farnoosh, N., Polimeridis, A.G., Klemas, T. ; Daniel, L. , "Accelerated Domain Decomposition FEM-BEM Solver for MRI via Discrete Empirical Interpolation Method", VLSI Design, Automation and Test Conference, pp. 1-4, 2014.

[6] M. Newman, "Networks, An Introduction", Oxford, Oxford University Press, 2010.

[7] http://dod.hawaii.gov/blog/news-release/new-partners-join-collaborative-effort-to-explore-methods-to-improve-energy-efficiency-and-grid-operations/, October 2014.

[8] http://dod.hawaii.gov/blog/news-release/state-of-hawaii-department-of-defense-office-of-homeland-security-hawaiian-electric-ibm-mehta-tech-incpacific-disaster-center-and-u-s-pacific-command-explore-methods-to-improve-energy-efficien/, February 2014.

[9] http://www.to-hawaii.com/natural-disasters.php, 2014.

[10] http://www.honolulu.hawaii.edu/instruct/natsci/geology/brill/gg101/Programs/program11%20Tsunami/program11.html, 2014.

[11] http://www.voanews.com/content/hawaii-vulnerable-yo-tsunamis-prepares-for-the-worst/2631262.html, 2015.

 ISBN: 978-1-61208-423-7

Hold the Drones

Fostering the Development of Big Data Paradigms through Regulatory Frameworks

Robert Spousta III
Dr. Steve Chan Center for Sensemaking, AIRS
Swansea University's NSRC and Hawaii Pacific University
Swansea, Wales
Email: spousta@mit.edu

Steve Chan
Dr. Steve Chan Center for Sensemaking, AIRS
Swansea University's NSRC and Hawaii Pacific University
Swansea, Wales
E-mail: s_chan@mit.edu

***Abstract*—We are at a critical phase in the proliferation of unmanned aircraft systems as a transformative technology, and the shape of regulatory policy for the broad civil use of these systems will be a determining factor in our ability to leverage pervasive remote sensing as a strategic national capability. In this paper, we explore the state of policy for civil unmanned aircraft systems and employ historical hindcasting of trends for comparably transformative technologies to gain insights into the role of public policy and regulation in the development of strategic capabilities. While the absence of a regulatory framework for unmanned aircraft operations has been a** blind spot **negatively impacting the growth of non-military unmanned aircraft capabilities to date, a prospective framework must strike a difficult balance between freedom and security. On the one hand, the American unmanned aircraft industry requires the freedom to experiment with innovative designs and applications. On the other hand, the American citizenry demands security against the potential threats posed by the misuse and malicious use of these systems. As we demonstrate with the example of space exploration, a clear vision of the goals to be achieved with a strategic capability is needed to drive the development and sustainment of that national capability, lest resources be wasted and control over it be ceded to competing nations. Similarly, the history of car making illustrates the danger of establishing policy that facilitates technological stagnation and systemic** brittleness **by absolving private industry of the imperative to innovate competitively and in the public interest. In light of these lessons, we find that a** resilient **regulatory framework must capitalize on the potential benefits of this promising technology while respecting the danger it poses.**

Keywords- Big Data, Blind Spots, Brittleness, Pervasive Remote Sensing, Resilience, Unmanned Aircraft System

I. INTRODUCTION

Pervasive remote sensing is a significant enabling capability for conducting critical infrastructure protection and other vital missions in a Big Data paradigm [1]. In turn, the rise of Unmanned Aircraft Systems (UAS) is the primary driver of the transition from satellite-based remote sensing to a **pervasive remote sensing** capability, and represents an area of rapidly evolving technology around the world [2]. While the United States has enjoyed a relative monopoly on such technology for military applications in the first decade of the 21st century, the slow development of a regulatory framework for their broader domestic use represents a **blind spot** that has hampered the nation's ability to maintain a qualitative edge over the use of UAS as a critical enabler for a variety of strategic capabilities. While the Federal Aviation Administration (FAA) and other U.S. Government (USG) entities have limited the use of UAS for public and commercial use for the time being, the development of a regulatory framework that fosters UAS growth and outlines a strategic vision for their broader role in national capabilities will generate wealth and serve the public good. While closely related and often complementary, national and commercial strategic capabilities are distinguishable primarily by their ultimate purpose; whereas commercial capabilities are developed to generate financial profit, national capabilities are developed in order to serve a public need, such as defense. Commercial capabilities can and frequently are marketed to governments in support of a national capability (i.e., the defense industrial base, commercial satellite imagery providers, contractors and private consultants for many Information and Communication Technology (ICT)-related functions, etc.).

Nations that embrace UAS through the development of robust regulatory frameworks will be postured to leverage the benefits of pervasive remote sensing and to mitigate the threats posed by the employment of UAS for malicious purposes. Such frameworks must incorporate a wide variety of social and technical considerations, from the potential for misuse of UAS platforms and the significance of individual air rights, to the latent **brittleness** of next-generation communications infrastructure that relies upon a particular frequency of the radio spectrum that is highly sensitive to atmospheric conditions (e.g., Ka Band). The current gap in U.S. policy with regard to UAS represents both a lost commercial economic opportunity and a potential erosion of national security.

In this paper, we aim to demonstrate how the development of policy and regulation regarding UAS impacts the U.S. at a national strategic level, in particular its ability to employ pervasive remote sensing within a Big Data paradigm. We begin in Section II by establishing a systemic context for understanding the impact of policy and regulation on the advancement of transformative technology

 ISBN: 978-1-61208-423-7

through historical hindcasting of automobile manufacturing and space exploration. The history of car making illustrates the danger of establishing policy that facilitates technological stagnation and systemic brittleness by absolving private industry of the imperative to innovate competitively and in the public interest. Similarly, the example of space exploration demonstrates the need for long term strategic vision to drive the development and sustainment of national capabilities, lest resources be wasted and control over them be ceded to competing nations. We go on in Section III to survey past and present implementation of UAS, and in Section IV we conduct a comparative analysis of national and international legal precedents which may bear relevance for UAS regulation. We find that while UAS have a significant military deployment history, as applications have expanded for their public and commercial domestic use, a commensurate regulatory framework has taken longer to develop in the U.S. While the absence of a regulatory framework for unmanned aircraft operations has been a blind spot impacting growth of non-military unmanned aircraft capabilities to date, a prospective framework must strike a difficult balance between freedom and security. On the one hand, industry requires freedom to experiment with innovative designs and applications. On the other hand, citizens demand security against threats posed by misuse and malicious use of these systems. We explore the consequences of this trend, and propose ways to improve leverage over UAS as a key enabling technology. We conclude in Section V that while current UAS policy is negatively impacting the economy and security of the U.S., such a trend is reversible. We also begin turning towards additional areas of strategic import in which a Big Data paradigm could be beneficially applied.

II. FROM CARS TO SPACESHIPS: SYSTEMIC CONTEXT FOR TRANSFORMATIVE TECHNOLOGY AND CAPABILITY DEVELOPMENT

In order to better appreciate the influence of policy on the development of pervasive remote sensing as a strategic national capability; it is illuminating to hindcast similar historical parallels. In doing so, we consider the rise of comparably transformative technologies; outlining the role they have played in national security and economic welfare. In particular, we take automobile manufacturing and space exploration as two areas which exemplify the importance of sustained innovation and forward-looking policy development. In both of these cases, we see that large investments fueled — initially — significant U.S. accomplishments, followed by a decrease in progressive momentum perpetuated by a mutually interactive combination of lax regulatory policy and industry malaise. The resulting lack of sustained innovation in both space capability and automobile manufacturing offered footholds for international competitors to capitalize on adaptations or expansions of early American achievement. In turn, the rise of international competition in both space endeavors and car making has born significant economic and national security consequences for the U.S. that help to illustrate the importance of fostering hospitable conditions to expand UAS capabilities in a Big Data Paradigm.

A. Automobile Manufacturing: the engine of innovation

The 20th century was a breakout era for mankind's advance in technological invention and critical problem-solving, which reached a crescendo with our arrival on the moon. Yet before mankind could reach into space, the car had to take him down the road. The production of the automobile begins as a story of individual rivals locked in a heated yet solitary contest to innovate, and unfolds as a lesson in the strength of group decision engineering. As illustrated below in Figure 1, automobile manufacturing was dominated by U.S. firms going into the second half of the last century, and yet the North American auto industry's doom appeared all but certain a few short years ago. The events that transpired during the intervening period show that while incremental innovation by individuals can yield significant technological breakthrough, it takes a whole society integrated around the technology's processes to truly maximize its value.

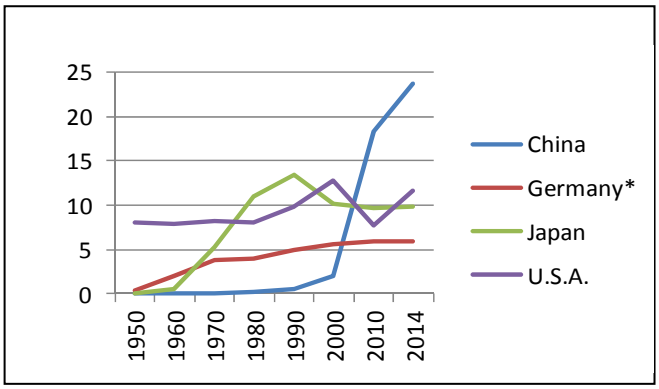


Figure 1. Annual Motor Vehicle Production (millions)
Sources: OICA; JAMA; U.S. Bureau of Transportation Statistics
**West Germany prior to 1990*

The first car was born out of competition to unify chemistry with physics and mathematics to achieve combustion-driven transportation. Whereas steam engines, wind power, and other power sources have remained common in transportation and other human processes, the first combustion engine fundamentally transformed individual human mobility. Two Germans, Carl Benz and Gottlieb Daimler each invented their own versions of an internal combustion engine mounted on wheeled vehicles within months of each other in 1896, working less than 100 miles apart [3]. However, it was roughly 4000 miles west

 ISBN: 978-1-61208-423-7

and 20 years later that Henry Ford's vision of the Model T truly revolutionized transportation by socializing the construction of vehicles on a massive scale.

Ford's breakthrough was in making cars affordable and widely available by adapting mass production techniques from other industries in his design of a modular platform [4]. Early car making was a time consuming and expensive process that resulted in a product which only the wealthy few could afford. However, by the early 1920s, Ford was producing 2 million Model Ts per year at a price that average citizens could pay for. Yet, such a breakthrough would not have been possible without the advent of the electric utility industry, the socialization of production, and the development of global supply chains, which facilitated the transition from belt-shaft networks of water wheels and coal-powered steam engines to more efficient unit drive assembly lines powered by large teams of skilled workers and electric motors [5]. Ford's role as an innovator is particularly notable not for his technological inventions, but for his integration of existing technologies and human skills that allowed him to achieve unprecedented production levels at a low cost. Similarly, Edward Budd's development of metal stamping improved assembly line efficiency, and Alfred Sloan's development of a comprehensive business model for the auto industry established the blueprint for how car makers could best market their products and maximize profits by employing ever larger groups in the auto ecosystem [6]. A single individual invented the first car on Earth, but now the global auto industry system comprises 50 million members of networked teams that bring 165 thousand new cars to market each day.

For the first half of the 20th century, American car manufacturers led the global auto industry by adhering to the model established by early leaders like Ford and Sloan, but their inability to sustain innovation compromised their position as a world leader. After World War II, the Japanese government instituted policies to protect the growth of Japanese auto makers by limiting the import of foreign cars to 1% of the domestic market, while manufacturers continually improved production efficiency through the adoption of just-in-time production techniques and decreasing worker specialization in favor of flexibility [7]. By the mid 1960s, Japanese productivity levels matched and surpassed that of its U.S. competitors. A critical factor for maximizing Japanese productivity was the horizontal integration of a highly organized network of component suppliers and assemblers, or keiretsu [8]. By engendering trust through exclusive transactions, close coordination, and information sharing, these keiretsu facilitated high levels of cooperative specialization between sectors of Japan's auto industry [8]. The keiretsu also enhanced the resilience of Japan's auto industry, as evidenced in 1997 by the Toyota group's ability to coordinate the actions of over 200 individual firms and quickly redirect production of a crucial brake system component after a fire destroyed the plant that had been the component's sole producer [9]. Meanwhile, U.S. production, characterized by vertically integrated and comparatively disorganized supplier-assembler networks remained largely constant into the 1980s, at which time Japanese production efficiency levels were vastly superior. U.S. manufacturers were path dependent, falsely assuming that their production efficiency either could not or did not need to be improved.

By the time of the worldwide economic crisis of 2007, the decline of the U.S. auto industry was drawn into sharp relief in contrast to skyrocketing Chinese production, begging the question of government's role in private industry. The bankruptcy of America's Big Three car makers (General Motors, Chrysler, and Ford) threatened to inflict the loss of one million jobs on the national economy, and the USG was forced to intercede with the Automotive Industry Financing Program, an $80 billion conditional industry bailout in 2009 [10]. Following in the tradition of technology-forcing legislation, such as the Clean Air Amendment Act of 1970 that mandated a reduction in carbon emissions [11], the conditional nature of the bailout enabled the USG to further induce U.S. automakers to embrace areas of innovation, particularly hybrid and electric vehicles, in order to increase their global competiveness. Nonetheless, the potential for a government bailout was itself a component of the American car industry's **brittleness**, in that the Big Three knew they could safely rely upon the precedent of bailouts established by the 1980 Chrysler Loan Guarantee Act, the post 9/11 airline industry bailout, and many other instances of the USG rescuing private companies from financial collapse [12]. Having established a universally known precedent for bailouts, the USG — in effect — dis-incentivized car makers from adapting their production to meet an evolving market.

The American experience in automobile manufacturing illustrates the imperative for continuous innovation, and the consequences for failing to heed that imperative. The early success of American auto makers led the U.S. to become a car-dependent society, but the ability of foreign auto makers to produce better cars at a cheaper price ultimately undermined the U.S. economy. Events like the 2009 bailout demonstrate that while industries cannot be forced to act strategically, government action and public policy play an important role in the development of technology. The history of car making also demonstrates the value of complementary technology, in that just as electricity facilitated mass production, UAS can facilitate pervasive remote sensing in a Big Data paradigm.

B. Outer Space: the sky is not the limit

The space race of the mid 20th century pushed the U.S. to achieve one of humanity's greatest accomplishments in

 ISBN: 978-1-61208-423-7

successfully journeying onto the moon and back, via the Apollo Program. Yet, little more than half a century later, the cession of American supremacy in space appears to be a near-term inevitability. What happened?

Driven by the Cold War urgency of winning the battle in space against the Soviet Union, the Apollo Program was a massive research and development effort with a single focus; getting to the moon first. However, the U.S. lacked a strategic vision of what to do with its hard-won space capability after achieving that feat, and was therefore challenged to follow up its huge investment with coherent progression. Although successive U.S. space programs have benefited from a more deliberate approach, they have also generally continued on Apollo's trajectory of increasingly complex and aggregated projects, which are expensive and subject to long development timelines [13].

Meanwhile, with the help of U.S. policies, other countries have developed notable space capabilities of their own. During the 1960s, the U.S. led the development of a regulated commercial space industry, with universal standards promoted by organizations like the International Telecommunications Satellite Organization (Intelsat). However, beginning in the early 1970s with the launch of the Open Skies initiative, the progressive deregulation of the satellite industry fueled the growth of global competition in space and gave rise to an increase in the number of small private firms in favor of large conglomerates like Intelsat [14]. At the same time that U.S.-led deregulation helped to increase the number of countries venturing into space, stringent export control laws severely limited the ability of American companies to capitalize on the expanding global market [15]. In addition, the refusal to carry foreign satellites into orbit aboard U.S. launch vehicles forced other countries to develop their own launch capability. A prime example of this dynamic is France's Arianespace, which was the first and remains among the world's largest commercial space launch providers [16].

While the development or acquisition of a space capability still requires significant national resources, including robust scientific and technological human capital, over 50 countries now have satellites in space and 12 have demonstrated a space launch capability [17]. To determine America's standing in this celestial mix, a review of two basic indicators is informative: where spacecraft are built and where they are launched from. Of the spacecraft launched in 2013, only 27% were manufactured in the U.S., compared with 41% in 2009 [18]. In the period 2000-2011, 80% of commercial low-earth orbit satellites and 90% of commercial geosynchronous earth orbit satellites were launched outside the U.S. [19]. These trends produce interesting outcomes, such as when the Department of Defense (DoD) is forced to rely on Chinese satellites to meet the communications requirements of U.S. Geographic Combatant Commands [20]. Yet, as commercial space operations have expanded and the nature of space capabilities have transformed, the U.S. has demonstrated its ability to continue making important breakthroughs in space. In contrast to other U.S. strategic space capabilities that rely on a small amount of large and hard to defend assets, the Global Positioning System (GPS) developed by the DoD leverages a distributed architecture consisting of a variety of assets that lend to the system's resilience by avoiding single points of failure [21]. Yet after 20 years in development, and despite becoming the world's primary navigation utility, GPS has not generated revenue to help offset U.S. investments in space and the system is vulnerable to a variety of threats including spectrum encroachment, jamming, spoofing, and space weather [22]. In addition, competing systems like Europe's Galileo, Russia's Global Navigation Satellite System (GLONASS) [23], and China's BeiDou Satellite Constellation [24] are all competing technologies with the potential to overtake the now aging GPS in the areas of accuracy and reliability.

Today, space assets are more vital to national security than ever before for their role in collecting and distributing information, but the U.S. ability to safeguard these assets is also more challenged than ever before [25]. While products of the Cold War space rivalry have been combined to achieve a monumental feat of global scientific and technological cooperation in the form of the International Space Station [26], emerging rivalries threaten to upset the extraterrestrial balance of power. In particular, China's rapidly expanding space program represents a potentially significant destabilizing force for U.S. space operations [27]. Since terminating its manned space shuttle program in 2011 in exchange for commercial crew and cargo programs, the U.S. has adopted a space strategy that relies on the cooperation and capabilities of private industry and other nations [28]. This policy shift has introduced a potential **blind spot** for the USG, in that it has divested itself of an engineering capability which took several generations to attain, and would ostensibly take several generations to reclaim. Meanwhile, China's national space program continues to progress along a deliberate and independent trajectory, gaining in sophistication with each mission [29]. Although the consequences of these divergent approaches to space have yet to fully materialize, it is clear that space is an area of increasing vulnerability for U.S. national security.

As unmanned aircraft technology advances, several key lessons from the ongoing American saga in space remain salient. First, a strategic vision of the broader capability to be achieved is a prerequisite for guiding the incremental development of scalable technology that will ultimately lead to that capability. Second, establishing a robust regulatory framework that accounts for both national security and revenue generation will ensure that a critical defense

 ISBN: 978-1-61208-423-7

capability does not have to sacrificed, because it is too expensive. This includes the ability to reconcile export control restrictions and allow industries to compete globally by marketing their technology overseas. Humanity's arrival in outer space is arguably among the most historically significant events in Earth's history, and the ecosystem of teams that can harness the potential of unmanned aircraft will propel the trajectory of exploration and capability into even as-yet unknown moments of innovation [30].

III. A Brief History of Unmanned Aircraft systems

Having seen how automobiles transformed ground transportation, we now move on to explore how the rise of unmanned aircraft and related systems is transforming aviation. Similar to the development of space capabilities, we will see how UAS grew from a national security tool into a ubiquitous technology. We will first trace the roots of early UAS application in war fighting and proceed to enumerate the diverse variety of devices and applications that have since evolved. Unmanned flight is not a recent development, but the increasing omnipresence of unmanned systems and their continually expanding functionality is novel. UAS, which include Unmanned Aerial Vehicles (UAVs) or drones, Remotely Piloted Aircraft (RPA), and other related technology refer to an aircraft and its associated elements that can operate without a human pilot onboard [31].

The history of unmanned flight is closely tied to international conflict and the evolving requirements of military operations. Indeed, the genesis of Unmanned Aircraft (UA) dates back nearly a century, to when American, British, and German inventors worked to develop aircraft like the Curtiss Speed-Scout and Kettering Bug for use in World War I [32]. During World War II, the British Queen Bees, American Denny Drones and German V-1 Buzzbombs were employed as pilot training aids in target practice and explosive ordinance delivery systems [33]. As the conclusion of the Second World War segued to a more protracted Cold War, Intelligence, Surveillance, and Reconnaissance (ISR) became a vital national capability. With the downing of U2 spy planes and capture or death of their pilots in 1959 over the Soviet Union and Cuba in 1962, the U.S. was forced to recognize the value of unmanned reconnaissance aircraft, and the Air Force and Central Intelligence Agency coordinated through the National Reconnaissance Office (NRO) to develop multiple variants of the Ryan Firebee, which were flown extensively during the Vietnam War in order to conduct surveillance and battle damage assessments [34]. While the intelligence community was a significant contributor to the development of unmanned capability, via the NRO, through the 1970s and into the 1980s, the U.S. reduced its focus on UAS in favor of satellite reconnaissance, and by 1991, the U.S. looked to Israel's Pioneer unmanned platform for ISR support over Iraq [35]. While satellites are a vital component of national intelligence capability, they are constrained in their ability to adapt to mobile objects of interest. The re-commissioning of SR-71 Blackbirds into military service in the mid 1990s demonstrates the unchanging need for a responsive and flexible reconnaissance capability, which satellites simply cannot fulfill in light of their fixed orbits [36].

As a result of Pioneer's significant contributions during the Persian Gulf War, the DoD increased its own research and development efforts for unmanned systems, and fielded the Predator in operations over the Balkan Peninsula in the mid 1990s. Imagery generated by the Predator and other remote sensing assets was so useful during negotiations of the 1995 Dayton Peace Accords that the National Imagery and Mapping Agency (NIMA) was created the following year, combining personnel from eight agencies to lead the integration of cartographic imagery and intelligence analysis [37]. The USG continued to increase its investment in UAS into the new millennium, and NIMA's transformation into the National Geospatial-Intelligence Agency (NGA) in 2003 represents the vital role that remote sensing has come to play in national security.

While NGA is the USG's lead integrator of remote sensing imagery, including that collected with unmanned aircraft, each of the military services now employ a large and diverse fleet of UAS for a variety of long-endurance and high-risk missions. These include ISR, force protection, resupply, signals collection, and direct strikes [38]. In fact, the DoD's inventory of UA is fast approaching that of manned aircraft, at roughly 7,500 and 10,700, respectively [39]. And these unmanned assets are generating vast amounts of data; at the height of U.S. campaigns in both Iraq and Afghanistan, UAS generated 24 years' worth of surveillance in a single year [40]. The operation of just one Global Hawk UAS generates 500 megabits of data per second, which is about five times the satellite-relayed data flow or bandwidth used by the entire U.S. military during the Persian Gulf War [41]. The explosion in data throughput requirements brought on by UAS capability has introduced its own set of challenges, as the expansion of fiber optic cable networks have stunted the growth of satellite bandwidth. During early deployments at the onset of Operation Enduring Freedom in Afghanistan, operators of the Global Hawk frequently had to lower its video resolution and cope with fuzzier images in order to avoid overwhelming the capacity of communication systems. Indeed, the availability of satellite bandwidth will continue to be an important consideration for both military and civil UAS operations going forward.

While the technical achievements of UAS in war are significant, it is important to note that their use for kinetic operations or direct strike missions is not without

 ISBN: 978-1-61208-423-7

controversy [42]. The United Kingdom's Ministry of Defence has acknowledged that unmanned direct strikes may actually undermine military campaigns by giving adversaries a "potent propaganda weapon" [43]. The precedent which the U.S. and its coalition partners have established by using UAS overseas for targeted killings raises important questions about international regulation in light of recent developments in Pakistan and elsewhere [44]. We will explore this issue further in the following section.

While the military service record of UAS for carrying out dull, dirty, and dangerous missions is well-established, their employment for non-military use represents an area of potentially enormous expansion. As demonstrated below in Figure 2, military applications continue to dominate UAS sales, and the civil UAS market is controlled by a small number of manufacturers. Within non-military UAS applications, the FAA delineates three broad civil categories: public (i.e. governmental), commercial, and private. UAS use is growing rapidly in each of these areas, as we will further explore below.

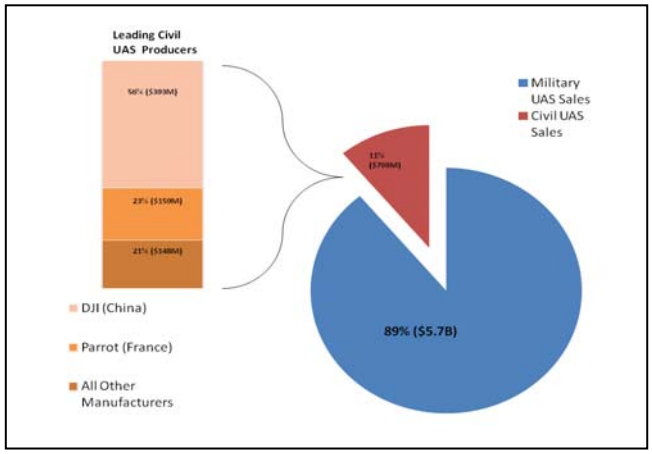


Figure 2. Estimated 2014 UAS Market Characteristics, Sources: Bloomberg News; Teal Group Corporation

Employing UAS as remote sensors holds promise for many public services; because it enables civilian government agencies to collect information that otherwise would be prohibitively expensive to gather using manned aircraft or satellite surveillance. Such a capability can be particularly valuable in safeguarding critical infrastructure and responding to natural disasters. For example, the early detection and continuous tracking of forest fires is a perennial challenge due to the inaccessible and mountainous terrain in which many fires occur. However, by using UAS to detect the outbreak and monitor the path of forest fires, state and federal responders are able to safely and more effectively stop their spread [45]. Similarly, law enforcement officers are beginning to use drones to detect illegal activities and track perpetrators, a capability that was historically limited by the cost of manned helicopters [46]. The Department of Homeland Security has been using UAS since 2004 to help close the gap in its ability to monitor isolated portions of the southern U.S. land and littoral borders, and today operates a fleet of 10 UAS platforms, with plans to expand the program in the future [47]. UAS can also play a pivotal role in environmental monitoring and enhancing our ability to understand and predict extreme weather phenomena by enabling scientists to collect more precise and complete climatic data, with the National Aeronautic and Space Administration's Helios project being one notable example [48]. Similarly, natural resource management efforts, including analysis of the effects of livestock grazing on the health of rangeland ecology are benefitting from UAS capabilities [49]. Remote sensing via UAS is also enabling federal and state Departments of Transportation to conduct traffic surveillance, assess road conditions, analyze travel patterns, and detect emergencies [50]. These examples are only a glimpse of the many potential benefits to be gained by the public use of UAS.

Commercial applications for UAS are equally varied, with only a small portion of potential uses having been realized thus far. In addition to the potential for UAS to enhance critical infrastructure protection, which combines aspects of public safety and commercial benefit, there are many opportunities for improved business efficiency. In Japan, 90% of all precision pesticide-spraying is done with a fleet of over 2500 unmanned helicopters [51]. Other examples include real estate mapping, aerial news and sporting event coverage, movie and television production, and cargo transportation. As UAS technology becomes more affordable, it is reasonable to expect that pervasive remote sensing itself will be marketed as a commodity in much the same way that smart phones have given rise to novel data-driven services [52].

Private UAS use carries on a well-established tradition of model aircraft piloting for recreational purposes, but also represents a significant threat if used for malicious purposes. As we have demonstrated in earlier research, UAS represent an important component in Improvisational Malignant Devices (coined as IMDs), which are characterized by low levels of sophistication and required resources, yet can yield significant destabilizing impact on complex systems such as critical infrastructure. While the U.S. has demonstrated some success in averting plans to employ UAS in malicious acts [53], events like the recent White House fly over and crash landing underscore the challenges associated with quickly detecting and responding to such acts as they occur [54].

IV. COMPARING U.S. AND INTERNATIONAL LEGAL PRECEDENT TO INFORM UAS REGULATION

Having established the comparatively long history of UA operations, and the wide variety of applications into which their employment has expanded, we now turn to the policies and regulations which govern their use. While Congress has mandated that regulations be developed to govern the

 ISBN: 978-1-61208-423-7

operation of UAS in the National Airspace System (NAS) before the end of this year, the policy of the FAA for the last ten years has been to broadly prohibit the operation of UAS for public or commercial purposes, instead regulating their exceptional limited use by issuing special air worthiness certificates and certificates of waiver or authorization [55].

This tact contrasts sharply with the U.S. Commercial Remote Sensing Policy, which asserts that maintaining the nation's leadership in remote sensing activities and enhancing the industry will protect national security and foster economic growth [56]. A more deliberate policy linkage between remote sensing and UAS could go a long way to reconciling this divergence, and promoting the advance of national capabilities in pervasive remote sensing. The FAA's recent release of a notice of proposed rulemaking for operation and certification of small UAS is a promising first step towards opening a sliver of the NAS to commercial unmanned activities [57]. The proposal reflects a balanced incremental approach, as it would place narrow limits on UAS operations and institute safe guards such as security threat assessments for prospective operators and mandatory device registration.

With regard to private operations, FAA's guidance for model aircraft from 1981 has been applied to UAS, advising that aircraft be operated away from populated areas at no higher than 400 ft above the ground, at least three miles from airports [58]. However, as with the proposed small UAS rule, such an advisory relies largely on the ability of local law enforcement to detect the misuse of UAS, and does not establish a systematic mechanism for addressing misuse or malicious use. As Figure 3 illustrates, there are a variety of complex dynamics at play in UAS regulation.

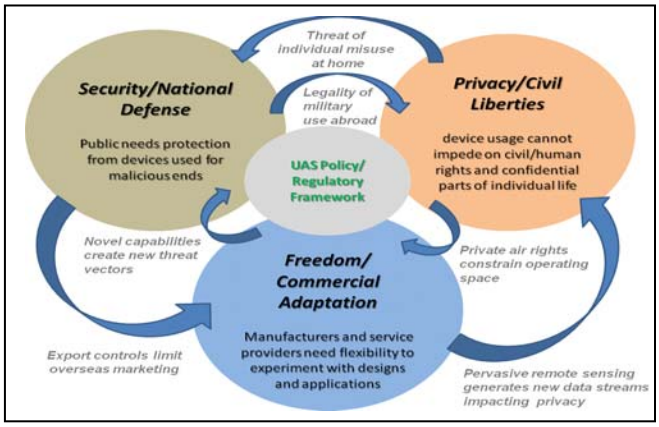


Figure 3. Sample of competing systemic factors impacting the development of comprehensive UAS policy

Indeed, any robust regulatory framework for unmanned aircraft operations must address the **blind spot** of maliciously-employed UAS as an emerging threat vector. To be sure, the development of policy for the broad civilian use and commercialization of domestic unmanned flight is no simple task. The difficulty of this task is compounded by the need to ensure harmony with a variety of contending issues as depicted in Figure 3, not to mention the technical complexity of UAS themselves.

While the FAA has rightly focused on the practical mechanics of safe operation, such as sense and avoid protocols, airworthiness standards, and pilot certification [59], a host of broader existential challenges also loom. For example, the case law for air rights establishes that the owner of a property also owns and is entitled to exclusive use of as much of the uncontrolled airspace above that property as they are reasonably capable of using [60].With the advent of UAS, property owners are now capable of using much more of their airspace. Therefore, a careful balance must be struck to ensure that public and commercial UAS are able to operate effectively without infringing on citizens' rights to their own airspace. Meanwhile, defining what constitutes acceptable use of one's airspace is also a central concern. As the State Department has encountered resistance from host nations regarding the U.S. authority to collect and disseminate data from the airspace above its embassies [61], it is clear that enhanced data collection capability will require more sophisticated forms of regulation.

In addition to reconciling potential conflict with existing law, UAS regulations must also complement the FAA's larger Next Generation Air Transportation System (NextGen) transformation effort [62]. NextGen aims to leverage satellite communication to supersede the currently overburdened radar systems in order to increase air traffic volume, safety, and efficiency. But, how to achieve these goals while integrating UAS is an open question, albeit one that appears to lend itself well to a Big Data paradigm based on effective management of increased data availability. In the NextGen system, more networked communication between air traffic controllers, aircraft pilots, and aircraft themselves will result in much larger amounts of data being generated, which raises important socio-techno concerns. Broadly speaking, we must determine how the roles of man and machine in air traffic control operations should evolve. More specifically, we must determine whether trends such as the Federal Communications Commission's support of an industry-wide shift from the Ku to Ka frequency bands for satellite links with UAS and other earth stations is introducing **brittleness** into the national communications infrastructure in light of Ka band's demonstrated vulnerability to signal attenuation in moist atmospheric conditions [63].

Although the tenets of international UAS regulation are perhaps even more ambiguous than those of U.S. policy, a review of legal precedent is instructive. The basic freedoms of the air established in the Chicago Convention and promoted by the United Nations (UN) International Civil Aviation Organization (ICAO) address issues of passenger aircraft, providing that states may grant each other the privileges of flying across, landing in, taking on, and putting

 ISBN: 978-1-61208-423-7

down traffic between states [64]. The ICAO has identified preliminary steps to bring UAS under the Chicago Convention rubric, but the transformative nature of the technology may warrant an even more fundamental restructuring of the framework governing air operations.

In this regard, the principles guiding maritime affairs potentially offer insight. In particular, Admiralty Law governing maritime navigation and shipping establishes that a ship's flag determines the source of law, such that vessels traveling outside their own national waters remain subject to the laws of their home nation. Assuming the U.S. and other nations develop regulations for the use of UAS in their own borders, applying the Admiralty principle to unmanned operation in international airspace appears logical. In addition, the UN Convention on the Law of the Sea (LOS) [65] establishes territorial seas in which the sovereignty of a state is extended 12 miles beyond its shore, including airspace above the water. Foreign vessels are permitted innocent or transitory passage through another nation's territorial waters, but solely for the purpose of traversal. Notably, conducting any survey activities during the passage of another nation's territorial waters is construed as prejudicial to the peace of that nation, and therefore illegal. The Convention also establishes Exclusive Economic Zones (EEZs) extending 200 nautical miles from a sovereign nation's shore in which that nation enjoys exclusive commercial and exploratory rights. Any area outside the territorial seas and EEZs are designated as the high seas, and are open to all states for peaceful purposes.

Extrapolating from the LOS, international airspace correlates neatly to the high seas and controlled national airspace correlates to territorial seas, but what about exclusive economic zones? As remote sensing capabilities expand with UAS, public and commercial applications requiring global circumnavigation will undoubtedly emerge. U.S. national airspace above 60,000 feet is currently designated Class E, the least regulated of any of the six airspace classes. Looking above the atmosphere, the Outer Space Treaty establishes that all nations and non-governmental organizations have the right to freely explore outer space without any discrimination [66]. From this context, an upper limit of nationally controlled airspace above which nations could freely navigate UAS is conceivable.

The employment of UAS across international borders for military operations is governed by established laws of armed conflict such as the 1949 Geneva Convention, yet new precedent is unquestionably being established by the U.S. amidst its global pursuit of Al-Qa'ida and affiliated entities [67]. Whether the protracted deployment of UAS for worldwide low intensity applications of force is indeed conducive to a stable international system is somewhat doubtful. In contrast, the Antarctic Treaty System (ATS) offers a more viable alternative. It establishes that as the only continent with no recognized or disputed claims of sovereignty, Antarctica will be used solely for peaceful purposes, namely scientific investigation and cooperation between its 50 signatories. While conflicts regarding the ATS do arise, such as the dispute between militant conservationists and whale "research" vessels [68], the cooperative spirit of the ATS lends credibility to a similarly open arrangement for globally operating UAS. Enabling the use of UAS for pervasive remote sensing increases our data collection capacity, this in turn increases our understanding of complex phenomena and contributes to enhanced **resilience**. However, addressing the privacy and security ramifications of a global pervasive remote sensing capability will be of chief importance to future international UAS regulations.

Although the exact form of UAS regulation has yet to crystallize, several facts are clear. First, the de-facto ban on public and commercial operations in the U.S. has confined the development of non-military UAS production. It is estimated that growth of the civil UAS industry will generate 70 thousand jobs in the first three years of integration and $80 billion over the next ten years, with each day of non-integration representing nearly a $28 million loss [69]. Indeed, the world's top two producers of commercial UAS are outside the U.S., and in an ironic turn of events, the platform being touted as the "Model T of unmanned aircraft" – The DJI Phantom – is being produced in the Silicon Valley of the East; Shenzhen, China [70]. Second, as UAS become more widely available, their potential to destabilize **brittle** systems through accidental misuse or deliberate malicious action will increase. Although they are areas for future research, geo-fencing and mandatory device registration are two possible components of a technical solution to UAS malicious use. More generally, developing policies and regulations that foster innovation and harness UAS as pervasive remote sensors can both mitigate the potential threat of **blind spots** posed by such technology while leveraging it to enhance **resilience**. Most importantly, creating a strategic vision that builds on the military and intelligence value of UAS by incorporating the technology into each of the remaining elements of national power can strengthen the nation's economy and expand its diplomatic reach.

V. CONCLUSION

From the assertion that Big Data is essential to building critical infrastructural **resilience**, we have come to the question centering upon how that capability is actually developed at a national strategic level through public policy and regulation. We are at a critical phase in the proliferation of unmanned aircraft systems as a transformative technology, and the shape of regulatory policy for the broad civil use of these systems will be a determining factor in the fate of pervasive remote sensing as a strategic national capability. UAS offer a potential doorway to pervasive remote sensing in a Big Data Paradigm. But, in order to unlock the door, public policy must catch up with

 ISBN: 978-1-61208-423-7

technology. Our historical hindcasting of trends in international space capability and automobile manufacturing underscore the influence that policy and regulation exert on the development of transformative technology. Through these cases, the potential for **blind spots** in public policy to introduce **brittleness** into critically important national capabilities is clear. A resilient civil UAS regulatory framework can and must capitalize on the potential benefits of this promising technology while respecting the danger it poses. Unmanned aircraft systems have shown significant success as a tool for generating Big Data to inform overseas military and intelligence operations, yet as applications are quickly expanding worldwide for their civil use, a commensurate regulatory framework for the systematic integration into the national airspace system has taken longer to develop. This constitutes a significant **blind spot** that is resulting in a loss of economic opportunities and degradation of national security.

While unmanned aircraft pose a unique set of policy challenges, the development of a robust regulatory framework for their civil operation is not an insurmountable task. In order to be effective, such a framework must outline a strategic vision for employing UAS as a national capability while directly addressing the security threats posed by such technology. In particular, sound UAS policy will include mechanisms that incentivize industry to develop technology that is both commercially competitive in the global marketplace, and complementary to national strategic priorities. In turn, the technological advantages presented by unmanned aircraft systems have the potential to yield vast increases in the amount of data available to engineer more sound decisions, including decisions regarding the prevention and mitigation of UAS malicious use.

This increase in available data and enhanced decision engineering is at the core of a Big Data Paradigm for pervasive remote sensing, and can improve our approach to a variety of missions, including critical infrastructure protection, homeland defense, law enforcement, resource management, environmental stewardship, and disaster response. Pervasive remote sensing will drive the advance of analytics in a host of commercial and research fields, as it makes more data available. However, this potential can only be realized if the proliferation of UAS is managed proactively and wisely. In light of a Big Data paradigm's value for these issues, we look forward to future work exploring what other areas of strategic interest might similarly benefit from such a paradigm.

ACKNOWLEDGMENT

The authors would like to thank the Cyber Futures Center, an initiative of the Sensemaking-U.S. Pacific Command Fellowship, and the Dr. Steve Chan Center for Sensemaking — one of the centers of the Asia-Pacific Institute for Resilience and Sustainability (AIRS), which is jointly anchored at Swansea University's Network Science Research Center and Hawaii Pacific University — for the opportunity to study the challenges facing Hawaii and other archipelagos, and to contribute towards the various Public Private Partnership Initiatives aimed at developing solutions to overcome those challenges.

REFERENCES

[1] A. Zaslavsky, C. Perera, and D. Georgakopoulos, "Sensing as a service and big data," arXiv preprint arXiv:1301.0159, 2013.

[2] F. Viani, P. Rocca, G. Oliveri, and A. Massa, "Pervasive remote sensing through WSNs," in Antennas and Propagation (EUCAP), 2012 6th European Conference on, 2012, pp. 49-50.

[3] L. Dorrington. (2011, January 24, 2011) 125th Anniversary of the Automobile: Karl Benz and Gottlieb Daimler put the world on wheels Autoweek. Available: http://autoweek.com/article/car-news/125th-anniversary-automobile-karl-benz-and-gottlieb-daimler-put-world-wheels-0 accessed February 15, 2015

[4] F. Alizon, S. B. Shooter, and T. W. Simpson, "Henry Ford and the Model T: lessons for product platforming and mass customization," Design Studies, vol. 30, pp. 588-605, 2009.

[5] R. B. D. Boff, "The Introduction of Electric Power in American Manufacturing," The Economic History Review, vol. 20, pp. 509-518, 1967.

[6] P. Wells and P. Nieuwenhuis, "Transition failure: Understanding continuity in the automotive industry," Technological Forecasting and Social Change, vol. 79, pp. 1681-1692, 2012.

[7] M. A. Cusumano, "Manufacturing innovation: lessons from the Japanese auto industry," Sloan Management Review, vol. 29, 2013.

[8] J. H. Dyer, "Specialized Supplier Networks as a Source of Competitive Advantage: Evidence from the Auto Industry," Strategic Management Journal, vol. 17, pp. 271-291, 1996.

[9] T. Nishiguchi, A. Beaudet, and B. M. Strategy, "The Toyota group and the Aisin fire," Image, 2012.

[10] Treasury, "The Department of the Treasury Office of Financial Stability – Troubled Asset Relief Program Citizens' Report Fiscal year 2014," D. o. t. Treasury, Ed., ed. Washington, D.C., 2014.

[11] C. Berggren and T. Magnusson, "Reducing automotive emissions—The potentials of combustion engine technologies and the power of policy," Energy Policy, vol. 41, pp. 636-643, 2012.

[12] J. N. K. K. Schmidt. (2009, April 15, 2009) History of U.S. Government Bailouts. ProPublica. Available: http://www.propublica.org/special/government-bailouts#tarp accessed February 10, 2015

[13] E. Pawlikowski, D. Loverro, and T. Cristler, "Space: Disruptive Challenges, New Opportunities, and New Strategies," Strategic Studies Quarterly, 2012.

[14] B. Warf, "International Competition Between Satellite and Fiber Optic Carriers: A Geographic Perspective," The Professional Geographer, vol. 58, pp. 1-11, 2006/02/01 2006.

[15] R. J. Zelnio, "Whose jurisdiction over the US commercial satellite industry? Factors affecting international security and competition," Space Policy, vol. 23, pp. 221-233, 2007.

[16] H. R. Hertzfeld, "Globalization, commercial space and spacepower in the USA," Space Policy, vol. 23, pp. 210-220, 2007.

[17] B. R. Early, "Exploring the Final Frontier: An Empirical Analysis of Global Civil Space Proliferation," International Studies Quarterly, vol. 58, pp. 55-67, 2014.

[18] SIA, "State of the Satellite Industry Report," Satellite Industry Association, Washington, D.C.2014.

[19] HPA, "The Impact of U.S. Space Transportation Policy on the Commercially Hosted Payload Enterprise," Hosted Payloads Alliance, Deerfield, Illinois2012.

[20] M. Gruss, "Pentagon Lease of Chinese Bandwidth Arouses Concern " in Space News, ed, 2013.

[21] P. Enge and P. Misra, "Special Issue on Global Positioning System," Proceedings of the IEEE, vol. 87, pp. 3-15, 1999.

[22] GAO, "GPS Disruptions: Efforts to Assess Risks to Critical Infrastructure and Coordinate Agency Actions Should Be Enhanced," U. S. G. A. Office, Ed., ed. Washington, D.C., 2013.

[23] S. Cojocaru, E. Birsan, G. Batrinca, and P. Arsenie, "GPS-GLONASS-

 ISBN: 978-1-61208-423-7

GALILEO: a dynamical comparison," Journal of Navigation, vol. 62, pp. 135-150, 2009.
[24] J. C. Moltz, "Technology: Asia's space race," Nature, vol. 480, pp. 171-173, 2011.
[25] D. Rumsfeld, D. Andrews, R. Davis, H. Estes, R. Fogleman, J. Garner, et al., "Report of the Commission to Assess United States National Security Space Management and Organization," Government Printing Office, Washington, DC, 2001.
[26] L. J. DeLucas, "International space station," Acta Astronautica, vol. 38, pp. 613-619, 1996.
[27] W. C. Martel and T. Yoshihara, "Averting a Sino-U.S. space race," The Washington Quarterly, vol. 26, pp. 19-35, 2003/09/01 2003.
[28] J. M. Logsdon, "Change and continuity in US space policy," Space Policy, vol. 27, pp. 1-2, 2011.
[29] E. Strickland, "The next space super-power," Spectrum, IEEE, vol. 51, pp. 48-51, 2014.
[30] S. Burleigh, V. G. Cerf, J. Crowcroft, and V. Tsaoussidis, "Space for Internet and Internet for space," Ad Hoc Networks, vol. 23, pp. 80-86, 2014.
[31] ICAO, "Circular 328, Unmanned Aerial Systems," I. C. A. Organization, Ed., ed. Montreal, Quebec, Canada, 2011.
[32] K. L. B. Cook, "The Silent Force Multiplier: The History and Role of UAVs in Warfare," in Aerospace Conference, 2007 IEEE, 2007, pp. 1-7.
[33] L. R. Newcome, Unmanned aviation: a brief history of unmanned aerial vehicles: Pen and Sword, 2005.
[34] J. M. Sullivan, "Revolution or evolution? The rise of the UAVs," in Technology and Society, 2005. Weapons and Wires: Prevention and Safety in a Time of Fear. ISTAS 2005. Proceedings. 2005 International Symposium on, 2005, pp. 94-101.
[35] T. P. Ehrhard, "Air Force UAV's: The Secret History," Mitchell Institute for Airpower Studies, Arlington, VA2010.
[36] SASC/HASC, "National Defense Authorization Act and Military Construction Authorization Act for Fiscal Year 1995 - Conference Report," U. Congress, Ed., ed. Washington, D.C.: Congressional Record, 1994.
[37] NGA, "The Advent of the National Geospatial-Intelligence Agency," O. o. t. H. Historian, Ed., ed. St. Louis, MO, 2011.
[38] OSD, "Unmanned Aircraft Systems Roadmap 2005-2030," D. o. Defense, Ed., ed. Washington, D.C., 2005.
[39] OSD AT&L, "Department of Defense Report to Congress on Future Unmanned Aircraft Systems Training, Operations, and Sustainability ", D. o. Defense, Ed., ed. Washington, D.C., 2012.
[40] A. Bleicher, "Eyes in the Sky That See Too Much [Update]," Spectrum, IEEE, vol. 47, pp. 16-16, 2010.
[41] G. Jaffe. (April 10, 2002) Military Feels Bandwidth Squeeze As the Satellite Industry Sputters. Wall Street Journal. Available: http://www.wsj.com/articles/SB1018389902229614520 accessed February 12, 2015
[42] Stanford/NYU, "Living Under Drones: Death, Injury, and Trauma to Civilians from U.S. Drone Practices in Pakistan," International Human Rights and Conflict Resolution Clinic at Stanford Law School and Global Justice Clinic at NYU School of Law2012.
[43] U.K. MoD, "Joint Doctrine Note 2/11 The U.K. Approach to Unmanned Aircraft Systems," M. o. Defence, Ed., ed. Shrivenham, United Kingdom: Development, Concepts and Doctrine Centre, 2011.
[44] M. Mazzetti and M. Apuzzo, "Deep Support in Washington for C.I.A.'s Drone Missions," in New York Times, ed. New York, NY, 2015.
[45] D. W. Casbeer, R. W. Beard, T. W. McLain, L. Sai-Ming, and R. K. Mehra, "Forest fire monitoring with multiple small UAVs," in American Control Conference, 2005. Proceedings of the 2005, 2005, pp. 3530-3535 vol. 5.
[46] J. Horgan. (2013, March 2013) The Drones Come Home. National Geographic. Available: http://ngm.nationalgeographic.com/2013/03/unmanned-flight/horgan-text accessed February 18, 2015
[47] C. H. J. Gertler, "Homeland Security: Unmanned Aerial Vehicles and Border Surveillance ", C. R. Service, Ed., ed. Washington D.C., 2010.
[48] M. Dunbabin and L. Marques, "Robots for Environmental Monitoring: Significant Advancements and Applications," Robotics & Automation Magazine, IEEE, vol. 19, pp. 24-39, 2012.
[49] A. Rango, A. Laliberte, J. E. Herrick, C. Winters, K. Havstad, C. Steele, et al., "Unmanned aerial vehicle-based remote sensing for rangeland assessment, monitoring, and management," Journal of Applied Remote Sensing, vol. 3, pp. 033542-033542-15, 2009.
[50] A. Puri, "A survey of unmanned aerial vehicles (UAV) for traffic surveillance," Department of computer science and engineering, University of South Florida, 2005.
[51] NRC, Autonomy Research for Civil Aviation: Toward a New Era of Flight. Washington, DC: The National Academies Press, 2014.
[52] W. Hongwei and H. Wenbo, "A Reservation-based Smart Parking System," in Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on, 2011, pp. 690-695.
[53] J. Bidgood, "Massachusetts Man Gets 17 Years in Terrorist Plot," in New York Times, ed, 2012.
[54] M. S. M. Shear, "A Drone, Too Small for Radar to Detect, Rattles the White House," in New York Times, ed. New York City, NY, 2014.
[55] H. G. Wolf, "Unmanned aircraft systems integration into the national airspace," in Aerospace Conference, 2013 IEEE, 2013, pp. 1-16.
[56] "White House U.S. Commercial Remote Sensing Policy," ed. Washington, D.C., 2003.
[57] Operation and Certification of Small Unmanned Aircraft Systems, U. S. D. o. Transportation RIN 2120–AJ60, 2015.
[58] FAA, "Department of Transportation Federal Aviation Administration Advisory Circular 91-57 Model Aircraft Operating Standards," U. S. D. o. Transportation, Ed., ed. Washington, D.C., 1981.
[59] FAA, "Integration of Civil Unmanned Aircraft Systems in the National Airspace System Roadmap," U. S. D. o. Transportation, Ed., First Edition ed. Washington D.C. : FAA Communications, 2013.
[60] A. Madrigal. (2012, October 25, 2012) If I Fly a UAV Over My Neighbor's House, Is It Trespassing? The Atlantic. Available: http://www.theatlantic.com/technology/archive/2012/10/if-i-fly-a-uav-over-my-neighbors-house-is-it-trespassing/263431/ accessed February 20, 2015
[61] S. S. S. Olivier. (2012, June 13, 2012) China Has No Good Answer to the U.S. Embassy Pollution-Monitoring. The Atlantic. Available: http://www.theatlantic.com/international/archive/2012/06/china-has-no-good-answer-to-the-us-embassy-pollution-monitoring/258447/ accessed February 20, 2015
[62] R. N. Van Dyk, D. H. Pariseau, R. E. Dodson, B. T. Martin, A. T. Radcliffe, E. A. Austin, et al., "Systems integration of Unmanned Aircraft into the National Airspace: Part of the Federal Aviation Administration Next Generation Air Transportation System," in Systems and Information Design Symposium (SIEDS), 2012 IEEE, 2012, pp. 156-161.
[63] S. Sala, M. Zennaro, L. Sokol, A. Miao, R. Spousta, and S. Chan, "Mitigation of Rain-Induced Ka-Band Attenuation and Enhancement of Communications Resiliency in Sub-Saharan Africa," 2013.
[64] ICAO, "Manual on the Regulation of International Air Transport - Second Edition," I. C. A. Organization, Ed., ed. Montreal, CA, 2004.
[65] United Nations Convention on the Law of the Sea, U. Nations, 1982.
[66] R. S. Jakhu and J. N. Pelton, "The Global Legal Guidelines Governing Satellite Deployment," in Small Satellites and Their Regulation, ed: Springer, 2014, pp. 43-48.
[67] C. Heyns, "Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions " vol. A/68/382, U. N. G. Assembly, Ed., ed. New York, NY: U.N. General Assembly, 2013.
[68] D. R. Rothwell, "The Polar Regions and the Development of International Law: Contemporary Reflections and Twenty-First Century Challenges," The Yearbook of Polar Law Online, vol. 5, pp. 233-251, 2013.
[69] D. J. B. Vasigh, "The Economic Impact of Unmanned Aircraft Systems Integration in the United States," Association for Unmanned Vehicle Systems International, Arlington, VA2013.
[70] J. N. C. Murphy. (2014, November 10, 2014) Who Builds the World's Most Popular Drones? The Wall Street Journal. Available: http://www.wsj.com/articles/who-builds-the-worlds-most-popular-drones-1415645659 accessed February 20, 2015

 ISBN: 978-1-61208-423-7

Space Race 2.0

Expanding Global Internet Accessibility

Robert Spousta III, Steve Chan
Dr. Steve Chan Center for Sensemaking, AIRS
Swansea University NSRC and Hawaii Pacific University
Swansea, Wales, U.K.; Honolulu, HI, USA
Email: spousta@mit.edu, stevechan@post.harvard.edu

Bob Griffin
General Manager and Co-Director
IBM and IBM Center for Resiliency & Sustainability
Armonk, NY; San Diego, CA, USA
Email: (care of) alkassam@us.ibm.com

***Abstract*—In this paper, we examine cyber electromagnetic activities and avenues for expanding Internet accessibility. Delivering fast and reliable Internet access to people, whose physical isolation precludes connectivity by wires and other traditional means, is challenging. In response, key industry players are racing to take on this challenge both with serious financial backing as well as their own methodologies for creating a more accessible global Internet. However, each approach must overcome its own set of technological hurdles. By way of example, satellites can deliver Internet access to sparsely populated areas, but the cost of using satellite data connections can be very high. Drones, in comparison, can reach those customers at a much lower cost. Yet, both platforms rely on the Ku-band, which is essentially saturated during the day and thereby subject to low throughput and long delays. Even if satellites and drones utilize another part of the electromagnetic spectrum, such as Ka-band, rain fade remains a persistent problem. High-altitude planes that fly above commercial airlines and the weather can utilize lasers, which are at the cutting edge of connectivity research for their incredible accuracy and high throughput; however, laser beams get scattered by clouds. Finally, the challenges of keeping a network of balloons — that are traveling on the edge of space — on course and without leaking are plentiful. The main technological challenge to the underpinning mesh network backbone is the frequency of power outages that disrupt the network. In any case, the next generation Generativity Principle holds promise not only for optimizing the flow of data throughout the Internet, but also for maximizing the primary infrastructure of Internet access.**

Keywords-Big Data, brittleness, generativity principle, net neutrality, Internet accessibility

I. INTRODUCTION

With the number of networked devices surpassing the number of humans on the planet in 2008 [1], the Internet of Things (IoT) has become a ubiquitous aspect of daily life for one third of the world's population, the majority of whom live in developed countries [2]. The creation of the Internet was driven by a need to share resources, and today there is a challenge to share the Internet itself as a resource. In bringing the Internet to the remaining two thirds of the global citizenry, there is a balance to be struck between reach and resilience. On the one hand, the Internet is a valuable tool [3] for supporting the universal human right of accessing information [4], and maximizing its reach is clearly beneficial. On the other hand, the Internet must operate reliably in order to be of value, and efforts to expand its reach quickly should not come at the expense of the system's resilience. As collective reliance on Internet connectivity increases and a variety of actors endeavor to expand global Internet accessibility, the underlying communications infrastructure remains **brittle** in key areas. This brittleness is a potential **blind spot** compromising the **resilience** of essential functions such as international commerce, national defense, and disaster preparedness, which have become highly dependent on the Internet. In the effort to increase Internet accessibility, infrastructural resilience must remain a primary consideration. In this regard, much discourse has centered around smart cities [5]. Although urban centers are home to 80% of the global population and appropriately are a major focus of infrastructure improvement and protection [6], the importance of internet connectivity for rural and physically isolated areas cannot be overlooked. Indeed, for archipelagoes like Hawaii maintaining connectivity is a very real challenge, as broadband capacity is projected to run out by as early as 2017.

Therefore, we explore what options are available for expanding broadband Internet capacity for isolated populations such as those in Hawaii. We survey the current variety of endeavors being undertaken to expand access, and identify challenges related to each approach. In Section II, we begin with an overview of terrestrial and fixed broadband, including direct subscriber lines, fiber-optic cable, and mobile broadband. In Section III, we explore the capabilities and limitations of satellite broadband. In Section IV, we identify additional methods of broadband delivery being pursued by various actors, including the use of unmanned aircraft systems (UAS), high altitude platform (HAP), and balloons. In Section V, we evaluate efforts to expand internet accessibility in the context of the generativity principle and net neutrality debate, and we conclude in Section VI.

II. A BRIEF HISTORY OF THE INTERNET AND TERRESTRIAL CONNECTIVITY

Since the inception of the four-node Advanced Research Projects Agency Network (ARPANET) was connected by terrestrial hard line with 50 kbps of bandwidth in 1969 [7], the Internet has grown into a vast web of diverse connections spanning land, air, sea, and space with over 160,000 gbps of bandwidth powering over 185 million active websites [2]. Just as ARPANET's architects envisioned a network of only a few hundred national-level resources, and had to adapt

 ISBN: 978-1-61208-423-7

their operating principles to address the unforeseen growth in traffic precipitated by local area networks, today myriad stakeholders are contemplating how the Internet can be expanded to reach every individual on the planet. Although the foundational building blocks of packet switching and the attendant protocol suite (e.g., transmission control protocol (TCP), internet protocol (IP), and user datagram protocol (UDP)) remain in place and much work has been done to unite what were once fragmentary networks [8], the modern Internet is so large and complex that obtaining a clear picture of how data are flowing through it is no longer feasible [9]. Whereas it is difficult to analyze what is happening inside the Internet at any given moment, certain facts about its accessibility are clear. The Internet has fundamentally changed the nature of human communication by providing a platform for truly novel developments such as the World Wide Web, yet over four billion people in the world are living without Internet access [10]. As stylized in Figure 1, below, how to bring it to them is a key question.

Figure 1. Artistic Rendering of Global Internet Connectivity

The first potential course of action is continued expansion of the land and subsea-based fixed connections upon which the Internet was originally built. However, as the International Telecommunications Union (ITU) notes in its most recent report [2], the growth of fixed broadband in the form of asynchronous direct subscriber line (ADSL) and fiber-optic cable is leveling off at over 700 million subscriptions, or roughly ten percent of global penetration in favor of mobile broadband, with 85% of the fixed network's 11 million km of underwater lines in the Asia Pacific region. While the worldwide average price for fixed broadband subscriptions has dropped 70% in the past eight years [2], a digital divide has persisted in that developed countries with the highest connection speeds enjoy the lowest cost subscriptions, while consumers in developing countries with less robust backbone infrastructure must pay more for slower connection speeds. By way of example, in Serbia consumers pay the highest broadband subscription rates in Europe at 3.8% of Gross National Income per capita (GNI pc) for 5 mbps of bandwidth, while in many African countries the cost of an entry-level fixed broadband subscription with speeds of 1 mbps or less can cost more than 100% of the GNI pc due to low income generation and a limited number of cables linking the continent to the international Internet [2]. Conventional fixed Internet access through fiber-to-the-premises, cable modem, and direct subscriber lines represents a large investment on the part of industry as it is expensive to lay and maintain, which translates to high subscription rates for consumers.

Such high expense is particularly cost-prohibitive for physically isolated communities who are located far from central infrastructural hubs and unable to attract investment. Submarine cable systems that provide the crucial intercontinental connections linking the global Internet are dominated largely by only three groups; Alcatel-Lucent, TE SubCom, and NEC. With such limited competition, there is little incentive to improve upon the cost or durability of undersea cable networks, as the cost of constructing new systems has remained relatively fixed at approximately $35 thousand per kilometer and designed for a 25-year lifespan, with the first cables laid in 1988 nearing the end of their service life [11]. Part of this low competition and high cost is due to the inconsistent growth of the industry, which saw a brief period of over-investment during the dotcom boom, particularly in 2001, which led to a largely dormant cable industry until recent years. Undersea cables connect all but 15 countries, providing 87 tbps in global transoceanic bandwidth. However, in light of the huge capital required for their development, cable routes cater to the world's financial hubs, as depicted below in Figure 2. At the same time, fixed Internet routed through submarine cable is not without its vulnerabilities. Indeed, shark attacks, anchor snares, fishing accidents, and other unintentional anthropogenically-induced damage accounts for as many as 150 outages per year [12]. For physically isolated locations that are unable to attract investment in the form of cable landing sites, an alternative to fixed broadband is required, which takes us back to the Internet's early days.

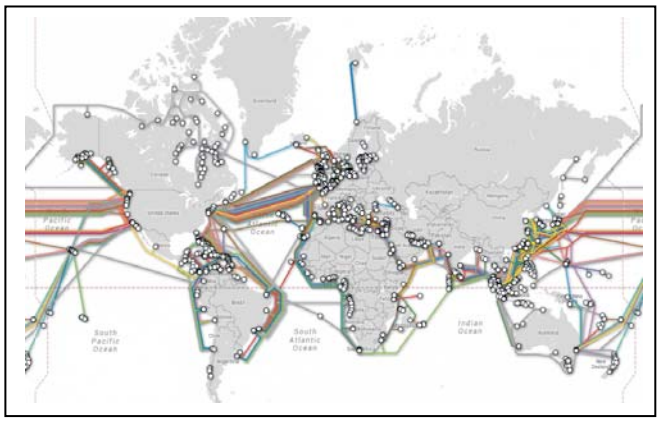


Figure 2. TeleGeography Global Submarine Cable Map

Research on wireless packet switching networks emerged simultaneously with the development of ARPANET, and

since then wireless complements to the fixed Internet have continued to evolve. The ALOHANET was the first instantiation of wireless computer communication, whereby University of Hawaii computing resources dispersed over the islands exchanged data through radio channels with each other and eventually ARPANET [13][1].This concept was adapted to achieve a lightweight and mobile packet switching capability for military applications through the Packet Radio Network (PRNET) in the 1970s [14]. Advances in wireless communications capability continued through the 1980s and 90s with cellular technology, and in 2010, the number of mobile broadband subscriptions surpassed that of fixed broadband subscriptions [15].

However, broadband Internet delivered through universal mobile telecommunications systems is constrained in several fundamental ways. First, mobile broadband access consumes significantly more power per user than fixed connections [16]. Second, mobile broadband connections are more inconsistent than fixed connections and subject to increased disruptions in service [17]. Just as fixed broadband relies on the abundance of backbone infrastructure, mobile connections rely on the strength of cellular signals and the proximity of towers. Finally, and perhaps most significantly, mobile broadband networks suffer from comparatively limited bandwidth capacity due to finite spectrum availability [18], which necessitates data offloading to fixed network connections through various means, including wireless fidelity (WiFi), femtocells, and IP flow mobility [19]. In fact, in 2014 26.4 exabytes of data representing 46% of total mobile broadband traffic was offloaded to fixed connections [20], demonstrating that mobile broadband is a complementary extension of the wired Internet, not a stand-alone replacement for it. Similarly, the Worldwide Interoperability for Microwave Access (WiMAX) Forum is helping to provide broadband without the need for each user to have a fixed connection, but the networks still require a large infrastructure of base stations that would be cost-prohibitive in isolated communities [21]. Therefore, we find that without significant infrastructural investment, neither fixed nor mobile broadband are viable options for connecting isolated populations.

III. INTERNET IN SPACE: SATELLITES

Although science fiction writer and futurist Arthur C. Clarke's first speculation about the potential for achieving a global broadcast capability through extraterrestrial relays seemed far-fetched to skeptical audiences in 1945, his vision has proved to be truly prophetic [22]. Indeed, the use of satellites for Internet connectivity is nearly as old as the Internet itself, however there are significant technical challenges associated with bringing the Internet to space. Geostationary orbiting (GSO) satellites have been used for some time to provide backbone connections for regional networks, with the Atlantic SATNET being an early example of implementing satellite communication for Internet protocols, as it provided a 64 kbps connection between the ARPANET and research networks in Europe from the late 1970s into the mid 1980s [23]. However, GSO satellites suffer from two major drawbacks; their enormous cost, and the latency of their communications. GSO satellites are located 35, 786 km from the Earth's surface, which can enable a single satellite to have broadcast coverage over a third of the planet's surface, yet also takes a signal approximately 280 milliseconds to travel each way, which amounts to over half a second in latency for roundtrips [24]. In contrast, low Earth orbiting (LEO) satellites located up to 3,000 km from the Earth's surface can overcome this latency problem, but given the proximity to the surface, their coverage area is severely limited, and therefore a network of satellites is required to relay signals. However, as depicted in Figure 3, below, early satellite Internet communications relied on a bent pipe configuration, whereby Earth uplinks were concatenated to Earth downlinks, and the satellites served as little more than signal relay points incapable of signal processing or dynamic routing [25]. Due to these limitations, early instantiations of the satellite-provided Internet were characterized by relatively high cost and low quality [26].

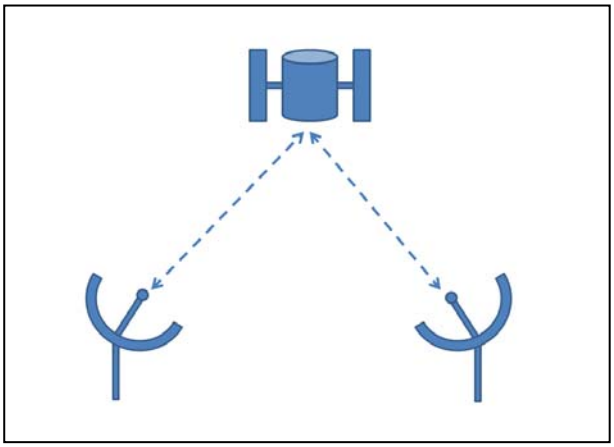


Figure 3. Bent Pipe Satellite Communications Architecture

With the explosive growth of internet users in the 1990s, demand for bandwidth appeared to be outpacing terrestrial network providers' ability to lay new wire and cable, and in response, companies endeavored to leverage satellites as an augmentation to the wired Internet. Hughes Network Systems' DirectPC Satellite System developed in 1996 enabled users to request data by phone line and modem, and download the results through 400 kbps direct link with Hughes' Galaxy GSO leveraging digital video broadcast by satellite (DVB-S) and very small aperture terminals (VSAT) [27]. Hughes expanded its services by incorporating medium Earth orbit (MEO) satellites in its Spaceway system in 2002, but was not alone in delivering broadband internet through GSO and MEO satellites on the DVB-S platform, as Cyberstar, ISky, Lockheed Martin's Astrolink, the European satellite conglomerate SES's Astra, Eutelsat's Hot Bird, Inmarsat, and Matra Marconi Space's Wideband European Satellite Telecommunications (WEST) represented some of the competition. GSO satellites using DVB-S platforms in

[1] Incidentally, this packet broadcasting technique also gave rise to Ethernet technology for local area networking.

 ISBN: 978-1-61208-423-7

the Ku Band have been very effective for broadcast and non-time sensitive communication, but simply cannot facilitate the kind of broadband switched services required for the peer-to-peer networking applications that individuals and business have come to rely upon. Nevertheless, for rural and other physically isolated communities, satellite broadband is the only option. Hence, regionally-focused GSO providers such as Eutelsat's Tooway in Europe, IPSTAR in the Asia-Pacific, and North American providers Hughes, Telesat Canada, and ViaSat have all optimized their capabilities by migrating to the Ka Band and are able to deliver service with as little as a single satellite. Many of these networks having undergone recent upgrades boosting bandwidth to as high as 130 gbps, and GSO satellites will certainly remain significant components in the Internet ecosystem [28].

In the late 1990s and early 2000s, there was considerably less competition in delivering internet communication through satellites in low Earth orbit, in light of the many technological challenges associated with ensuring continuity of connections, quality of service (QoS), and achieving operational intersatellite links. However, there was great enthusiasm in the telecommunications community for leveraging LEO satellites, with the development of the Iridium and Globalstar communication systems representing over $8 billion in investment and 1000 new patents at the close of the 20th century [29].

The two early experiments in global broadband LEO networks, Teledesic and Skybridge each approached the myriad technical challenges from different angles. Teledesic, a $9Billion joint venture between Microsoft's Bill Gates, cellular technology magnate Craig McCaw, Boeing, and Motorola aimed to provide an affordable worldwide core and access broadband network with a constellation of 288[2] LEO satellites operating in the Ka Band with an equivalency to optical fiber networks. Their only competitor, Skybridge was a venture between the world's leading submarine cable manufacturer Alcatel, Loral, and Qualcomm, which utilized an 80 satellite constellation to provide a broadband access network, operating in the Ku Band frequency at a cost of $4.2Billion, and relying largely on terrestrial gateway networks for switching and routing procedures [30]. For both systems, the estimation of available resources and the dynamic allocation of resources, or bandwidth was a key technical challenge to be overcome, as the systems aimed to accommodate traffic from as many as 20 million simultaneous users [31]. However, just as Iridium and Globalstar were unable to develop a sufficient customer base amidst a burgeoning terrestrial cellular market and were forced into bankruptcy and restructuring, Teledesic and Skybridge were unable to realize a cost-effective LEO alternative to expanding ADSL and cable networks, and both folded before the systems were put in place [32].

Although the unfulfilled promise of global satellite broadband networks such as Teledesic and Skybridge stymied further investment in non-GSO platforms during the early 21st century, continued research and advances in technology have illuminated alternate pathways for extraterrestrial networking while leaving the door open for LEO platforms. With an increased focus on deep space travel, delay-tolerant networks (DTN) and bundling protocols were developed to mitigate the latency issues of disparate networks such as GSO satellites, with the ultimate intent of achieving a future interplanetary internet [33]. Historically, the U.S. National Aeronautics and Space Administration (NASA) developed unique communications systems for each of its missions, a tendency which became clearly untenable as NASA's systems were incompatible with the many emerging international, military, and commercial satellite capabilities [34]. As a result, the Consultative Committee for Space Data Systems was formed to promote shared infrastructure and develop universal Space Communications Protocol Standards (SCPS) based on Internet protocols and modified for the unique operating conditions of outer space [35]. International collaboration on the development and refinement of the SCPS has helped to facilitate rapid advances in technology that have significantly improved satellite capability. In particular, the development of onboard processing (OBP), switching (OBS), and routing (OBR), performance enhancing proxies (PEP) boosting the TCP slow start algorithm, and the integration of asynchronous transfer mode (ATM) protocols have led to improved intersatellite links, signal regeneration, error correction, and dynamic data routing [24]. Such advancements have led to renewed interest in LEO constellations for Internet connectivity, as depicted below in Figure 4.

Figure 4. Stylized Depiction of Global LEO Constellations

In light of this recent progress, a new set of actors is taking the stage to deliver global broadband Internet through satellites. First is O3B (The Other 3 Billion) Networks, which began emplacing its initial constellation of eight MEO satellites in 2010 and became commercially operational at the end of 2014. It will ultimately scale up to a network of 16 satellites, each of which deliver 10 spot beams in the Ka Band, building on the design and operating principles developed for Teledesic and Skybridge and capable of delivering a total capacity of over 160 gbps of bandwidth [36]. Orbiting at 8,000 km above the Earth, signals from O3B's satellites have a roundtrip time of 150 milliseconds, still considerably less than that of GSO counterparts. With

[2] Teledesic's system designers originally envisioned a constellation of 840 satellites, which was reduced to 288 as designs progressed, and finally down to 30 satellites before the project was cancelled.

 ISBN: 978-1-61208-423-7

significant financial support from investors such as the Virgin Group and Qualcomm, O3B's creators are in the process of launching OneWeb, a 648-LEO satellite constellation with aspirations that are reminiscent of Teledesic's vision to construct a satellite constellation to match the fixed Internet in terms of coverage, capacity, and reliability by 2019 [37]. While OneWeb is in the early stages of development, a critical step forward has been its ability to secure Ku Band wireless spectrum rights from the ITU [38], through which the network will interface with mobile broadband network operators and individuals with OneWeb receivers. Yet, constructing a satellite constellation is resource-intensive as a single launch costs $300 million, let alone the cost of building hundreds of satellites. Despite the resource-intensive prospect, Space Exploration Technologies (SpaceX) also recently announced plans to field an LEO satellite constellation to provide global broadband Internet access [39]. With over $1 Billion raised thus far from investors such as Google, Fidelity Investments, and Founders Fund, SpaceX plans to construct a network of roughly 4,000 satellites beginning in the next 5 years [40].

Whereas these recent developments in expanding Internet accessibility through large scale satellite networks are certainly promising, the success of such endeavors is challenged by significant technical obstacles. For systems such as OneWeb, the saturation of the Ku Band and need to develop viable spectrum sharing mechanisms among various satellite networks remains an open area of investigation [41]. In addition, although the vulnerability of both the Ku and Ka Bands to signal attenuation in moist atmospheric conditions has been well known for some time [42], effective rain fade countermeasures have yet to be developed. The details of SpaceX's LEO network design remain unclear, but one immediately apparent concern is that of how a constellation of 4,000 assets can operate sustainably and reliably in an environment characterized by significant amounts of residual orbiting debris [43], as little as 1 cm's worth of which is capable of inflicting significant damage to small assets. At the same time, the affordability and quick turn-around time in production make micro-satellites an attractive option for populating fleets of orbiting devices that number near the thousands [44].

IV. OUTSIDE OF THE BOX: ALTERNATE METHODS FOR EXPANDING INTERNET ACCESSIBILITY

Just as the open and collaborative spirit of the Internet Engineering Task Force has led to the achievement of a global information infrastructure through the request for comments (RFC) process and the dynamic exchange of ideas [45], the effort to expand Internet accessibility can benefit from collaboration between actors. However, the potential benefits of collaboration and rough consensus have to be balanced against the economic imperative of profit making for the investors involved. In that vein, it should come as no great surprise that some of the most innovative research with regard to increasing global Internet access are being pursued separately in parallel by two of the most influential and revenue-generating forces on the World Wide Web; the search engine Google and the social networking application Facebook.

In addition to its investment in a future satellite network, Google has pursued several equally ambitious avenues for improving and expanding Internet accessibility. First, it is constructing its own fiber-optic cable network in the United States, delivering broadband service speeds that drastically outperform existing Internet Service Providers (ISP) [46]. Second, it has conducted pilot programs to deliver 4G LTE wireless broadband access to remote areas via high-altitude superpressure envelope balloons equipped with a payload of solar panels, a battery, flight computer, altitude control system, radio, and antennae. Dubbed Project Loon, the mesh network of balloons receive and relay mobile broadband signals from telecommunications operators' cell towers and down to anyone in range with a 4G-capable device while cruising at an altitude of just over 20,000km, staying aloft for up to 100 days in the stratosphere, where temperatures can be as low as -117º Fahrenheit (-83º Celsius) [47]. Although Google has collaborated with manufacturers to produce special-purpose lithium-ion batteries for the balloons' electronic systems, such cold temperatures remain a particular challenge for sustaining power, in addition to keeping the balloons aloft for longer durations [48]. Project Loon builds on earlier concepts such as the Israeli ConSolar/Rotostar system and Sky Station stratospheric telecommunications platform, developed privately by former U.S. Secretary of State Alexander Haig in cooperation with NASA's Jet Propulsion Laboratory in the late 1990s [49]. Whereas Sky Station was unable to get off the ground, Project Loon has completed a pilot experiment with 30 balloons communicating to specialized ground antennae (pictured below in Figure 5) in 2013 over New Zealand's South Island and more recently in Brazil and Nevada [50]. Although Google is continuing to expand the project, aiming to have 100 balloons aloft by the end of 2015, a global network will have to overcome the aforementioned technical challenges as well as negotiate international over-flight rights and spectrum usage licenses in each country the system crosses over, political obstacles which ultimately proved insurmountable for Sky Station and similar platforms [51].

Figure 5. Project Loon Signal Receiver

While Google moves forward with its mesh balloon network, Facebook has announced a similar pursuit of expanded broadband access. In addition to GSO and LEO satellite options, it is exploring ways to leverage high-altitude solar powered unmanned aircraft systems (UAS) and free space optical (FSO) communication to deliver the Internet to isolated populations [52]. UAS such as Helios, Pathfinder, and Proteus have demonstrated the ability for aircraft to remain aloft for extended periods and to serve as viable high altitude platform stations (HAPS) for telecommunications, however signal attenuation and spectrum allocation have been among the greatest challenges to implementation [53]. The ITU has allocated a small section of the Ka Band for HAPS broadband services, however rain fade remains a persistent problem, particularly for systems deployed over isolated tropical areas that experience significant precipitation [54]. At the same time, FSO communication systems that utilize lasers as an alternative signal medium are attractive for their high bandwidth capacity, license-free usage, and immunity to electromagnetic interference, however they cannot effectively transmit through clouds and remain subject to atmospheric absorption, attenuation, and backscatter [55]. Although methods to mitigate atmospheric effects that include photon counting receivers and coherent reception techniques have been identified, for the moment they remain sufficiently resource-intensive so as to preclude widespread commercial application [56]. Indeed, HAPS such as Project Loon and those envisioned by Facebook hold great promise for increasing global Internet accessibility; however significant technical and political hurdles remain to be overcome.

V. GENERATIVITY AND NET NEUTRALITY

As the Internet's vast communicative power lies in the generativity of its many diverse data pathways and content contributors, it is appropriate that a multitude of gateways exist for its accessibility. However, the openness of these gateways is an important consideration that impacts both the network's reach and resilience. In order to maximize the network's reach, barriers to entry such as identity verification are minimized. Meanwhile, to ensure the resilience of the network, access control and security protocols are needed to prevent malicious activity.

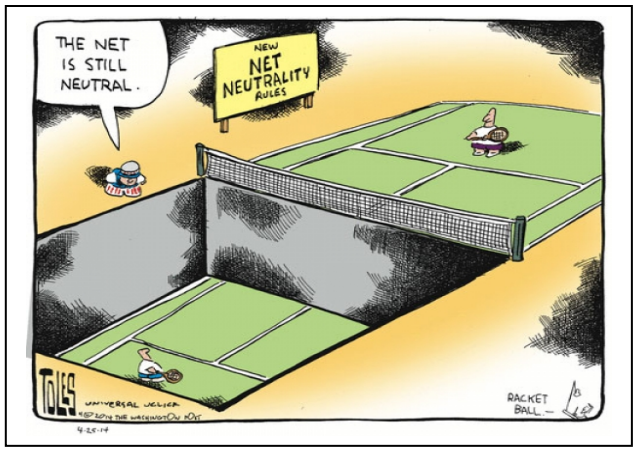


Figure 6. Satirical Commentary; Toles, The Washington Post

While any of the aforementioned methods can help increase connectivity in physically isolated areas, a separate question remains as to providing Internet access for underprivileged communities who simply cannot afford to pay. If access to information is truly a human right, then perhaps the economic principles governing the Internet's for-profit development warrant revision in light of its role in global information dissemination, and vehicles for delivering free Internet to the underprivileged bear increased consideration [57].

The business model of capitalizing on free labor in the form of user-generated content has allowed a small number of individuals to amass vast fortunes [58]. However, is there a fundamental conflict in the Internet being both a gold mine for industrious prospectors and the primary delivery mechanism for a basic human right? Whereas initiatives such as Facebook's Internet.org appear helpful for bringing the Internet to the world's underprivileged populations through negotiated access with certain regional ISPs, these efforts are facing stiff resistance for the role they play in privileging certain web sites and services [59]. Indeed, the prerogative to maintain net neutrality is in question if the aforementioned efforts to expand Internet accessibility only translate into expanding access to circumscribed pieces of the Internet. Amidst the debate over net neutrality, the Federal Communications Commission (FCC)'s Open Internet Order categorizes broadband as a Title II telecommunications service and is therefore subject to the terms of the 1934 Communications Act [60]. As a result, blocking, throttling, or paid prioritization of any content provider is illegal for the time being. However, efforts to circumvent the neutrality of the Internet ecosystem by providing free access to certain limited content undermines the technology's main function of facilitating end-to-end communication. Expanding Internet accessibility to the as-yet connected portion of the world's population will undermine the system's overall resilience if it comes at the cost of creating a two-tiered Internet for those who can afford to buy access to all information and others who can only afford free access to some information.

VI. CONCLUSION

Although alternative means for expanding global Internet accessibility are in order, current endeavors to do so remain challenged by the same limitations that have hindered conventional Internet service delivery methods in the past. High infrastructural investment costs, signal attenuation, and efficient power generation are among the most notable obstacles to be overcome in making the Internet a truly global resource. Ambitious developments such as Project Loon, OneWeb, and Green Networking [61] all have the potential to enhance the resilience and reach of the Internet, provided they can surmount numerous remaining obstacles.

 ISBN: 978-1-61208-423-7

ACKNOWLEDGMENT

The authors would like to thank the Cyber Futures Center, an initiative of the Sensemaking-U.S. Pacific Command Fellowship, and the Dr. Steve Chan Center for Sensemaking — one of the centers of the Asia-Pacific Institute for Resilience and Sustainability (AIRS), which is jointly anchored at Swansea University's Network Science Research Center and Hawaii Pacific University — for the opportunity to study the challenges facing Hawaii and other archipelagos, and to contribute towards the various Public Private Partnership Initiatives aimed at developing solutions to overcome those challenges.

REFERENCES

[1] C. Aggarwal, N. Ashish, and A. Sheth, "The Internet of Things: A Survey from the Data-Centric Perspective," in Managing and Mining Sensor Data, C. C. Aggarwal, Ed., ed: Springer US, 2013, pp. 383-428.

[2] ITU, "Measuring the Information Society Report 2014," International Telecommunications Union, Geneva, Switzerland2014.

[3] V. G. Cerf, "Internet access is not a human right," New York Times, vol. 4, p. 55, 2012.

[4] M. McDonagh, "The Right to Information in International Human Rights Law," Human Rights Law Review, vol. 13, pp. 25-55, March 1, 2013 2013.

[5] H. Chourabi, N. Taewoo, S. Walker, J. R. Gil-Garcia, S. Mellouli, K. Nahon, et al., "Understanding Smart Cities: An Integrative Framework," in System Science (HICSS), 2012 45th Hawaii International Conference on, 2012, pp. 2289-2297.

[6] G. Bugliarello, "Urban security in the United States: An overview," Technology in Society, vol. 27, pp. 287-293, 2005.

[7] B. M. Leiner, V. G. Cerf, D. D. Clark, R. E. Kahn, L. Kleinrock, D. C. Lynch, et al., "A brief history of the internet," SIGCOMM Comput. Commun. Rev., vol. 39, pp. 22-31, 2009.

[8] Toward a National Research Network. Washington, DC: The National Academies Press, 1988.

[9] R. Yuan and W. Gong, "On the complexity and manageability of Internet infrastructure," Frontiers of Electrical and Electronic Engineering in China, vol. 6, pp. 424-428, 2011/09/01 2011.

[10] D. M. West, "Digital divide: Improving Internet access in the developing world through affordable services and diverse content," The Brookings Institution, Washington, D.C.2015.

[11] Terabit, "Submarine Telecoms Industry Report," Submarine Telecoms Forum, Cambridge, MA2014.

[12] J. K. Crain. (2012) Assessing Resilience in the Global Undersea Cable Infrastructure. *Naval Postgraduate School*. Available: http://www.dtic.mil/cgibin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA562772, accessed June 10, 2015

[13] N. Abramson, "Development of the ALOHANET," Information Theory, IEEE Transactions on, vol. 31, pp. 119-123, 1985.

[14] M. Gerla and L. Kleinrock, "Vehicular networks and the future of the mobile internet," Computer Networks, vol. 55, pp. 457-469, 2011.

[15] B. Commission, "The State of Broadband 2014: broadband for all," International Telecommunications Union, Geneva, Switzerland2014.

[16] K. Hinton, J. Baliga, M. Z. Feng, R. W. A. Ayre, and R. Tucker, "Power consumption and energy efficiency in the internet," Network, IEEE, vol. 25, pp. 6-12, 2011.

[17] D. Baltrunas, A. Elmokashfi, and A. Kvalbein, "Measuring the Reliability of Mobile Broadband Networks," presented at the Proceedings of the 2014 Conference on Internet Measurement Conference, Vancouver, BC, Canada, 2014.

[18] R. N. Clarke, "Expanding mobile wireless capacity: The challenges presented by technology and economics," Telecommunications Policy, vol. 38, pp. 693-708, 2014.

[19] A. Aijaz, H. Aghvami, and M. Amani, "A survey on mobile data offloading: technical and business perspectives," Wireless Communications, IEEE, vol. 20, pp. 104-112, 2013.

[20] CISCO, "Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019," San Jose, CA2015.

[21] S.-I. Chakchai, R. Jain, and A. K. Tamimi, "Scheduling in IEEE 802.16e mobile WiMAX networks: key issues and a survey," Selected Areas in Communications, IEEE Journal on, vol. 27, pp. 156-171, 2009.

[22] A. C. Clarke, "Extraterrestrial relays," Wireless world, vol. 51, pp. 305-308, 1945.

[23] H. D. Clausen, L. Linder, and B. Collini-Nocker, "Internet over direct broadcast satellites," Communications Magazine, IEEE, vol. 37, pp. 146-151, 1999.

[24] H. Yurong and V. O. K. Li, "Satellite-based Internet: a tutorial," Communications Magazine, IEEE, vol. 39, pp. 154-162, 2001.

[25] C. Metz, "IP-over-satellite: Internet connectivity blasts off," Internet Computing, IEEE, vol. 4, pp. 84-89, 2000.

[26] A. Botta and A. Pescape, "On the performance of new generation satellite broadband internet services," Communications Magazine, IEEE, vol. 52, pp. 202-209, 2014.

[27] M. Williamson, "Can satellites unblock the Internet?," IEE Review, vol. 45, pp. 107-111, 1999.

[28] J. P. Conti, "Satellites bring broadband home," Engineering & Technology, vol. 5, pp. 60-63, 2010.

[29] J. Beesemyer, R. Adam, and R. Donna, "Case Studies of Historical Epoch Shifts: Impacts on Space Systems and their Responses," in AIAA SPACE 2012 Conference & Exposition, ed: American Institute of Aeronautics and Astronautics, 2012.

[30] D. J. Bem, T. W. Wieckowski, and R. J. Zielinski, "Broadband satellite systems," Communications Surveys & Tutorials, IEEE, vol. 3, pp. 2-15, 2000.

[31] J. Farserotu and R. Prasad, "A survey of future broadband multimedia satellite systems, issues and trends," Communications Magazine, IEEE, vol. 38, pp. 128-133, 2000.

[32] E. W. Ashford, "Non-Geo systems—where have all the satellites gone?," Acta Astronautica, vol. 55, pp. 649-657, 2004.

[33] S. Burleigh, A. Hooke, L. Torgerson, K. Fall, V. Cerf, B. Durst, et al., "Delay-tolerant networking: an approach to interplanetary Internet," Communications Magazine, IEEE, vol. 41, pp. 128-136, 2003.

[34] K. Bhasin and J. L. Hayden, "Space Internet architectures and technologies for NASA enterprises," International Journal of Satellite Communications, vol. 20, pp. 311-332, 2002.

[35] A. J. Hooke, "Towards an interplanetary internet: A proposed strategy for standardization," in Proceedings of Space Operations Conference and World Space Congress, 2002.

[36] M. Williamson, "Connecting the other three billion," Engineering & Technology, vol. 4, pp. 70-73, 2009.

[37] A. Vance. (2015, January 22, 2015) The New Space Race: One man's mission to build a galactic internet. Bloomberg Business. Available: http://www.bloomberg.com/news/features/2015-01-22/the-new-space-race-one-man-s-mission-to-build-a-galactic-internet-i58i2dp6 accessed April 18, 2015

[38] P. B. d. Selding. (2015, January 23, 2015) SpaceX-Google Matchup Sets Up Satellite Internet Scramble Space News. Available: http://spacenews.com/spacex-google-matchup-sets-up-satellite-internet-scramble/ accessed April 18, 2015

[39] R. Winkler, E. Rusli, and A. Pasztor. (2015, 2015 Jan 20) SpaceX Gets $1 Billion From Google, Fidelity; Investment in Musk's Company Aimed at Helping to Provide Internet Access Via Satellites. Wall Street Journal (Online). Available: http://www.wsj.com/articles/spacex-gets-1-billion-from-google-fidelity-1421795584 accessed April 18, 2015

 ISBN: 978-1-61208-423-7

[40] J. Hsu. (2015, January 27, 2015) SpaceX Raises $1 Billion from Google and Fidelity for Satellite Internet Project. IEEE Spectrum. Available: http://spectrum.ieee.org/tech-talk/aerospace/satellites/spacex-raises-1-billion-from-google-fidelity-for-satellite-internet-project accessed April 18, 2015

[41] T. D. Cola, D. Tarchi, and A. Vanelli-Coralli, "Future trends in broadband satellite communications: information centric networks and enabling technologies," International Journal of Satellite Communications and Networking, 2015.

[42] A. Safaai-Jazi, H. Ajaz, and W. L. Stutzman, "Empirical models for rain fade time on Ku- and Ka-band satellite links," Antennas and Propagation, IEEE Transactions on, vol. 43, pp. 1411-1415, 1995.

[43] D. J. Kessler, R. C. Reynolds, and P. D. Anz-Meador, "Orbital debris environment for spacecraft designed to operate in low Earth orbit," National Aeronautics and Space Administration, Houston, TX1989.

[44] C. d. Aenlle. (2001, June 19) U.K. Firm Finds Niche in 'Discount' Satellites. *The New York Times*. Available: http://www.nytimes.com/2001/06/19/news/19iht-rsat.html, accessed June 9, 2015

[45] S. D. Crocker, "How the Internet got its rules," The New York Times, p. 29, 2009.

[46] J. Davidson. (2015, April 14, 2015) Google Fiber has Internet providers scrambling to improve their service. Fortune. Available: https://fortune.com/2015/04/14/google-fiber-has-internet-providers-scrambling-to-improve-their-service/ accessed April 18, 2015

[47] N. Davidson. (2015, April 14, 2015) How it Works: Project Loon's Global Internet. Popular Science. Available: http://www.popsci.com/how-it-works-project-loon-global-internet accessed April 18, 2015

[48] A. Barr. (2015, April 10) Google Gets Into Battery Arms Race. Wall Street Journal. Available: http://www.wsj.com/articles/google-gets-into-battery-arms-race-1428694613 accessed April 18, 2015

[49] L. Yee-Chun and Y. Huanchun, "Sky Station Stratospheric Telecommunications System, a high speed low latency switched wireless network," in 17th AIAA International Communications Satellite Systems Conference and Exhibit, ed: American Institute of Aeronautics and Astronautics, 1998.

[50] S. Katikala, "Google Project Loon," Rivier Academic Journal, vol. 10, Fall 2014 2014.

[51] K. Maney. (2014, April 19) Dark Side of the Loon. Newsweek. Available: http://www.newsweek.com/2014/04/18/dark-side-loon-248107.html accessed April 18, 2015

[52] B. Fung, "Lasers, satellites and drones: How Facebook plans to deliver Internet to the developing world," in Washington Post, ed. Washington, D.C., 2014.

[53] A. Widiawan and R. Tafazolli, "High Altitude Platform Station (HAPS): A Review of New Infrastructure Development for Future Wireless Communications," Wireless Personal Communications, vol. 42, pp. 387-404, 2007/08/01 2007.

[54] A. Mohammed, A. Mehmood, F. N. Pavlidou, and M. Mohorcic, "The Role of High-Altitude Platforms (HAPs) in the Global Wireless Connectivity," Proceedings of the IEEE, vol. 99, pp. 1939-1953, 2011.

[55] D. Killinger, "Free Space Optics for Laser Communication Through the Air," Optics and Photonics News, vol. 13, pp. 36-42, 2002/10/01 2002.

[56] V. W. S. Chan, "Free-Space Optical Communications," Lightwave Technology, Journal of, vol. 24, pp. 4750-4762, 2006.

[57] A. Sathiaseelan and J. Crowcroft, "The free Internet: a distant mirage or near reality?," University of Cambridge, Technical Report, vol. 814, 2012.

[58] M. Finn, A. Mathew, S. Yeo, L. Irani, S. Kelkar, A. Chia, et al., "(Invisible) Internet infrastructure labor," Selected Papers of Internet Research, vol. 3, 2013.

[59] S. Rai. (2015, April 16, 2015) Facebook's Internet.org Faces Heat In India Over Net Neutrality. Forbes. Available: http://www.forbes.com/sites/saritharai/2015/04/16/facebooks-internet-org-faces-heat-in-india-over-net-neutrality/ accessed April 18, 2015

[60] J. Pil Choi and B.-C. Kim, "Net neutrality and investment incentives," The RAND Journal of Economics, vol. 41, pp. 446-471, 2010.

[61] A. P. Bianzino, C. Chaudet, D. Rossi, and J. Rougier, "A Survey of Green Networking Research," Communications Surveys & Tutorials, IEEE, vol. 14, pp. 3-20, 2012.

 ISBN: 978-1-61208-423-7

Automating Clustering Analysis of Ivory Coast Mobile Phone Data

Deriving Decision Support Models for Community Detection and Sensemaking

Thomas J. Klemas
MIT Lincoln Laboratory
tklemas@alum.mit.edu

Steve Chan
Network Science Research Centre
Swansea University
stevechan@post.harvard.edu

Abstract—**Sensemaking involves numerous levels of processing and logic in order to achieve automated decision support. Many of these concepts derive from the realm of pattern recognition. The data under consideration frequently is observed in a noisy environment and so one of the first steps involves preprocessing the data to suppress noise and isolate the data signal. Patterns within the data are often used to improve signal detection and aid identification of the data in the quest to produce actionable information. A critical step of making sense from raw or partially processed data and other aspects of decision support is to organize information, which frequently involves grouping, partitioning, or clustering objects. However, there is typically an assumption that structure exists within the data, and the number of clusters is a required parameter for many of the clustering algorithms. A common approach to determine the best number of clusters is to iterate across a set of potential values the for number of clusters and evaluate the quality of the resulting clusters using some metric. In this paper, we present an automated approach to detect structure and improve automation of clustering algorithm parameters. We apply our approach to analyze a complex, dynamic multiple edge set network that was used to model call data from the Ivory Coast compiled from France Telecom/Orange anonymized call records over a 5 month period.**

Keywords - Sensemaking; adaptive clustering; spectral clustering; network theory; silhouette; k-means; unsupervised; partitioning; proximity measure; similarity measure; decision support; iterative; randomized singular value decomposition.

I. INTRODUCTION

When data sets are extremely large, very complex, or the data is changing rapidly, analysis requirements reach a level beyond which humans are unable to consider the full scope of the data and lack the capacity to keep up, derive insights, and make decisions. In today's world of increasingly smart and interconnected systems, more and more sensors are deployed in civilian, medical, industrial, and military systems and these systems are frequently networked in some manner to allow programming (automation), monitoring, and remote control, to facilitate software updates, to enable interactivity, and related objectives. Accompanying the rapid rise of networked sensors is a flood of available new data. However, to maximize the value of this data it is critical to attach appropriate labels, meta data, and links that enable combining and synchronization of this data with other suitable data for analysis.

The ultimate objective for Sensemaking technologies is to make sense of the raw data. Automated decision support and Sensemaking tools apply machine learning, pattern recognition, expert logic, and other algorithms in order to detect and identify patterns in the mass of data that contain information that supports and enables decisions. Typically, there are many steps required before one is able discern actionable information from the raw data. These steps may include numerous preprocessing routines that may eliminate data outliers that have the potential to distort decision making algorithms, normalize the data to improve sensitivity, inserting values for missing data items, and similar manipulations to "clean up" the data in preparation for subsequent mainstream processing stages. Depending on the specific methods to be used, feature vectors will be computed from the raw data and metrics will be calculated from the feature vectors to support various classification decisions.

Our objective is targeted to detect and identify important but non-evident structural groupings, develop insights based on the structure to resolve community clusters. In this paper, as in the previous paper [5], we focus on pattern recognition algorithms that provide a mechanism for grouping objects detected in the data channel based on features vectors or measurable quantities of interest that are selected to help distinguish different objects. When training data is available, the grouping of objects is often called partitioning. When no training data is available, grouping of objects is termed clustering. Classical methods for clustering include *k*-means, spectral, Kerninghan-Lin, and other algorithms. A variety of proximity measures can be used to determine whether data points and corresponding objects share either similarity or dissimilarity, based on feature vector values in 1 or more dimensions. Examples include Euclidean distance, silhouette values, Pearson coefficients, Saltine's cosines, or other proximity measures [1]. In particular, we will examine unsupervised clustering techniques that group data without the benefit of training data containing truth information.

This research will present and explore performance of an automated approach to detect if structure is present in the data and also to select a number of clusters and cluster

 ISBN: 978-1-61208-423-7

objects from new, unknown data sets. If no data is present within a data set the various clustering algorithms can produce unusual, potentially nonsensical results. We adapted methods, based on spectral decomposition, to achieve clustering in a multiple edge set network that we generate to model the Ivory Coast France Telecom/Orange call records.

In our previous work [5], we observed that the evolutionary approach that we adopted to model the drift of parameters of the associated proximity measures required either careful selection of parameters [2] [3] [6] or iterative solves to choose a parameter value, such as number of clusters. Related to this issue, it is important to determine that structure exists before applying clustering algorithms or risk nonsensical results when attempting to interpret the results of clustering analysis. Additionally, we observed that solves were computationally intensive, so in this work we explore an approach to automate detect structure and improve parameter selection. Also in this paper, although it is obviously not the primary focus of this work, we illustrate the steps involved to apply randomized [4] hybrid methods to accelerate clustering algorithms in our problem space. This sort of computational efficiency becomes increasingly important especially when contemplating community detection in much larger countries or regions of the world.

The remainder of this manuscript is arranged as described herein. Section II describes the technical details of our clustering algorithms and how they accomplish analysis of a multiple edge set network. Section III provides a brief description of the data set and also outlines how the key data elements are aligned as inputs to the analysis. Section IV described the performance and provides results of applying our automated clustering approach to the multiple edge set network data modeled from the Ivory Coast France Telecom/Orange call records. Section V offers our conclusions. Finally, the acknowledgment and reference sections complete the manuscript.

II. TECHNICAL DETAILS

We start by reviewing the fundamental clustering methods and the notations that we will be using throughout the manuscript. First of all, we will model sub-prefectures in which our callers access cell towers to make and receive calls as nodes and the call records between 2 sub-prefectures as result of a call between 2 callers (one in each sub-prefecture), as an edge. In our case, we also were able to construct travelers from the call data as we observed callers that switched cell towers and even sub-prefectures as they traveled by car, bus, train, or airplane. Thus, our nodes have traveler edge connections between them as well, as another type of edge set interconnecting our social network.

$$G = (V, E_1, E_2) \quad (1)$$

In the graph G, the V nodes represent sub-prefectures, the E_1 edges represent calls between sub-prefectures, and the E_2 edges represent travelers between sub-prefectures Furthermore, for completeness, the additional induced graph relating the various cell towers and sub-prefecture centers by geographical distance should really be incorporated into this graph as well, but for now we ignore this layer. For the sake of simplicity, we will represent the travel with an undirected edges connecting the graph model. Since the callers use cell towers that are distributed geographically within sub-prefectures of the Ivory Coast, our algorithms incorporate a mapping layer to translate between cell towers and sub-prefectures As our goal is to detect hidden structure withing the call data that may correspond to communities, our notation provides corresponding terms. The term S_i defines a community or cluster of nodes, in this case sub-prefectures, that are disjoint to all other communities. Thus, a vertex can exist in only one community.

$$V = \bigcup S_i, \ \forall\, i,j, i \neq j, \ S_i \cap S_{j=\varnothing} \quad (2)$$

Following well established methods [1], by selecting a feature vector, f_a, in this case the accumulated calls between sub-prefecture a and every other sub-prefecture, and choosing a proximity measure, in this case the euclidean distance between two feature vectors, f_a and f_b, we can utilize this dissimilarity metric to compare two sub-prefectures on the basis of the associated call records. Furthermore, extending this concept to the entire set of sub-prefectures, we can applying a variety of pattern recognition and network science techniques to attempt to cluster the sub-prefectures based on the call records, as well. In this research, we employed several clustering approaches, derived from k-means, spectral decomposition, and aggregation, and developed modifications aimed to enable improved automation, improve computational efficiency, improved ease of implementation, and facilitate comparison study of clustering performance,

First, we briefly review these approaches to augment our notation prior to modification and enhancement of the algorithms. The classical k-means algorithm [1] requires a parameter, k, which specifies the number of clusters into which the objects, in our case sub-prefectures, should be grouped. Then the algorithm, randomly selects k centroids in the space in which feature vectors reside. The objects are then clustered into the cluster with the nearest centroid using the similarity measure and centroids are recomputed. This process continues iteratively until the cluster centroids converge or cease to change.

To effectively utilize k-means and the other algorithms to cluster the caller records, based on selected features, as an precursor aid to facilitate community detection, it is typical to iteratively solve for a new clustering of the system and determine the best number of clusters based on a suitable metric. In our previous paper, we adopted silhouette values [8] as such a metric to facilitate choice of the number of

 ISBN: 978-1-61208-423-7

clusters. The silhouette value concept was constructed to characterize the degree of community structure that is present in a clustering induced from a set of interrelated objects, such as the sub-prefectures of the Ivory Coast. Briefly reviewing the mechanics of this approach, the silhouette function is defined as:

$$silhouette(i) = (b(i) - a(i))/max(a(i),b(i)) \quad (3)$$

the value *a(i)* represents the intra-cluster dissimilarity of sub-prefecture *i* or, in other words, the mean value of a chosen dissimilarity measure for the sub-prefecture *i* with respect to the other sub-prefectures that are members of the same cluster. The value *b(i)* represents the smallest average dissimilarity between sub-prefecture *i* and the clusters of which it is not a member. For this research, we adopt euclidean distance between two feature vectors as the proximity metric, in this case a measure of dissimilarity between the two nodes. A silhouette value is assigned to the entire clustering, as well,

$$silhouette(k) = mean_i(silhouette(i)) \quad (4)$$

which is simply the mean value of the silhouette values of each node or sub-prefecture Using these definitions, the silhouette values will vary between 1, indicating high degree of community structure and -1, which suggests the absence of community structure.

Next, we describe how clustering can be accomplished using classical spectral methods for graph decomposition. The adjacency matrix, **A**, indicates a measure of the amount and duration of calls that were exchanged between each pair of particular sub-prefectures, forming connections between the corresponding nodes in the graph where inter-sub-prefecture calling was recorded in the data set. If the call pairing vectors that arise from the columns of the adjacency matrix are compared with a "proximity" measure (in particular a similarity measure) then it is possible to determine the extent to which nodes share common connectivity patterns. Thus, illustrating this concept, the similarity matrix, **W**, which is the target for the spectral decomposition, is computed from the adjacency matrix simply as the inner product of the column vectors forming the adjacency matrix,

$$W_{ij} = a_i^T a_j \quad (5)$$

Thus, the entries indicate similarity between columns of the adjacency matrix and highlights node pairs with similar connectivity patterns. Traditional spectral clustering methods involves computing the singular value decomposition (SVD) of a matrix related to the similarity matrix, such as the Laplacian matrix,

$$L = D - A \quad (6)$$

and the matrix is decomposed as described in equation (7).

$$L = U \Sigma V^T \quad (7)$$

There are schemes that determine clusters based on the eigenvectors corresponding to largest or smallest eigenvalues of the Laplacian matrix and related systems. However, we have selected a technique described in [7] in which the first K eigenvectors are retained $[U_1, U_2, \ldots, U_K]$ and then the rows of the retained K eigenvectors are partitioned in K clusters using the k-means algorithm. Since K is a parameter that needs to be selected, the silhouette values can be used to select a suitable value of K. However, this approach requires repeated iteration to compute the clusters that correspond to each value of K and determine the resulting silhouette values for each clustering. The clusterings are compared by silhouette values to determine the optimal number of clusters.

In this paper, we describe results from a method that we used to improve the overall approach automation by obviating the need for extensive clustering iterations to determine the best number of clusters, k. While many metrics can be used to select the best number of clusters, the silhouette metric [8], described previously, was used in this research, because its definition best captured our goals for clustering. As one might surmise, since we are using a hybrid spectral algorithm to accomplish clustering, the basis for selecting the k parameter derives from analysis of the larger singular values associated with the singular vectors that contribute the most towards defining the nodes sharing the most similar communication patterns in the graph subspace.

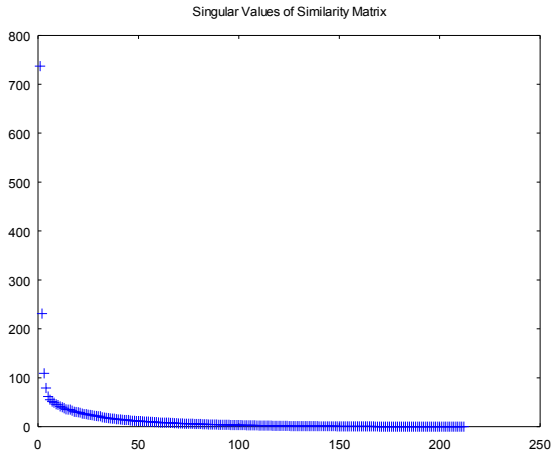


Figure 1: Singular Values for 255 by 255 Similarity Matrix indicate structure exists within sub-prefecture caller communication records.

We have developed a dual filtering analysis engine that attempts to detect the last large gap in the singular values and then determine the corresponding index of the singular

 ISBN: 978-1-61208-423-7

value preceding the identified gap, in order to select this index value as an upper bound on the k parameter, thereby significantly reducing the number of iterations to kth clustering and the few clusterings preceding it. Silhouette values or a similar metric can be used to choose the optimal value among the few that are explored. We note that this singular value gap analysis approach, based on the described dual-filtering method, requires a bounding parameter for the maximum number of clusters expected, but this does not necessitate additional clustering iterations and we expect that in many application, as in ours, this required parameter does not decrease utility of the method relative to the primary objectives. In testing, as we will show subsequently in the results section, the clustering decisions seem to compare favorably with the typical approach based on the iterative computation of cluster quality metrics, such as silhouette values.

Finally, to reduce the cost of computing the SVD, it is possible to use a randomized technique to decrease the overall computations. The technique we selected and implemented, [4], involves stimulation of the system input by multiplying the similarity matrix by a random matrix, Θ, with a reduced number of columns to elicit the corresponding output matrix, as described by $W*\Theta=Y$. Then, by computing an orthogonal decomposition or QR factorization of the output matrix, $Y=Q*R$, and multiplying the original matrix by Hermitian of Q, $C=Q^H*W$. Computing a reduced subset of the SVD of C $C\simeq\tilde{U}*\tilde{\Sigma}*\tilde{V}^H$, is much less expensive, and we can approximate A as $W\simeq Q*Q^H*W$. Thus, the singular value decomposition of A can also be approximated $W\simeq Q*\tilde{U}*\tilde{\Sigma}*\tilde{V}^H$ and $U\simeq Q*\tilde{U}$. The overall cost of this SVD is significantly less expensive, $O(N^2)$, and by using a lower cost SVD, spectral clustering methods are significantly more feasible for numerous applications with extremely large numbers of nodes.

An added improvement to spectral clustering methods, developed in [7], is to use K-means to cluster the row space of the singular vectors U. The theory behind this approach is developed nicely in that paper. We incorporated the same technique into our approach in order to compute spectral clustering, and we explore its performance relative to the traditional approaches.

III. Application Space

The cell tower call record data used in this research is described in the document prepared by Blondel and Esch et al [9]. The records consist of several categories of differing types. Throughout this research, we primarily analyzed the first and second subsets within the Data for Development (D4D) data record collection. In the first subset of call data records, cell tower to cell tower connections were accumulated for each hour, including frequency and duration attributes for each pairing. The subsequent subset of data was comprised of sub-prefecture indexing information for a random sample of 500,000 individual callers (as opposed to pairs) over a limited 2 week period. Finally, the third and last subset of the data contained call records for a smaller number of 50,000 individual records that endured over the entire 5 month D4D data collection period. Our research focused on the the second subset as the source of traveler information. Additional data files included information that specified center locations of sub-prefectures and locations for antennas. With this auxiliary data it is possible to map antennas to closest center locations for sub-prefectures, and, coupled with the file registering the locations of the sub-prefecture centers, the combination enables graphical result data plots revealing geographical trends.

IV. Results

To evaluate the efficacy of our algorithms we developed feature vectors that captured the content of sub-prefecture call records between February 7, 2012 and Feb 14, 2012. Our feature vectors comprised the cell tower connection data between callers that was recorded during this period and is mapped to corresponding sub-prefectures to which they belong geographically, of the 255 total sub-prefectures Thus, with this approach, our feature vector is 255 element vector, Υ_m, in which $\Upsilon_m(n)$ is the complete extent of calls between a pair of sub-prefectures indexed by m and n accumulated over the recording duration. Next, this sub-prefecture connection data is accumulated between each pair of sub-prefectures by stacking the columns Υ_m to generate an adjacency matrix with elements Υ_{mn}, representing the degree of connectivity between each of the associated pairs of sub-prefectures The adjacency matrix contains 255 rows and 255 columns.

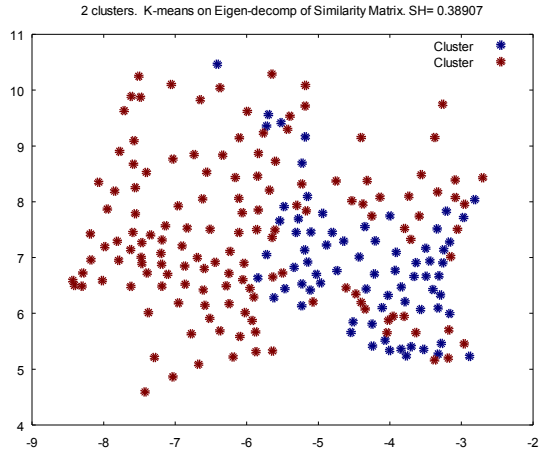


Figure 2: Clustering into 2 groups by applying K-means algorithm to similarity matrix

Then, in the final step of the setup stage, the similarity matrix is constructed from the adjacency matrix such that the value of each element is the selected proximity measure computed for the feature vectors of the sub-prefecture pair

 ISBN: 978-1-61208-423-7

corresponding to the indices of the associated element of the similarity matrix. Thus, in this fashion, an appropriate similarity matrix, also containing 255 rows and 255 columns, is generated for clustering analysis using the algorithms referenced and described earlier. We would like to note that while the population density within the Ivory Coast has geographical dependence we were not able to obtain sub-prefecture specific population data to utilize to normalize our feature space relative to population.

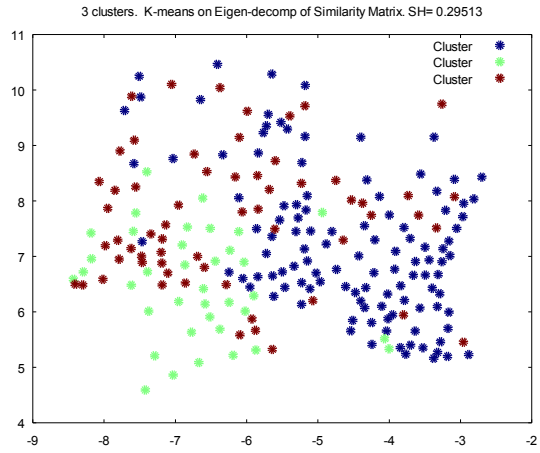


Figure 3: Clustering into 3 groups by applying K-means algorithm to similarity matrix. Note the decreasing silhouette score relative to the 2 group clustering.

As discussed earlier in section II, the spectral algorithm involves analysis of the singular values and vectors of the similarity matrix. In figure 1, we see the spectral value distribution for the similarity matrix constructed as described above. Note the steep drop-off in singular values

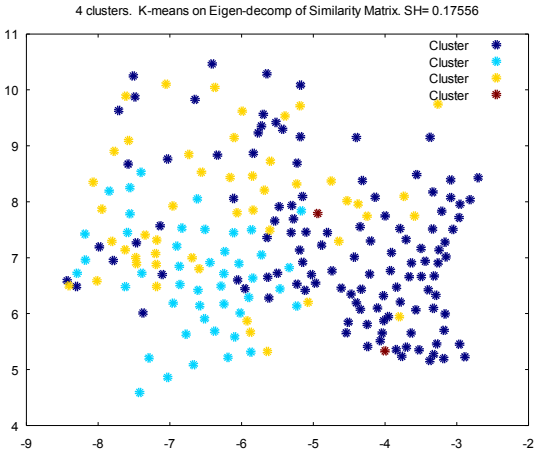


Figure 4: Clustering into 4 groups by applying K-means algorithm to similarity matrix. Note the decreasing silhouette score relative to the 2 and 3 group clusterings.

as well as the rapid decrease in gaps between subsequent singular values.

These observed phenomena are associated with the degree of structure within the data. Our dual-filtering algorithm analyzes the singular value gaps to determine suitable bounds on the number of clusters, k, and thereby significantly reduce the iterations required to generate strong clusterings comprised of the tightest possible clusters dependent on the selected clustering algorithm. Once we have selected the number of groups in which we wish to cluster the data, the algorithm for generating the clusters can be utilized in a straightforward manner as will be shown in our subsequent figures.

Figure 2 plots large points at the geographical centroids of each of the sub-prefectures that contain call data within the period of our study, so although there is some slight skewing of the relative scale of the axes, the points still align fairly well with a current political map of the Ivory Coast. The 2 colors, red and blue, in figure 2 are used to distinguish the 2 groups into which the sub-prefectures were clustered based on the call records. The silhouette score of approximately .40 indicates a significant degree of community structure. Silhouette values range between minus 1 and positive 1, whereby negative silhouette values indicate lack of structure and positive silhouette values indicate presence of structure within the analyzed data set.

As we see in figure 3, the 3-group clustering of sub-prefectures computed from the recorded caller data has a lower silhouette value than the previous figure depicting the 2-group clustering with parameter number of clusters selected automatically by the dual-filtering approach.

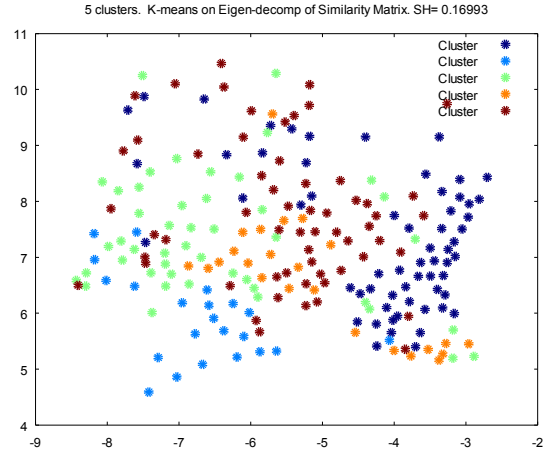


Figure 5: Clustering into 5 groups by applying K-means algorithm to similarity matrix. Note the decreasing silhouette score relative to the 2, 3, and 4 group clusterings.

This indicates that the structure present to the data is more indicative of a 2-group clustering than a 3-group clustering. In fact, the subsequent figures 4-6, each present an increasing number of groups, while the corresponding silhouette values decrease. These facts suggest that the structure present in the data best matches a 2-group clustering.

 ISBN: 978-1-61208-423-7

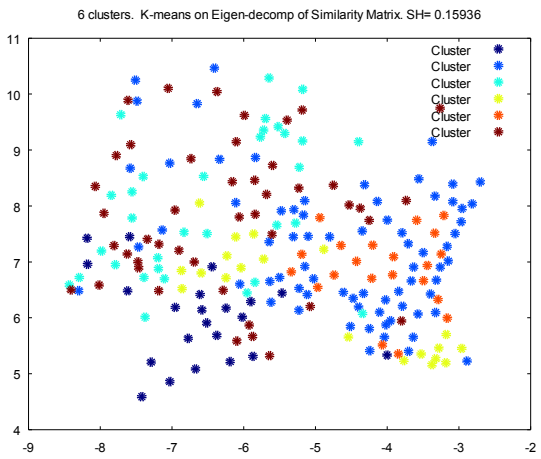


Figure 6: Clustering into 6 groups by applying K-means algorithm to similarity matrix. Note the decreasing silhouette score relative to the 2, 3, 4, and 5 group clusterings.

Other observations become apparent from examining the sequence of six figures generated from the results of the spectral decomposition based clustering algorithms. For one, the quality of the clusterings, measured in silhouette values, falls rapidly with the clusterings corresponding to 3 and then 4 clusters, but the rate of deterioration in quality slows dramatically between the 4th and 6th clusterings. This behavior matches the eigenvalue curve in figure 1. Also, we note that the region in and around Abidjan is differentiated from the other sub-prefectures in almost every one of the figures. This arises from the fact that Abidjan is the largest city by population within the Ivory Coast nation. Finally, every one of the clusterings (all of the figures) has positive silhouette values, revealing that the even with the differing number of groups in each clustering, there is evidence for corresponding structure hidden in the sub-prefecture call records, even if the degree of that particular structure (number of clusters) varies between clusterings.

V. CONCLUSION

In this research, we have explored the efficacy of using spectral information, revealed by singular value decomposition, to detect clustering structure and guide parameter selection for automated clustering algorithms. We utilized an independent measure, in the form of silhouette values, to characterize the quality of the clusterings generated by the spectral decomposition. The results support the conclusion that the singular value decomposition can aid in determining the presence of structure and selecting appropriate clustering parameters. As a result, these concepts seem like good candidates to improve automated clustering.

ACKNOWLEDGMENT

The authors would like to thank the Sensemaking/PACOM Fellowship and Swansea University's Network Science Research Center for challenging us and others to engage in researching this topic and also for the opportunity to examine this data set. Also, the authors would like to thank Daniel Rajchwald for his previous efforts to develop software input and preprocessing routines, which we reused, to prepare the data for analysis. Finally, the authors are very grateful to, Jan-Kees Buenen and Stef van den Elzenhave for advice regarding this paper.

REFERENCES

[1] M. Newman, Networks, An Introduction. Oxford : Oxford University Press, 2010.

[2] Y. Chi, X. Song, D. Zhou, K. Hino, and B. Tseng, “Evolutionary spectral clustering by incorporating temporal smoothness,” in KDD Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug 2007, pp. 1-5.

[3] P. Grindrod and D. Higham, “Evolving graphs: Dynamical models, inverse problems, and propagation,” in Proceedings of the Royal Society A, 2009, pp. 753-770.

[4] N. Halko, P.G. Martinsson, J. Tropp, “Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions” arXiv.org report 0909.4061, 22 Sep 2009 .

[5] T. Klemas and D. Rajchwald, “Evolutionary Clustering Analysis of Multiple Edge Set Networks used for Modeling Ivory Coast Mobile Phone Data and Sensemaking”, Data Analytics 2014, Third International Conference on Data Analytics.

[6] H. Jo, R. Pan, and K. Kaski, “Emergence of bursts and communities in evolving weighted networks,” PLOS ONE, 2011, pp. 1-3.

[7] A. Ng, M. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” Advances in Neural Information Processing Systems (NIPS), vol. 14, 2002, pp. 1-6.

[8] P. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” Computational and Applied Mathematics , vol. 20, 1987, pp. 53-65.

[9] V. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, and E. Huens, “Data for Development: The d4d challenge on mobile phone data,” arXiv:1210.0137v1 [cs.CY], Sep 2012, pp. 5-9.

 ISBN: 978-1-61208-423-7

Blind Spots and Counterfeits in the Supply Chain

Lessons from Haiti that can be well applied to the Philippines and Hawaii

Alison Kuzmickas, Steve Chan, Robert Spousta III, Simone Sala
Dr. Steve Chan Center for Sensemaking, AIRS
Swansea University NSRC and Hawaii Pacific University
Swansea, Wales, U.K.; Honolulu, HI, USA
Email: akuzmic@mit.edu, stevechan@post.harvard.edu, spousta@mit.edu, salas@mit.edu

***Abstract*—Collaborative Big Data Analytics involve a variety of techniques for information gathering and source authentication, including crowdsourcing data. In turn, the development of crowdsourced and participatory mechanisms for a more transparent supply chain is pivotal to identify blind spots and mitigate their impact on global citizens' health and safety. Such effort is also instrumental in reducing cyber risks from the digital world that currently represent a major threat to the stability of national and international economic systems. The paper reviews the lessons learned from the earthquake that devastated Haiti in 2010, whereby the successful combined application of crowdsourced crisis mapping with standard disaster relief operations was equalized by the challenges of data verification for the authenticity of humanitarian products. By overviewing the threats to supply chains coming from the digital world, the paper proposes the adoption of next-generation network science tools to enhance transparency of global supply chains and reduction of cyber risks on the global society.**

Keywords-Big Data, Cyber-Physical Supply Chain, Decision Engineering, Social Complexity Science, Technological Innovation

I. INTRODUCTION

At first glance, the Haiti Earthquake and Tsunami of 2010 demonstrates the extraordinary value of modern day geolocative technological responses. Ushahidi, a geospatial platform that became the go-to for a number of 2010 extreme-scale disasters (including the January 12th and February 27th earthquakes in Haiti and Chile, respectively) propelled two ideas onto a global stage. First, it demonstrated the merit of Gartner Research Vice President Anthony Bradley's 2008 blog post that proclaimed "Every Twitterer as a Sensor" for the reporting of timely information, via Twitter posts and Short Message Service (SMS) text messages from disaster areas, so as to assist boots-on-the-ground officials in locating and prioritizing those victims with the most time-sensitive needs as well as to effectuate the point-of-need delivering of humanitarian and medical relief more rapidly. Second, it formulated the intersection graph for Kevin Ashton and the Massachusetts Institute of Technology Auto-ID Lab's ubiquitous computing mantra of the "Internet of Things" with Graham Cluley's descriptor of the current state of the Web — the World Where Web — in which everything is increasingly being tagged, tracked, mapped, and construed as part of a global supply chain.

With this study space clearly illuminated, we began to get a handle on the incredibly complex Wurlitzer of supply chain logistics that runs the gamut from inbound helicopters precisely choreographed with the whistling of extending flaps by landing aircraft, at the U.S. Air Force-operated Toussain Ouverture International Airport in Port-au-Prince (P-au-P), to the orchestral grinding of gears amidst the convoy of United Nations (U.N.) vehicles and the sharp metallic sounds of the local Haitian trucks, which serve as supply transports from P-au-P to the abutting rural areas of need in Carrefour, Leogane, Delmas, and Jacmel. However, just as the U.N. was about to plant the pennant of victory so as to memorialize its successful distribution of essential drugs and medical treatments to throngs of grateful Haitians, a locally well-known, but little advertised phenomenon (and **blind spot**) arose from an obscured subterranean position to a prominent surface location, and the ensuing tectonic shift sent a high magnitude shock wave through the entire humanitarian world; many of the crates chock-full of emergency supplies and medicines were filled with counterfeit pharmaceuticals [1].

In this paper, we explore vulnerabilities in the global cyber-physical supply chain, and offer mitigation strategies. In Section II, we begin by establishing the danger of counterfeit goods in the global supply chain through various recent examples. In Section III, we discuss the economic impact of counterfeiting on corporate brand images, and the need for increased private sector focus on cyber risk mitigation. In Section IV, we offer the "See Something, Say Something" mantra of citizen vigilance and homeland security as a means for enhancing resilience in the cyber-physical supply chain. In Section V, we discuss how increased transparency, when coupled with citizen vigilance and technological innovation can yield more resilient global supply chains, and we conclude in Section VI.

II. COUNTERFEITS IN THE SUPPLY CHAIN

Whereas having a fake Louis Vuitton bag does not pose any personal risk per se, counterfeit drugs pose a clear and present danger to both the patient and the provider of medical materiel [2]. In one stroke, the integrity of the savior white knight's supply chain was called into question, and as we obtained an increasingly deeper understanding of the P-au-P supply chain and engaged in a hermeneutic examination of the actual machinations for the supplying of

 ISBN: 978-1-61208-423-7

the much needed humanitarian aid from the concerned-community-at-large to Haiti's devastated regions, our sense of organizational triumph was swiftly punctured. We quickly discovered that even with our dedicated and sustained efforts towards this worldwide-attention-receiving mission, our incredible preponderance of logistical force conjoined with the aggregate of multinational "no-expenses-spared" herculean technological muscle, with plenty of technological safeguards, simply was not sufficient to prevent the fact that large supplies of medicinal drugs in Haiti, in many instances delivered under the haloed imprimatur of a respected non-governmental organization (NGO) or sanctioned sovereign force, still turned out to be false and potentially harmful to those disaster victims, who desperately needed these supplies.

Even though P-au-P is no stranger to counterfeits (such as when it happily received from New York City, in April 2010, approximately $10 million worth of NYPD-seized knockoff footwear and clothing [3], which sported spurious labels ranging from Nike to Ralph Lauren), P-au-P also retains painful lingering memories of the death of eighty nine of its children who died from bogus cough syrup containing antifreeze [4]. The ever-increasing prevalence of these reported horrific incidents involving harmful counterfeit medicines is frightening: anti-inflammatories that contain leaded road paint [5], antibiotics that are made of talcum powder [6] or flour [7], and other purported life-saving pharmaceuticals that contain atrocious ingredients such as floor polish [8], sawdust [9], and rat poison [10].

The cry, "The evil of [fraudulent] fake drugs is worse than the combined scourge of malaria, HIV/AIDS, armed robbery, and illicit drugs" [11] echoes throughout the developing world, and some experts have estimated that there are about a million deaths a year from the consumption of counterfeit drugs [12]. Even in the cases for which there actually are some active ingredients in the sham drugs, the trace amounts are not sufficient to function effectively and, ironically, actually induce the virus (because there is insufficient potency to kill it) to mutate into an entirely new strain, thereby causing the unwitting patient to develop an irreversible resistance against subsequent treatments by legitimate medicines. It turns out that these pestilent counterfeits not only irreparably harm these innocent patients, but the fraudulent mislabeling and ensuing breach of trust for the alleged brand also tarnishes the reputation of the victim company.

Despite valid mitigating factors in each of these counterfeiting cases, the stigma and intensely negative perception attached to the incidents cannot be displaced or dispelled by the victim companies. The counterfeiting state of affairs has become a force onto itself, and the World Health Organization (WHO) has estimated that up to 10% of the world's pharmaceutical market is now comprised of these spurious drugs [13], and for some cases in Africa, Latin America, and Asia (including the Philippines), these counterfeits congest up to 30% of those markets [14].

The counterfeit pharmaceutical market equates to approximately U.S. $75 billion [15], and pharmaceutical firms must now diligently maintain global intelligence efforts and actively collaborate with law enforcement to search out, seize, and destroy counterfeit products in order to protect the integrity and reputation of their brands. From the myriad of various jurisdictions from around the world, the dedicated and indefatigable anti-counterfeiting hounds of law are more than eager to assist in these mutually reinforcing Public-Private Partnership Initiatives (P3I) because the involved host nation's economic success and progress is predicated upon the notion of uninterrupted trade; any lack of confidence in iconic brand names most definitely constitutes a barrier to the flow of goods.

Many iconic brand names have suffered, and the traditional top-down supply chain approaches are fraught with issues of opacity, particularly when corporate annual reports only necessitate peering at the primary layer of suppliers. In essence, operations ranging from the U.N. operations in P-au-P to large distributors and manufacturers, such as Wal-Mart and Boeing, amidst the increasingly convoluted supply chain web in these hard economic times, can no longer readily identify who the suppliers of their suppliers are. Traditionally, transparency is divided along two dimensions. Given a more constrained product line, and particularly if there is an extremely popular product, firms might provide complete transparency about just that specific product. In contrast, given an enormous product line or a wide swathe of involved components, transparency might only go one or two levels deep. When transparency does not run deep, there are blind spots and things can go bump in the night, for the nation as a whole, such as when a large company like Boeing is impacted.

In the case of the U.S.-based Boeing Company, it not only has an iconic brand, but it is also one of several large companies whose success or failure can have an enormous impact on the U.S. economy; an interruption of just a few weeks in the company's production contributed to a 6.2% decline in the U.S. Gross Domestic Product (GDP) in the fourth quarter of 2008 [16]. This recital of national factual significance underscores a core tenet and forms the touchstone for not only the treatment of counterfeits and the explorations for a crowdsourced participatory mechanism for a more transparent supply chain, but also for the revealing of an unexpected opportunity to simultaneously tackle another ominous national priority — **cyber risks in the digital world**.

III. COUNTERFEITS, CORPORATE REPUTATION, AND THE NEED TO BETTER MANAGE CYBER RISK

Given today's litigious climate, organizations across the board now take **cyber risks** very seriously, and nothing is more valuable to a business than its reputation [17]. Hence, cyber brand attacks, which leverage a company's valuable brand for nefarious purposes, are particularly dangerous. Firms, such as Novartis, assert that their brand depends

 ISBN: 978-1-61208-423-7

upon their ability to assure patients that products bearing the Novartis label are, in fact, Novartis products, which are inherently underpinned by elevated unwavering standards of "quality, safety, and efficacy" [18].

The most vicious of cyber brand attacks is malware (malicious software) [19], and 38% of all cyber attacks use malware [20]; in fact, there are 60,000 new pieces of malware identified per day [21]. At a broad level, malware is best identified by us as simply the Google warning, "this web site may be harmful to your computer." Behind the scenes, malware is designed to target the contact list of the victim of attack. The contacts might start receiving Viagra spam (unsolicited e-mail containing, in many cases, a payload such as malware) and other unsolicited email messages pertaining to a variety of pharmaceutical drugs. According to the *Verizon Data Breach Investigations*, the majority of all corporate data breaches are effectuated by organized criminal groups [22]. This Poneman Institute study puts the average cost for a data breach at $202 for each customer record compromised [23], and the pinnacle of severe data breaches has ended up costing approximately $109 million in the case of Heartland Payment Systems [24] (the sixth largest credit card processor in the U.S.) and $4.5 billion in the case of TXJ Companies [25] (the parent company of T.J. Maxx, Marshalls, and Home Goods). To compound this situation, with regards to the contacts receiving the spam, in the cases for which medicines are purchased over the Internet, approximately 50% of the pharmaceuticals have been found to be counterfeit [26].

These phenomena have prompted the realization that apart from contending with the baseline preexisting legalities for **cyber risks** as specified under the Family Educational Rights and Privacy Act (FERPA) of 1974, Health Insurance Portability and Accountability Act (HIPAA) of 1996, Sarbanes-Oxley Act (SOX) of 2002, Federal Information Security Management Act (FISMA) of 2002, et al, the disciplinary scope of the cyber practitioner is broadening to encompass not just data breaches, but also risk management, supplier management, brand protection, perception management, and reputation management. This growing scope will very likely be accompanied by increasing liability associated with these expansion zones, such as product liability suits, even in those cases for which the company is simply a data breach victim. This is underscored by both the U.S. Foreign Corrupt Practices Act and the U.K. Bribery Act, which assert that companies are now responsible for bringing their supply chain partners into an overall compliance program [27]. Additionally, the degree of personal liability will directly correspond to the cyber practitioner's day-to-day operational involvement and tangible actions of due care. After all, we now live in an era of elevated standards, whereby it is not just the *letter* of the law that is critical, but the compliance with the *spirit* of the law. Suffice it to say, this presents serious problems for the **cyber risk** professional.

In the realm of physical security, if something is really a big threat, you can typically see it coming — the rabid dogs coming over the other hillside or the army crossing the isthmus. You can readily see these threats, and they do not constitute a surprise. In the cyber security world, that is simply not the case. By way of example, if your car is stolen, you will notice. If your data is stolen, you still have it. If the police do an exceptionally good job, and your car reappears in your driveway, you know that no one else has it. But if your data has been stolen, you will never again be able to say whether or not someone else has a copy [28]. This poses an ongoing liability for companies experiencing data breaches. Given this operating environment of **cyber risk**, cyber security, information assurance, risk management, or whatever term of art one wishes to use, the arena is deemed to be an incredibly challenging intellectual field to engage in, because the problem space changes so rapidly. Even the recognized subject matter experts worry that, every day, something that was true yesterday might no longer be true today. Each and every day, the corpus of network-attached peripherals (e.g., uninterruptible power supplies, printers, copiers, postage meters, digital signs, point-of-sale systems, et al) is distending in size and its constituents are becoming increasingly computerized and subject to cyber attack (thereby constituting increased **cyber risk**). This interrelation of factors is depicted below in Figure 1.

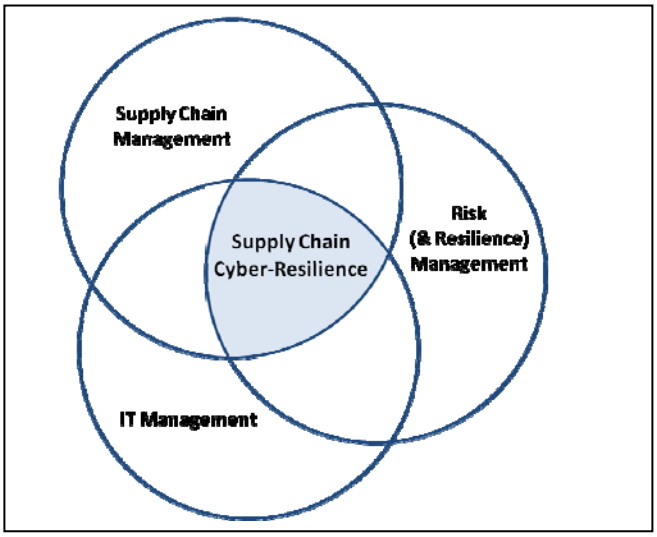


Figure 1. Main knowledge domains in supply chain cyber-risk management, Khan & Estay, Technology & Innovation Management Review, April 2015

The nexus between enhancing our **cyber risk** posture and increasing the transparency of the global supply chain is clear, as depicted above in Figure 1, and there are numerous researchers active in this space, including members of our team. First, cyber crime is the venue of choice for organized crime [29], particularly since the rule of law for Internet crime is — because of its transnational borderless nature — far more nebulous (e.g., there exists a non-unified motley

crew of international treaties for extradition [30]) than conventional crime (traditional, illegal behaviors that most people think of as crime). Generally, Internet-facilitated spam and the sale of bogus goods carries with it penalties of up to five years of incarceration under the Criminal Spam Act of 2003 [31] and a similar tenure exists under the Anti-Counterfeiting Consumer Protection Act of 1996 [32], respectively, which by comparison, represent a much lower threshold of penalties than selling fifty grams of crack cocaine or ten grams of lysergic acid diethylamide (LSD), which carries a minimum of ten years imprisonment without parole under the Anti-Drug Abuse Act of 1986 [33].

Malware and spam are the tools of opportunity for organized crime, particularly since the draconian mandatory minimum sentences under the Federal Sentencing Guidelines Act of 1984 did not anticipate and take into consideration the combination of cyber crime/fraudulent drugs, particularly as the World Wide Web was not yet launched, until six years later, in 1990. Similar to how the odds in gambling are on the side of the casino, the advantage in the cyber landscape is currently skewed in favor of the criminal element. As an example, the average sentence for running an international multi-million pound counterfeit drug operation between 2002 and 2005 was 2.5 years imprisonment. Thus, since the penalty, even in the event of being successfully prosecuted, is comparatively speaking, much lower than other forms of crime, organized criminal groups have gravitated toward this very comfortable and prolific venue of fraudulent drugs — the flagship product promoted by the malware delivered by massive spam campaigns.

The profitability and growth of the counterfeit drug market is currently running rampant [35] and will continue to grow as a juggernaut force until this avenue of corruption has run its course and deemed to be no longer profitable for the currently engaged sophisticated criminal actors. For society to successfully transform the fraudulent drug business from a highly lucrative proposition to an unprofitable venue, the number of law enforcement seizures must increase to dramatically raise the cost of counterfeiting [36]; for these “Title 18” [37] busts to percolate quickly, there need to be more eyes from disparate communities of interest and many more voices from active public participation (such as in the manner consistent with the New York City Metropolitan Transportation Authority’s security slogan, and the now U.S. Department of Homeland Security’s mantra, “see something, say something” [38]).

IV. See something, say something for a more transparent supply chain

Since the time when Google acquired the geospatial data visualization firm, Keyhole, Inc., in 2004 and rebranded the Keyhole software offering, EarthViewer 3D, as Google Earth, the public interest in geographic information systems (GIS) and encompassing geospatial technologies has increased tenfold [39] [40]. Paralleling this phenomenon, the interest in utilizing network science tools and methodologies for better understanding the global supply chain has increased as well. Now, more than ever before, network science practitioners are eagerly observing and reporting various aspects of pedigree/provenance (i.e., the origin or source). This is critical for three reasons. First, in 2006, two hundred and sixty thousand bottles of Panamanian cold medicine contained antifreeze and killed at least one hundred seventy four people [41]. Second, in 2007, bottles of toothpaste bearing the Colgate and Crest brand, which contained substitute ingredients (e.g., diethylene glycol instead of glycerin) provided by a Chinese subcontractor killed at least a hundred people [42]; and third, in 2009, there was a massive peanut butter recall linked to salmonella poisoning, and the list of recalled products included a spectrum of items, such as peanut butter-flavored cookies, crackers, cereals, and ice cream [43].

Ultimately, this recall highlighted the complication of keeping food safe as it makes its way through a complex supply chain from farms to grocery stores shelves to kitchen pantries. Many have long advocated for transparency within the supply chain to provide the all-important origin or provenance information for the consumer [44]. After all, as customers, we want to know the pedigree of what we are buying and using, and of tantamount importance is authenticity. However, we are also concerned with ethics and environmental impact, as in 2006, when Gap, Inc recalled one of its clothing lines after an outcry that one of its Indian subcontractors was using children as young as ten years old to work sixteen hours a day for no pay [45]; and in 2010, when Nestlé fired one of its Indonesian suppliers after Greenpeace revealed that the supplier was destroying vast tracts of rainforests to make way for palm plantations, which produce the palm oil used in Power Bar, Coffee Mate, Nestle Crunch, and other Nestlé products [46].

Figure 2. Example of New York City citizen vigilance campaign

So, what happens when the crowdsourced cylinders are all firing as in the example pictured above in Figure 2 amidst this era of a "Network Science Evolution" which is replete with a rich trove of geolocation and social media information, to help increase the overall transparency of the supply chain and combat counterfeit pharmaceuticals, et al?

V. SUPPLY CHAIN TRANSPARENCY FOR A BREAKTHROUGH IN CYBER SECURITY

Imagine this. Given an increased number of tips from the "see something, say something" mantra and the resultant seizures, the counterfeit drug business is no longer as profitable [47], relative to other criminal venues. Hence, organized crime groups abandon this increasingly stringent sector, as their predominant revenue source, and the mass quantities of malware and spam begin to dip and suddenly fall away. With fewer professional **cyber risk** resources necessary for allocation towards the malware and spam amalgam, the new question becomes, "Does this constitute a breakthrough in governmental cyber risk efforts?" The answer is yes, absolutely [48]. After all, our current "state of the practice" cyber defense systems are not completely automated like science fiction writer Sir Arthur C. Clarke's sentient computer, HAL 9000. There still exists the necessity and significant reality of the all-important human component within human-computer interactions, and a finite amount of manpower necessitates careful prioritization in dealing with the Pandora's Box of cyber risk concerns. Given the newfound excess capacity of human cycles, other cyber risk domains are now able to receive an infusion of much-needed dedicated cycles of attention, thereby segueing into our penultimate question: "Exactly how significant are these newly allocated human cycles?"

Consider the following. When a historically proud seafaring nation, such as Great Britain, retires its flagship earlier than planned and begins to actively shrink the size of its surface warfare fleet so as to increase expenditures on cyber risk, you begin to sense that this new battle space is of serious concern. When you start digging into the classification of national threats within the U.K. and discover that the highest ranking national threat, a "Tier One," is assigned to a devastating attack on computer networks while a "Tier Two" threat is assigned to a nuclear, chemical, or biological attack, the sinking feeling in your stomach provides an indicator that something is afoot.

This begs the riveting question of whether the aforementioned freshly available human cycles can be of value-add to this endeavor? Absolutely, it can. The pathway of transparency within the global supply chain for the partial resolution of the cyber risk problem will be one that effectively makes counterfeit pharmaceuticals an unattractive venue for organized crime, and a lion's share of the world's spam and malware amalgams can indeed be remanded to the past.

To actually realize this vision and to effectuate a transparent supply chain so as to explore and contextualize the pedigrees of the innumerous ingredients and materials (which are sourced from around the world, aggregated, and processed to become the medicines we take, foods we eat, the clothes we wear, the things we buy, and the infrastructure we rely upon), we need to take a deep dive into the world of Big Data where there are branches and sub-branches of information pertaining to the trillion things that we have made in the world.

For this envisioned world, we cannot simply rely upon a centralized top-down identification and tracking system, which may be vulnerable to failure or being compromised by a cyber attack. We must engage the bottom-up distributed democracy, such as the nearly a billion smartphone users in the world, and situated before us now, we have the real possibility of leveraging the "Network Science Evolution" to contextualize the swarm of information from all over the world. This way, we can build a common operating picture of where our morning coffee comes from, whether anybody under the age of fourteen has labored to weave the clothes we are wearing, that the components of the aircraft we are flying today are of the highest standards of engineering excellence, and that the medicine we provide during humanitarian relief efforts are authentic.

VI. CONCLUSION

The earthquake-induced crisis in Haiti in 2010 was exemplary in showing the potential contribution of a complementary application of crowdsourced crisis mapping with standard disaster relief operations. Validation as well as bottom-up editing and management of geographic information were made easier and more rapid than in the past, but in parallel, limitations and challenges of data and information verification did indeed also emerge. Issues of veracity not only contaminated the information supply chain, but also expanded into the whole supply chain logistics that were part of the humanitarian relief efforts. As a result, a relatively large share of the pharmaceuticals supplied in Haiti consisted in fraudulent counterfeit medicinal drugs.

This case highlighted how crucial it is to expand our understanding of supply chain dynamics so as to reduce **cyber risk** as well as to improve disaster preparedness. Within this framework, the notion of transparency is, axiomatically, critical. Within large-scale product supply chains, transparency might only have a two-step depth, and in such cases, the likelihood of **blind spots** occurring increases exponentially. These blind spots can profoundly impact a country's economy, as is exemplified by the case of Japan and Toyota in the aftermath of the 1997 Aisin Fire [49].

It is evident that the there exists a nexus between the physical supply chain and the cyber supply chain, and it is clear that a stable economy and society does pass through the combined the enhancement of transparency of global supply chains and reduction of cyber risks. The case of fraudulent drugs promoted by criminal groups, via cyber attacks (e.g., through massive spam and malware

 ISBN: 978-1-61208-423-7

campaigns) is paradigmatic of such nexus. Indeed, cyber attacks can not only include data breaches, impacts upon supply chain integrity, and other related attack vectors, but also represent a direct and profound danger to corporate reputation, particularly when the criminal group leverages a company's brand for nefarious purposes. Various governments have issued ad hoc laws stating that companies are now responsible for their entire supply chain, including partners; nevertheless, law enforcement is particularly arduous in the area of cyber crime, because the likelihood for criminal groups to be prosecuted is comparatively lower (given the Internets' transnational and borderless structure) and the possible penalty is much lower than other forms of crime.

The development of crowdsourced and participatory mechanisms for a more transparent supply chain is pivotal for identifying blind spots and mitigating their impact upon the integrity of supply chains. Concurrently, such efforts could reduce cyber risks in the digital world that currently represent a major threat to the stability of national and international economic systems. Thanks to the growing availability of crowdsourced and volunteered geographic information, more robust mapping and analytical tools have been developed and/or applied for bottom-up monitoring and mapping of socially or politically sensitive processes. Such results could represent the starting point to develop next-generation of network science tools, which lend to supply chain analytics, via pattern of life analyses, thanks to the integration of official and unofficial information produced (even involuntarily) across social networks and other collective intelligence feedback loops.

ACKNOWLEDGMENT

The authors would like to thank the Cyber Futures Center, an initiative of the Sensemaking-U.S. Pacific Command Fellowship, and the Dr. Steve Chan Center for Sensemaking — one of the centers of the Asia-Pacific Institute for Resilience and Sustainability (AIRS), which is jointly anchored at Swansea University's Network Science Research Center and Hawaii Pacific University — for the opportunity to study the challenges facing Hawaii and other archipelagos, and to contribute towards the various Public Private Partnership Initiatives aimed at developing solutions to overcome those challenges.

REFERENCES

[1] "Haiti Earthquake Victims to Receive $10 Million Worth of NYC's Counterfeit Goods," The Huffington Post 21 April 2010: 1.

[2] "Fatalities Associated with Ingestion of Diethylene Glycol-Contaminated Glycerin Used to Manufacture Acetaminophen Syrup," Center for Disease Control 2 August 1996: 1.

[3] "Counterfeit Internet Drugs Pose Significant Risks and Discourage Vital Health Checks," Science Daily 21 January 2010: 1.

[4] A. Gardner, "Fake Drugs Bought on the Web Pose Big Health Risks," U.S. News & World Report 29 January 2010: 1.

[5] A. Marshall, "The Fatal Consequences of Counterfeit Drugs," Smithsonian.com October 2009: 1.

[6] E. Clark, "Counterfeit Medicines: The Pills That Kill," The Telegraph 5 April 2008: 1.

[7] A. Coghlan, "Sawdust, Coffee and Dirt – Just About Anything Can End Up in Medicines Commonly Sold To The World's Poorest People. How Many More Will Die Before Proper Controls Are Put in Place?" New Scientist 29 March 1997: 1.

[8] S. Boggan, "Headache Pills Made of Rat Poison and Viagra Made of Chalk: We Reveal the Chilling Truth about Internet Drugs," Daily Mail 29 April 2009: 1.

[9] "Opening remarks by R. K. Noble, INTERPOL Secretary General," 2009 International Law Enforcement Intellectual Property Crime Conference 15 December 2010: 1.

[10] "Indian Start-up Strikes Deal to Combat Counterfeiting of Medicine," The Christian Science Monitor 14 December 2010: 1.

[11] "Counterfeit Drugs Pose Dangers in 90 Countries Worldwide," America.gov 14 October 2010: 1.

[12] Activities of the United Nations Office on Drugs and Crime to address emerging forms of crime," Conference of the Parties to the United Nations Convention against Transnational Organized Crime 18-22 October 2010: 15.

[13] J. Rothfeder, "Bumpy Ride," Portfolio.com 22 April 2009: 1.

[14] S. Narisi, "Feds Put IT in the Hot Seat for Security Breaches," Docucrunch.com 5 December 2010: 1.

[15] "Counterfeit Medicines," Novartis November 2005: 1.

[16] "The New World of eCrime: Targeted Brand Attacks and How to Combat Them," Mark Monitor March 2009: 1.

[17] "Organized Crime Wants Your Data," Docucrunch.com 5 December 2010: 1.

[18] L. Dignan, "Cyber Security by the Numbers: Malware Surges, Spam Declines in Third Quarter," ZDNet.com 17 November 2010: 1.

[19] "Expanded Study Finds More Insider Threats, Greater Use of Social Engineering, Continued Strong Organized Criminal Involvement," Verizon 28 July 2010: 1; Verizon's 2012 Data Breach Investigations Report.

[20] "Data Breach Costs Increase," Help Net Security 25 January 2010: 1.

[21] B. Krebs, "Payment Processor Breach May Be Largest Ever," The Washington Post 20 January 2009: 1.

[22] "Estimates Put T.J. Maxx Security Fiasco At $4.5 Billion," InformationWeek 2 May 2007: 1.

[23] "Growing threat from counterfeit medicines," Bulletin of the World Health Organization April 2010: 241-320: 1.

[24] S. Weber, "The U.K. Bribery Act 2010, Cheers!" Adfero Group 13 October 2010: 1.

[25] Interview with Dan Geer, Chief Security Officer for In-Q-Tel, the venture capital arm of the Central Intelligence Agency.

[26] B. Krebs, "Organized Crime Behind a Majority of Data Breaches," Washington Post 15 April 2009: 1.

[27] "UN Rejects International Cybercrime Treaty," ComputerWeekly.com 20 April 2010: 1.

[28] "Congressional Record-Senate," Congress 19 June 2003: 15564.

[29] "Trademark Counterfeiting – Introduction," Criminal Resource Manual 1997: 1701.

[30] E. E. Sterling," Drug Laws and Snitching: A Primer," Frontline, January 1999: 1.

[31] "U.K.'s Largest Counterfeit Drug Operation Concluded," Pharmaceutical Manufacturing 15 July 2009: 1.

[32] J. Schenker, "MPedigree's Rx for Counterfeit Drugs," Bloomberg Businessweek 3 December 2008: 1.

[33] "Report on Counterfeiting and Piracy in Canada: A Road Map for Change," The Canadian Anti-Counterfeiting Network March 2007: 39.

 ISBN: 978-1-61208-423-7

[34] E. Dou, "'See Something, Say Something' Goes National," Huffington Post, 1 July 2010: 1.

[35] GIS in Humanitarian Activities," Aid & International Development Forum 8-9 June 2011: 1.

[36] R. Hart, "Google Earth Popularity Booms," GeoCarta 25 January 2006: 1.

[37] "Panama Releases Report on '06 Poisoning," 14 February 2008: 1.

[38] J. Enoch "FDA Bans Toothpaste from China," ConsumerAffairs.com 24 May 2007.

[39] "How Does Salmonella Get in Peanut Butter? And Can You Kill It Once It's There?" Scientific American 13 January 2009: 1.

[40] Interview with Dr. Steve New, Program Director at the Centre for Corporate Reputation within the University of Oxford.

[41] "Gap: Report of Kids' Sweatshop 'Deeply Disturbing," CNN World 29 October 2007.

[42] D. Gutierrez, "Nestle Drops Indonesia Palm Oil Supplier After Greenpeace Report on Rainforest Destruction," Natural News 18 August 2010.

[43] "Fake Pharmaceuticals: How They and Relevant Legislation or Lack Thereof Contribute to Consistently High and Increasing Drug Prices," American Journal of Law & Medicine 22 December 2003: 1.

[44] Interview with Executive Director Maxim Weinstein, who leads StopBadware.org, a former Harvard Berkman Center for Internet and Society and Oxford Internet Institute project that develops new approaches to address malware.

[45] G. Wilson, "Fight Cyber War Before Planes Fall Out of the Sky," The Sun 19 October 2010: 1.

[46] M. Conner, "Sensors Empower the 'Internet of Things'," Electronics Design, Strategy, News 27 May 2010: 1.

[47] "One Billion Subscribers to Own Smartphone Devices in 2013," Informa 20 September 2010: 1.

[48] T. Nishiguchi and A. Beaudet, "The Toyota group and the Aisin fire," Massachusetts Institute of Technology Sloan Management Review, vol. 40, Fall 1998.

 ISBN: 978-1-61208-423-7

From Cheese to Fondue

A Sensemaking Methodology for Data Acquisition, Analytics, and Visualization

Robert Spousta III, Steve Chan
Dr. Steve Chan Center for Sensemaking, Asia-Pacific Institute for Resilience and Sustainability (AIRS)
Swansea University's Network Science Research Center and Hawaii Pacific University
Swansea, Wales
spousta@mit.edu, s_chan@mit.edu

Stef van den Elzen, Jan-Kees Buenen
Eindhoven University of Technology
SynerScope BV
Helvoirt, The Netherlands
e-mail: s.j.v.d.elzen@tue.nl,
jan-kees.buenen@synerscope.com

***Abstract*—Although Big Data are being leveraged through proprietary means by a host of private enterprises for significant financial gain, there are comparably fewer examples of how to harness the power of massive data through analytics in order to enhance societal resilience and directly serve the public good. In this paper, we present a three-layer framework for conducting Collaborative Big Data Analytics, including data selection and acquisition, steps comprising the analytic process, and considerations for informative data visualization. With regard to data selection, we discuss the primary characteristics of so-called Big Data, namely the Six Vs of data Variety, Volume, Velocity, Veracity, Value, and Volatility. Next, we discuss some of the various analytical tools and techniques available for processing data, as well as methods for effectively visualizing the products of data analytics. In order to illustrate the utility of such a framework, we summarize findings from our participation in Orange Telecom's Data for Development Challenges in the Republic of Côte d'Ivoire and Senegal. We conclude that while the field of Collaborative Big Data Analytics holds great promise, the development of open-source frameworks for conducting layered analytics, combined with the continuation of data challenges, such as those recently held in West Africa, will help to generate more and better uses of the Big Data that have come to dominate our world.**

Keywords*—*Collaborative Big Data Analytics; Decision Engineering; Data Visualization; Sensemaking Methodology

I. INTRODUCTION

Whereas the dot-com boom of the late 1990s and early 2000s ushered in a wholly novel industry, replete with information-based products and virtual services marketed via the Internet, collaborative approaches for conducting civil-centric data analytics have taken longer to develop [1]. This fact notwithstanding, the rise of the Internet of Things (IoT) has introduced unprecedented levels of artificial complexity within many cyber-physical systems, which demand constant attention, lest areas of brittleness and blind spots compromise the resilience of essential services and infrastructure that are the backbone of modern civilization. In order to adulterate this vacuity, we present a basic framework for treating data and gaining insight. This **Sensemaking Methodology** addresses three primary concerns, namely, where and how to get data, how to process and refine data into insight, and how to visualize insight in a way that supports Decision Engineering endeavors. In this manuscript, we briefly outline the system of methods that comprise our three layer framework.

The remainder of the paper is organized as follows. Section II introduces the first layer of our methodological framework; harvesting and generating data, and discusses some of the primary considerations for data selection and acquisition, including the variety of sensor platforms that are responsible for producing data. Section III presents the framework's middle layer of data analytics, and goes on to describe the basic categories of analytic tools and techniques available for data processing. Section IV addresses the framework's top layer; data visualization. Section V summarizes major findings and lessons learned from our participation in the first two Data for Development (D4D) Challenges as an exemplar of the Sensemaking Methodology for **Collaborative Big Data Analytics**. We conclude in Section VI with general thoughts on the state of the art with regard to Collaborative Big Data Analytics, and propose areas for future application of our Sensemaking Methodology.

II. DATA: PROSPECTING FOR THE GOLD OF THE INFORMATION AGE

We embark on our brief journey of discovery by posing two foundational questions. First, where do data come from? And second, how do we get those data? The answers to these primary questions will guide us to an optimal data harvesting strategy, and therefore, form the base of our methodological framework. However, in order to thoroughly appreciate the complexity of these seemingly simple queries, we must first explore the basic nature of data and massive datasets. At the core, we find that the phenomenon of Big Data revolves around the "Six Vs" of Volume, Variety, Velocity, Veracity, Value, and Volatility, depicted in Table 1 below.

The Big Data phenomenon is perhaps most commonly linked with the sheer amount or Volume of data being generated by a host of remote sensors, household appliances, mobile communication devices, and human content generators worldwide that totals over 2.5 quintillion bytes of data per day [2]. Although difficult to comprehend quantitatively, these reams of data come in many forms, from the millions of photos and videos shared daily from smart phones through applications like Instagram, Snapchat, and YouTube, to raw system measurements recorded by sensors and fed into synchrophasor data concentrators and

 ISBN: 978-1-61208-423-7

other industrial control systems [3]. In order to achieve quantitative exactitude whilst navigating complex problem sets, analysts must incorporate a maximally inclusive Variety of data types and sources. In this regard, a critical determinant in achieving perspicacity through the Sensemaking Methodology is the incorporation of diverse data. By way of example, in researching issues of infrastructural resilience, we utilize a host of data gathering mechanisms, including electric grid monitoring equipment such as Phasor Measurement Units (PMU) and Digital Fault Recorders (DFR), Unmanned Aircraft Systems (UAS), Ocean Data Acquisition Systems (ODAS), Synthetic Aperture Radar (SAR) and other weather observation tools, as well as human sensor networks in the form of crowdsourced event observation and reporting. In addition to harvesting a large variety of data, the speed with which data are generated is another equally important variable, as time-critical operations including critical infrastructure protection (CIP), emergency response, law enforcement, and national defense all must be able to sense the occurrence of anomalous events in near real-time in order to prevent loss of life and property [4]. In managing both emergency responses and routine system operations, all data consumers rely on the authenticity or Veracity of data in order to gain actionable insight. The consistency of data taxonomy is an important aspect of Veracity, and, in this regard, discovery standards for electronic resources such as the Dublin Core standards for Metadata are essential for datasets held by diverse curators to remain compatible with one another [5].

TABLE 1 CHARACTERISTICS OF DATA

V	The 6 Vs of Big Data	
	Description	***Units of measure / Dimensions***
Volume	Massive amounts of data	Bytes => Terabytes
Variety	Multiple forms / formats	video, sms, .pdf, .doc, .jpg, .xls, .rtf, .tif, PMU, etc
Velocity	Speed of data feeds	Event-driven / Streaming
Veracity	Trustworthiness of data	Provenance / Pedigree
Value	Usefulness of data	Ambiguity / Uncertainty; Correlation / Causation
Volatility	Shelf-life of data	Time-Sensitive / Static

a. An alternate V of Viability has also been proposed in [2], which we believe is subsumed above

A more persistent challenge for data Veracity is the ability to establish the provenance and pedigree of data, particularly in the context of data manipulation and spoofing, or counterfeiting in the information supply chain. While gathering redundant data from multiple sources, and cross-referencing particularly specious data are prudent strategies for mitigating the negative impact of false or corrupted data, ensuring data Veracity is a perennial problem that demands consistent attention and focus.

Two rather more subjective aspects of data are their Value and Volatility. In **Decision Engineering**, the Value of a given dataset loosely correlates to how much of any given decision can be built from it. In other words, can we decide a course of action based on a single dataset? If so, then that dataset could be said to be of high Value. If many disparate datasets are required in order to engineer a single decision, then each of those datasets is of comparatively low Value. Data's Volatility or duration of relevance depends largely on the nature of the decision it is serving to inform or build. Whereas certain digitally preserved historical records maintain their relevance or Value in perpetuity, other datasets that pertain to rapidly evolving circumstances may remain relevant for only a matter of days, if not seconds. Determining a dataset's Volatility is yet another important step in the process of Sensemaking.

Having established the basic nature of data, we return to the original question of where and how to acquire data. For all organizations - public, private, and any permutation in between - data accessibility and knowledge management remain areas of active research and constant improvement [6]. With the United Nations (UN) recently asserting that information in itself is a life-saving need for people in crisis, just as important as water, food, and shelter, the necessity of open source data is clearly a global one that now transcends the realm of scholarly open access [7]. So, the short answer to our question is that there is no comprehensive, authoritative single source for all data, and therefore, we get data from everywhere we can, however we can.

III. STACKING THE DECK: TOOLS AND TECHNIQUES FOR LAYERED DATA ANALYTICS

Next, we turn to the analytic component of the Sensemaking process, which includes algorithms, cognitive high performance computing, machine learning, signal resolution, allegorical engines, and the Unstructured Information Management Architecture (UIMA). Many of these components are rooted in mathematical concepts dating back centuries. Notable examples include the famous problem of the Bridges of Konigsburg and Graph Theory, Ada Lovelace's development of early programming instructions for Babbage's Decision Engine, the Pragmatists' precepts of indeterminacy, order in chaos, and long-run convergence; as well as Turing's Machine, and Weaver's Complex Systems Ontology [8].

The modern analytical toolkit is comprised of far too many instruments to concisely summarize here. However, there are fundamental components of the analytic process, which we will introduce in this manuscript. Upon identifying, generating, and acquiring data, the initial step in the analytic layer of our framework is data ingestion and refinement. By way of example, satellite imagery is unfortunately not as simple as an "eye in the sky" beaming down neat pictures to a computer console for analysis and distribution. The many 0's and 1's that make up the digital representation of a physical object must first be processed and translated into an intelligible picture. Once raw data are refined into a malleable commodity, that commodity can then be annealed into meaningful insight through a systematic layering of Analytics on Analytics (A2O). This process begins with a geospatial and or temporal matrix of

 ISBN: 978-1-61208-423-7

data points, and proceeds through a set of systematic organizational steps that include data clamping, normalization, and hierarchical clustering, in order to reveal traces of emergent phenomenon and achieve pattern recognition. Such patterns are the bedrock of insight, and serve to evaluate the role of myriad variables in the emergent outcomes of complex systems and networks, as depicted below in Figure 1. However, a fundamental prerequisite for effective A2O is the storage and management of massive datasets. In this regard, distributed computing architectures and parallel processing are also prominent features in the analytic layer of the Sensemaking Methodology [9].

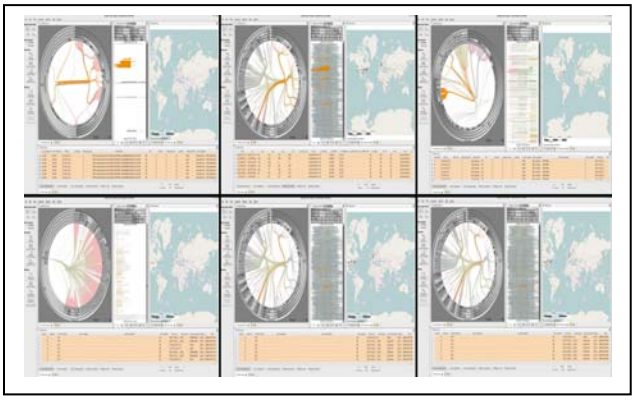


Figure 1. Example of SynerScope A2O Visualization Suite

Impressive though they may be, machine capabilities comprise but one half of the analytic layer of our methodological framework. The remaining half relies on the inherently human capabilities of contextual orientation and intuitive leaping [10]. Whereas machines are capable of generating, processing, and storing massive quantities of data, the human mind remains unique in its ability to superimpose context over data in order to discern relevance and meaning. Hence, the Sensemaking Methodology is characterized by its counterpoising and fusion of socio and techno perspectives. On the one hand, we leverage the technical advantages of machine capability to yield algorithmic insight. On the other hand, we also leverage inherent knowledge of the human social condition and sentient thought to arrive at heuristic insight. This socio-techno unification is at the heart of our methodology for pattern recognition and Decision Engineering. Going back to the example of satellite imagery, let us consider the case of the Global Earth Observing System of Systems (GEOSS) and the view of Somali villages at night as an illustration of counterpoising algorithmic versus heuristic insight. With the rise of both maritime piracy off the coast of the Horn of Africa, and the violent extremist organization Al-Shabaab in Somalia, international security organizations were keen to establish a link between the two groups [11]. As assets in the GEOSS satellite constellation observed significant variances in the night-time illumination of various towns along the Somali Coast and provincial capitals, analysts sought to employ the algorithmic insight as evidence for a correlation between the dispensation of pirate ransoms and the buildup of jihadi strongholds [12]. However, heuristic insight suggested that the ideological and religiously-motivated nature of Al-Shabaab was incompatible with the financially-driven motives of the criminal piracy network, and therefore a link was unlikely. The truth of this insight would later be established through data gathered by the International Criminal Police Organization (INTERPOL) and the United Nations Office on Drugs and Crime (UNODC) [13]. Such an example shows us that while technology and algorithmics are more than capable of identifying patterns of interest, we still need heuristic insight to decipher what those patterns actually mean.

IV. A PICTURE TELLS A THOUSAND WORDS: IMPARTING INSIGHT THROUGH DATA VISUALIZATION

Upon recognizing patterns of interest, we are now ready to move into the third and final phase in the **Sensemaking Process**; visualizing insights for Decision Engineering. The primary aim of the data visualization phase is to establish the relevance of insight gained through the A20 process, and ultimately answer the basic question of “So what?” Figure 1, above, displays output from one of our visualization platforms, the SynerScope. SynerScope and other similar tools use a coordinated multi-view approach with a scalable and flexible visual matrix in order to visualize key insights from massive datasets.

However, before we progress into any further detail with regard to contemporary visualization techniques, let us briefly consider the history of data visualization. The roots of visualization are as old as human knowledge and communication; from cave paintings, to pictographs, hieroglyphics, numerology, symbolic logic, and language. In order to understand what methods have been developed over time for effectively conveying knowledge and information, it is instructive to visit certain historical examples. One case in point is the work of the Mixtec civilization of Oaxaca, Mexico [14], depicted below in Figure 2.

Figure 2. Image from the *Codex Vindobonensis Mexicanus*

 ISBN: 978-1-61208-423-7

Although the figure above depicts the Mixtec's primordial cosmology and creation mythology, it is an early example of how human insights gained through observation of natural phenomenon (i.e., data analysis) were preserved for distribution and posterity. This and other similar precedents from early civilization remain germane to many data-related fields, including Education, the Arts, Public Information, Manufacturing, Product Advertisement, Device Instruction Manuals, Traffic Signage, Emergency Management, and Information Technology (IT) [15]. With the advent of the Internet, and eventually the World Wide Web, the tradition of data visualization has continued to evolve. Today, such professional disciplines as Cognitive Science, Behavioral Psychology, Computer-Assisted Design (CAD), and Strategic Communication all build on the work of early visualization specialists by combining machine capability with human insight to generate socio-techno innovations in how the brain senses and interprets information. In turn, our interpretation and assimilation of information drives our ability to engineer decisions and determine appropriate courses of action, as individuals in daily life, as agents in organizations, and as members of the global citizenry.

Nevertheless, this does not mean that modern data visualization is a perfected science. Rather, visualization is a principled art that requires both intelligence and intuition in its composition. In turn, efforts to visualize pseudo-insights that are not informed by robust A2O run the risk of proliferating misinformation, bias, conflict, and spoilage of resources [16]. In addition to these pitfalls, data-informed visualizations also can be subject to information overload, if insights are not concisely crystallized in a digestible form, as depicted in Figure 3 [17].

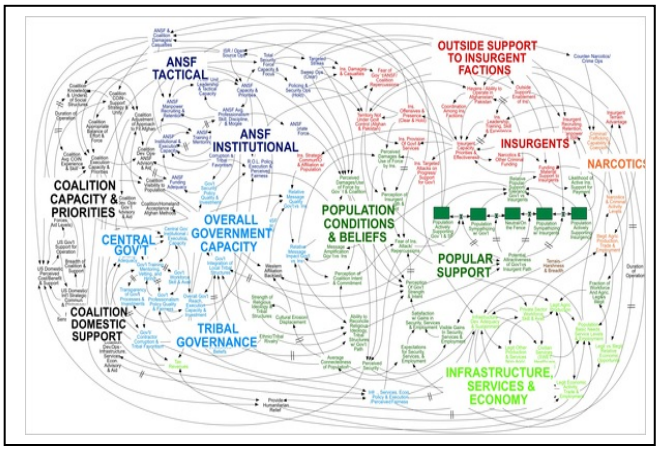


Figure 3. Example of Counterinsurgency Diagram

The design of any given data visualization is driven by two primary factors; the nature of the decision it serves to engineer, and the demographic characteristics of the audience or consumer. Firstly, is the aim of the visualization simply to impart generally useful information, or is it intended to inform a specific choice? If the aim is the former, then visualizations such as that in Figure 3 may be appropriate. However, decision-quality visualizations must clearly depict actionable intelligence, and offer tangible courses of action. Secondly, how much does the target audience for a given data visualization already know? An audience of laymen will require a significant amount of context in order to make sense out of visualizations. Conversely, too much context will be superfluous (and potentially distracting) to an audience of experts. Therefore, constructing an effective data visualization means striking a delicate balance between sufficient context and specific insight.

With this in mind, we turn to a final consideration regarding the value of data visualization; the identification of brittleness in complex systems. In light of the staggering layers of complexity and interdependence that characterize many of our most critical infrastructural systems (e.g., electric grids, the Internet, etc.), there is significant potential for percolation effects or cascading failure [18]. Therefore, to ensure the resilience of such systems, it is essential to identify areas of brittleness or weak links in the chain before they fail. With regard to the resilience of the Internet in particular, tools such as the SeeSoft System, pictured below in Figure 4, enable analysts to visualize statistics of interest in software code [19]. In the case of Figure 4, a color-coding scheme displays how recently lines of code have been changed, with red lines having been most recently changed, and green lines having remained unchanged the longest.

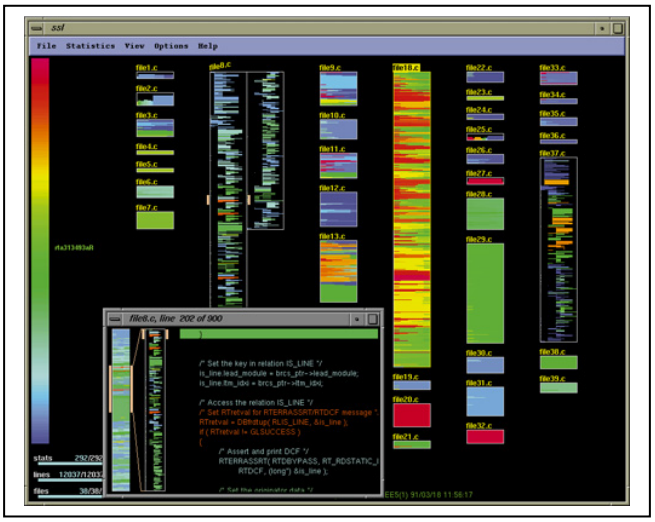


Figure 4. SeeSoft software code visualization system, Lucent Technologies

Visualization tools are invaluable assets that enable us to quickly and clearly see areas of potential brittleness in complex systems. In the case of Figure 4, above, we have a mechanism to visualize answers to questions such as whether software security improves with age, as lines of code not recently updated to address proliferating cyber threat vectors are likely brittle [20]. Therefore, visualization is not only a product of the analytic phase of the

 ISBN: 978-1-61208-423-7

Sensemaking Methodology, but can actually be a feedback loop that helps to inform the A2O process.

V. PROOFS OF CONCEPT: SYNERSCOPE AND THE DATA FOR DEVELOPMENT CHALLENGE

With our **Sensemaking Methodology** in hand, we finally come to the shores of West Africa and the Data for Development Challenge (D4D) [21]. Since its inauguration in 2012, the annual D4D Challenge has represented a unique opportunity for Big Data analysts to experiment with diverse tools and techniques for harvesting insight from mobile phone data. For each challenge, international competitors from academia and private industry are given the chance to analyze a multitude of datasets pertaining to mobile phone use in a designated country during a circumscribed portion of the year [22]. We have had the privilege to participate in both challenges thus far, in the Republic of Côte d'Ivoire and Senegal, with a sampling of our results displayed below in Figure 4 [23].

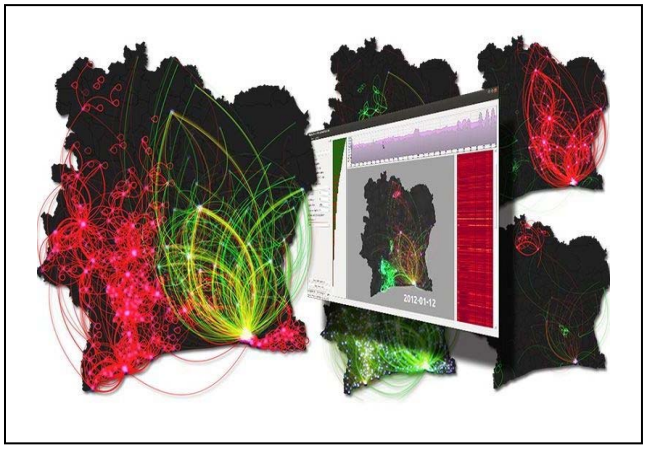


Figure 4. 2013 D4D Best Visualization prize winner: *"Exploration and Analysis of Massive Mobile Phone Data: A Layered Visual Analytics Approach"*

In conducting our analysis of the D4D datasets and generating the illustrations sampled above, two lessons became clear to us. First, we needed data Variety, through which to contrast and correlate mobile phone activity with other significant trends and events. For the first D4D in Côte d'Ivoire, we contrasted the given mobile phone data with UN reports of violent conflict and significant social disturbance, as well as meteorological data for the given timeframe. This helped to reveal regional political affiliations and ethnic enclaves, as violent events targeting certain political and ethnic groups in the capital city, Abidjan, catalyzed notable increases in call activity to specific communities elsewhere in the country. In addition, we observed that abundant rainfall in areas of significant cocoa and yam cultivation correlated with heightened call activity, likely indicating increased agro-business developments at specific points in the growth and harvest cycles in response to favorable weather conditions. Our second lesson learned was the need to adopt multiple perspectives from which to interrogate the datasets. Our normalization and clustering algorithms produced dendograms, with which we were able to sort items (e.g., cell towers) of similar behavior into groups for further investigation. By grouping cell towers of similar call behavior, we were then able to further explore what other commonalities linked these disparate regions.

Although such techniques are still relatively nascent, we believe that the work of our team and fellow D4D participants is a clear demonstration that Collaborative Big Data Analytics can help to increase insight into complex interrelated phenomenon, and thus improve Decision Engineering in a variety of social, political, and economic arenas. However, the implementation of our **Sensemaking Methodology** remains in the early stages, and inevitably there is room for improvement in such an approach. Specifically, increasing the Volume and Variety of data included in the A2O phase will yield greater insight in future D4D Challenges, and other applications of our methodological framework. In addition, the deliberate articulation of alternate frameworks for Collaborative Big Data Analytics will help to progress the state of the art, by revealing common best practices as well as shortfalls and gaps.

VI. CONCLUSION: STANDING ON THE THRESHOLD OF A BRAVE NEW WORLD

Our journey ends with the realization that humanity's quest for insight is by nature eternal. Although it is temporally little, the story of Big Data is truly epic. As machine capability continues to accelerate, the power and promise of data analytics will only grow. At the same time, our ability to make sense out of evolving circumstances quickly, and adapt social structures accordingly will be important determinants in the shape of things to come.

Our experiences with D4D and other instances of Collaborative Big Data Analytics are evidence that critical thinking is an inseparable ingredient in the recipe for Big Insight, and that socio-techno approaches are an indispensable element of complex problem solving. We believe that open and inclusive approaches such as the **Sensemaking Methodology** have the potential to enhance numerous dimensions of resilience, including those of cyber-physical systems, societies, and individuals. Systematic Decision Engineering is a practical way to identify latent Black Swan blind spots, Maginot Line-scale brittleness, and Pearl Harbor-level threat vectors. Similarly, we also hope that such a methodology can facilitate positive developments, such as the smart integration of green technologies into sustainable Blue Economies [24], and an improvement in our roles as both environmental stewards and engines of social progress. Each of these areas represents exciting and relatively unexplored realms of

 ISBN: 978-1-61208-423-7

research that we have designated as targets for future work. Specifically, we plan to demonstrate how technological advancements such as Pervasive Remote Sensing (PRS), Comprehensive Domain Awareness (CDA), and Cognitive Computing can be effectively integrated with human Sensemaking techniques to achieve increasingly useful insights and practical Decision Engineering solutions.

ACKNOWLEDGMENT

The authors would like to thank the Cyber Futures Center, an initiative of the Sensemaking-U.S. Pacific Command Fellowship, and the Dr. Steve Chan Center for Sensemaking — one of the centers of the Asia-Pacific Institute for Resilience and Sustainability (AIRS), which is jointly anchored at Swansea University's Network Science Research Center and Hawaii Pacific University — for the opportunity to study the challenges facing Hawaii and other archipelagos, and to contribute towards the various Public Private Partnership Initiatives aimed at developing solutions to overcome those challenges.

REFERENCES

[1] N. R. Council, *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press, 2013.

[2] N. Biehn. (2013, May) The Missing V's in Big Data: Viability and Value. *Wired*. Available: http://www.wired.com/2013/05/the-missing-vs-in-big-data-viability-and-value/ accessed May 20, 2015

[3] C. Alcaraz and J. Lopez, "Wide-Area Situational Awareness for Critical Infrastructure Protection," *Computer*, vol. 46, pp. 30-37, 2013.

[4] K. M. Chandy, "Sense and respond systems," in *Int. CMG Conference*, 2005, pp. 59-66.

[5] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf, "Dublin core metadata for resource discovery," *Internet Engineering Task Force RFC*, vol. 2413, p. 132, 1998.

[6] M. Alavi and D. E. Leidner, "Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues," *MIS Quarterly*, vol. 25, pp. 107-136, 2001.

[7] C. Hajjem, S. Harnad, and Y. Gingras, "Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact," *arXiv preprint cs/0606079*, 2006.

[8] W. Weaver, "Science and Complexity," *American Scientist*, vol. 36, pp. 536-544, 1948.

[9] G. S. Sureshrao and H. P. Ambulgekar, "MapReduce-based warehouse systems: A survey," in *Advances in Engineering and Technology Research (ICAETR), 2014 International Conference on*, 2014, pp. 1-8.

[10] N. Ford, "Information retrieval and creativity: towards support for the original thinker," *Journal of Documentation*, vol. 55, pp. 528-542, 1999.

[11] J. Stevenson, "Jihad and Piracy in Somalia," *Survival*, vol. 52, pp. 27-38, 2010/03/01 2010.

[12] A. Shortland, "Treasure mapped: using satellite imagery to track the developmental effects of Somali Piracy," *London: Chatham House*, 2012.

[13] S. Yikona, *Pirate Trails: Tracking the Illicit Financial Flows from Pirate Activities Off the Horn of Africa*: World Bank Publications, 2013.

[14] B. E. Byland and J. M. Pohl, *In the realm of 8 Deer: The archaeology of the Mixtec codices*: University of Oklahoma Press, 1994.

[15] J. Z. Gao, L. Prakash, and R. Jagatesan, "Understanding 2D-BarCode Technology and Applications in M-Commerce - Design and Implementation of A 2D Barcode Processing Solution," in *Computer Software and Applications Conference, 2007. COMPSAC 2007. 31st Annual International*, 2007, pp. 49-56.

[16] W. Neil Adger, N. W. Arnell, and E. L. Tompkins, "Successful adaptation to climate change across scales," *Global Environmental Change*, vol. 15, pp. 77-86, 2005.

[17] E. Bumiller. (2010, April 26) We Have Met the Enemy and He is Powerpoint. *New York Times*. Available: http://www.nytimes.com/2010/04/27/world/27powerpoint.html?_r=2 accessed May 20, 2015

[18] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, pp. 268-276, 2001.

[19] S. G. Eick, J. L. Steffen, and E. E. Sumner, Jr., "Seesoft-a tool for visualizing line oriented software statistics," *Software Engineering, IEEE Transactions on*, vol. 18, pp. 957-968, 1992.

[20] A. Ozment and S. E. Schechter, "Milk or wine: does software security improve with age?," in *Proceedings of the 15th conference on USENIX Security Symposium-Volume 15*, 2006, p. 7.

[21] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. M. T. Do, *et al.*, "From big smartphone data to worldwide research: The Mobile Data Challenge," *Pervasive and Mobile Computing*, vol. 9, pp. 752-771, 2013.

[22] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, *et al.*, "Data for development: the d4d challenge on mobile phone data," *arXiv preprint arXiv:1210.0137*, 2012.

[23] J. Poole. (2013, May 6) Winning Research from the Data 4 Development Challenge. *United Nations Global Pulse*. Available: http://www.unglobalpulse.org/D4D-Winning-Research accessed May 20, 2015

[24] G. Pauli, "The blue economy," *Our planet*, pp. 24-27, 2010.

 ISBN: 978-1-61208-423-7

Market Basket Analysis Using Heterogeneous Multivariate Probit Models for Groups of Product Categories

Harald Hruschka

Faculty of Economics
University of Regensburg, Germany
Email: harald.hruschka@wiwi.uni-regensburg.de

***Abstract*—Several heterogeneous multivariate probit models are used to analyze market baskets purchased by households. Each of these models is related to one group of product categories contained in seven prior partitions formed for a total of 25 product categories. The best model in terms of cross-validated log likelihood found considers all categories as one group, i.e., it does not split the 25 categories into two or more groups. In the next step of this project, we will compare this result to multivariate probit models which are related to a partition which is not fixed beforehand, but determined by stochastic model search.**

Keywords–Market basket analysis; Multivariate probit models, MCMC; Stochastic model search

I. INTRODUCTION

Using several heterogeneous multivariate probit models we analyze market baskets, i.e., multicategory purchases of households. Each of these models is related to one group of product categories contained in partitions formed from a total of 25 product categories. In the marketing literature, purchase incidence models as a rule either have a multivariate probit (MVP) or a multivariate logit (MVL) form. Papers applying MVP models typically take latent heterogeneity of households into account. To the best of our knowledge, Manchanda *et al.* [13] provide the first publication analyzing four product categories by MVP models. In their MVP models, Chib *et al.* [6] and Duvvuri *et al.* [9] consider a maximum of twelve and six categories, respectively. Russell and Petersen [15] as well as Boztuğ and Hildebrandt [3] estimate MVL models without latent heterogeneity for a maximum of four and six categories, respectively. Dippold and Hruschka [8] analyze 31 categories by one MVL model and account for latent heterogeneity.

We choose purchase incidences as response variables motivated by he expectation of Song and Chintagunta [16] that interdependences of categories emerge rather on this level than, e.g., for purchase quantities or expenditures. Error correlations are allowed only between categories belonging to the same group. In other words, error correlations are restricted to equal zero between categories which belong to different groups.

To the best of our knowledge, only three studies specify models for different category groups. Chib *et al.* [6] compare the parameter estimates of three MVP models (each model for one group with four categories) and one overall MVP model for all 12 categories. Category groups in [6] are formed by sorting category names in alphabetic order. Boztuğ and Reutterer [4] in a first step determine basket classes by online K-means of purchase incidence data. Then these authors estimate one MVL model for each class. In each MVL model, they consider as category group about five product categories which attain the highest class specific purchase frequencies using data of those households whose purchase incidences have the highest similarity to the relevant basket class.

The paper presented here differs from previous publications in two respects. Firstly, we form seven alternative prior partitions with category groups that reflect the typical uses of assigned categories by household members (e.g., drinking, eating, personal care, cleaning etc.). Then we evaluate the statistical performance of models implied by these seven partitions. Secondly, the total number of categories investigated is much higher compared to studies specifying models for different category groups.

In Section II, we introduce the basic heterogeneous MVP model and subsequently explain the overall model. We give an overview on model estimation in Section III. In section IV, we characterize the data used and present estimation results. In the final Section V, we summarize main results and mention the next step of the project presented here.

II. MODEL SPECIFICATION

The basic heterogeneous MVP model is characterized by the fact that category constants, coefficients, and residual correlations vary across households. J_g symbolizes the number of categories belonging to a category group g. Indices of product categories are denoted as $j = 1, \cdots, J_g$, indices of households as $i = 1, \cdots, I$, indices of baskets of household i as $t = 1, \cdots, T_i$. Household i purchases category j in basket t (symbolized by a purchase indicator $y_{jit} = 1$) if the stochastic utility U_{jit} of such a purchase is positive. If U_{jit} is negative, the household does not purchase category j in basket t (symbolized by a purchase indicator $y_{jit} = 0$). Stochastic utility U_{jit} results from deterministic utility V_{jit} (a linear combination of independent variables plus category constant $\beta_{1,ji}$) to which error ϵ_{jit} is added. We obtain the following expression:

$$U_{jit} = \beta_{1,ji} + \sum_{d=1}^{D} \beta_{1+d,ji}\, x_{1pi} + \sum_{m=1}^{M} \beta_{1+D+m,ji}\, x_{2mjit} + \epsilon_{jit} \quad (1)$$

The model includes two types of independent variables in (1). The first type consists of D predictors x_{1di} which differ across households, but assume the same value for all market baskets and categories of any household i. Socio-demographic household variables are examples of this type of

 ISBN: 978-1-61208-423-7

independent variable. Coefficient $\beta_{1+d,ji}$ indicates the effect of such a household-specific variable d on the utility for category j. The second type of independent variables are M marketing variables x_{2mkit} which differ across market baskets of household i and are specific to category k. Coefficient $\beta_{1+D+m,ji}$ measures the effect of marketing variable m on the deterministic utility of its category j.

We allow errors to be correlated across different categories belonging to the same group. By assuming that errors follow a multivariate normal distribution with zero mean vector and a (J_g, J_g) error covariance matrix the MVP functional form results. To attain identifiability we restrict the error covariance matrix to a correlation matrix [6].

To account for latent heterogeneity of households we use a Dirichlet process mixture (DPM) with MVP models as components. This way we allow for infinitely many household clusters in the overall population, with an unknown number of clusters observed in the finite sample [14]. The DPM is capable to reproduce multimodal and skewed distributions and determines the number of latent clusters alongside the estimation process (see, e.g., [1]). The prior of a DPM is a Dirichlet process, which in this case consists of two independent distributions. The first one is a multivariate normal distribution of category constants and coefficients [6]. The second one is a uniform distribution on the space of correlation matrices of dimension J_g which corresponds to a prior developed by Barnard *et al.* [2] on which Liu and Daniels [12] base an appropriate Metropolis-Hastings simulation step.

The overall model can be seen as union of several heterogeneous MVP models, each of which is specific to one of G groups of a partition of the total set of categories. Error correlations between categories assigned to different category groups are zero.

III. Model Estimation and Evaluation

Models are estimated by iterative Markov chain Monte Carlo (MCMC) simulation comprising the algorithm 7 of Neal [14] to construct household clusters, and additional sampling steps to estimate stochastic utilities, a correlation matrix, category constants and coefficients for each group and cluster. We evaluate performance of each overall model by the expected log likelihood over cross-validated predictive densities, which we briefly call cross-validated log likelihood (CVLL). Cross-validation predictive densities indicate which market baskets are likely if a model is fitted to all data with the exception of the respective observation, i.e., market basket t of household i [10]. To this end parameter samples $\boldsymbol{\theta}_{s,-it}$ with $s = 1, \cdots, 500$ are drawn from the density of parameters $f(\boldsymbol{\theta}_{s,-it})$ using the resampling approach described in Gelfand [10]. $CVLL$ values of a model are defined as:

$$CVLL = \sum_{i=1}^{I}\sum_{t=1}^{T_i}\sum_{j=1}^{J} 1/500 \sum_{s=1}^{500} [y_{jit} \ln p(\boldsymbol{\theta}_{s,-it})(1 - y_{jit}) \ln(1 - p(\boldsymbol{\theta}_{s,-it}))] \quad (2)$$

We compute the probability $p(\boldsymbol{\theta}_{s,-it})$ as relative frequency that the j-th element of 500 random number vectors is greater than zero. These random vectors are generated from a multivariate normal distribution with deterministic utilities as expected values and the error correlation matrix R_i all computed from parameter sample s and for the predictors of category j, household i and basket t.

IV. Empirical Study

A. Data

The data refer to 24,047 shopping visits of a random sample of 1500 households to one specific grocery store over a one year period composed from the IRI data set [5]. Each shopping visit is characterized by a market basket, which is a binary vector whose elements indicate whether a household made a purchase in each of 25 product categories

As predictors we consider two binary marketing variables, feature and display, showing whether any brand of the respective category is advertised by local newspapers and receives special placements in the store, respectively. The original data also include information on price reduction, which we omit because of high correlation with the feature variable. The other predictors are household size (number of persons) and a binary variable high income (set to 1, if income is above the median). Table I contains relative frequencies of purchases, feature and display for each category as well as overall means and standard deviations of the number of baskets, basket size (i.e., the number of categories purchased), and household size.

TABLE I. Descriptive Statistics

Category	Abbreviation	Relative purchase frequency	Relative frequency Feature	Relative frequency Display
Milk	milk	0.476	0.129	0.009
Carbonated beverages	carbbev	0.400	0.175	0.283
Salty snacks	saltsnck	0.351	0.154	0.267
Cold cereal	coldcer	0.280	0.151	0.114
Yogurt	yogurt	0.202	0.179	0.020
Soup	soup	0.197	0.112	0.061
Spaghetti sauce	spagsauc	0.181	0.169	0.072
Toilet tissue	toitisu	0.171	0.095	0.081
Margarine/Butter	margbutr	0.158	0.130	0.026
Paper towels	paptowl	0.140	0.067	0.071
Coffee	coffee	0.136	0.124	0.080
Laundry detergent	laundet	0.118	0.106	0.081
Frozen pizza	fzpizza	0.110	0.174	0.121
Mayonnaise	mayo	0.109	0.100	0.054
Frankfurters and hotdog	hotdog	0.103	0.094	0.034
Mustard/Ketchup	mustketc	0.102	0.041	0.054
Frozen dinner	fzdin	0.090	0.187	0.071
Facial tissue	factiss	0.084	0.119	0.048
Peanut Butter	peanbutr	0.080	0.133	0.053
Beer/Ale	beer	0.076	0.061	0.080
Toothpaste	toothpa	0.059	0.089	0.045
Shampoo	shamp	0.053	0.094	0.077
Deodorant	deod	0.040	0.083	0.034
Household cleaners	hhclean	0.030	0.041	0.016
Diapers	diapers	0.020	0.171	0.010

Variable	Mean	Standard Deviation
Number of baskets	16.05	13.47
Basket size	3.85	2.65
Household size	2.36	1.29

B. Estimation Results

Table II lists all prior partitions investigated. Groups of prior partitions differ with respect to the way that assigned categories are typically used by household members, e.g., for drinking, eating, personal care, cleaning etc. These partitions are also typical for category groupings, which grocery retailers use. $A5$ is the most detailed partition with five lowest level groups. We define higher-level prior groups as unions of lower level ones, i.e., Non Food as union of Personal Care and Cleaning, Other Food as union of Other Food Main and Other Food Additional, Food as union of Beverage and Other Food, and finally $A1$ which comprises all 25 categories as union of Food and Non Food. Note that in the case of $A3$ and $A4$ we

actually consider two alternative category partitions ($A3a$, $A3b$ and $A4a$, $A4b$) with three or four groups.

TABLE II. PRIOR PARTITIONS AND GROUPS

Prior partitions	Category groups
$A1$	One group
$A2$	Food, Non Food
$A3a$	Beverage, Other Food, Non Food
$A3b$	Food, Pers Care, Cleaning
$A4a$	Beverage, Other Food Main, Other Food Additional, Non Food
$A4b$	Beverage, Other Food, Personal Care, Cleaning
$A5$	Beverage, Other Food Main, Other Food Additional, Personal Care, Cleaning
Lowest Level Groups	**Categories**
Beverage	beer, carbev, coffee, milk
Other Food Main	coldcer, fzdin, fzpizza, hotdog, saltsnck, soup, yoghurt
Other Food Additional	margbutr, mayo, musketc, peanbutr, spagsauc
Personal Care	deod, diapers, factiss, shamp, toitisu, toothpa
Cleaning	hhclean, laundet, paptowl

Table III contains the best partition in terms of CVLL for a number of category groups varying between 2 and 5. It also contains the results for the model for which all categories belong to one group. Among prior partitions with at least two groups the most detailed one with five groups ($A5$) performs best. But the overall best performance is attained by $A1$, which treats all categories as belonging to one group. This model is, of course, the most complex one in terms of the number of parameters, as it includes 300 error correlations for all the pairs of the 25 categories.

TABLE III. CROSS-VALIDATED LOG LIKELIHOOD VALUES (CVLL)

# of category groups	prior partitions label	CVLL
1	$A1$	-135,028
2	$A2$	- 163,644
3	$A3a$	-153,520
4	$A4b$	-151,764
5	$A5$	-150,298

Values are rounded to nearest integer.

Parameter estimates are based on every 10th of 50,000 iterations, which are immediately consecutive to a burn-in phase of 50,000 iterations. The largest four household clusters are dominant. Vectors of average percentage shares of these four clusters are (58.5, 16.4, 10.1, 5.9) and (23.7, 21.3, 19.2, 16.3) for models $A1$ and $A5$, respectively.

In the following, we present a selection of higher parameter estimates for the two partitions $A1$ and $A5$. These estimates are averaged across households. Table IV shows all significant effects of the two marketing variables which are greater than 0.15 in absolute size for at least one of the two partitions.

These coefficients indicate positive effects of features and displays on utility. Effects of features are more frequent and as a rule higher compared to effects of display. For the most part, effects for partition $A5$ are higher (e.g., for features: coldcer, margbutr, yogurt, hhclean; for display: hotdog, shamp, coldcer, hhclean), a few become insignificant (features: deod, beer; display: beer).

Table V lists significant average error correlations for the two partitions which are greater than 0.200 in absolute size for at least one of the two models. Note that these correlations are all positive. Our interpretation of error correlations follows Song and Chintagunta [16]. In the case of a positive correlation a demand shock which increases (decreases) the utility of category j, also increases (decreases) utility of category j'.

TABLE IV. SELECTED COEFFICIENTS OF FEATURES AND DISPLAYS

Category	Partition $A1$	Partition $A5$	Category	Partition $A1$	Partition $A5$
			Feature		
coffee	0.352	0.367	laundet	0.287	0.332
hotdog	0.291	0.334	shamp	0.274	0.313
spagsauc	0.250	0.279	fzpizza	0.238	0.280
factiss	0.232	0.260	toothpa	0.223	0.247
deod	0.228	-	beer	0.213	-
peanbutr	0.209	0.213	mustketc	0.197	0.177
soup	0.205	0.224	margbutr	0.206	0.475
milk	0.201	0.206	yogurt	0.200	0.248
mayo	0.197	0.192	saltsnck	0.195	0.216
coldcer	0.196	0.356	hhclean	0.184	0.360
toitisu	0.156	0.178	fzdin	0.130	0.145
diapers	0.222	0.245	paptowl	0.120	0.154
			Display		
mustketc	0.247	0.269	beer	0.230	-
mayo	0.195	0.219	fzpizza	0.191	0.200
hotdog	0.186	0.235	shamp	0.185	0.229
coffee	0.181	0.197	toothpa	0.181	0.216
fzdin	0.171	0.177	peanbutr	0.176	0.197
soup	0.170	0.174	paptowl	0.169	0.170
factiss	0.163	0.161	laundet	0.160	0.179
toitisu	0.158	0.180	coldcer	0.133	0.243
hhclean	0.123	0.233			

all significant coefficients with absolute size > 0.150 in $A1$ or $A5$;
- indicates insignificance.

We obtain the highest correlation for toitisu & paptowl (0.489). Other correlations greater than 0.300 are found for the category pairs toitisu & factiss, musketc & mayo, shamp & deod, laundet & hhclean, paptowl & laundet, toitisu & laundet, and paptowl & factiss. To give an example, a positive demand shock associated with higher utilities of the two categories toitisu and factiss might be triggered by a household's decision to jointly purchase personal care items.

$A5$ restricts about 73% of error correlations to zero because it assigns the two categories involved to different groups. In addition, about 22% of error correlations are lower (including insignificant correlations) according to partition $A5$.

V. CONCLUSION

The models presented here can be used by retail managers to decide which product categories are appropriate for features and displays. In addition, management can on the basis of these models predict sales caused by these marketing decisions. Preliminary results suggest that the most accurate model A1 is preferable if management wants to predict sales. On the other hand, if managers only want to select categories for features and displays and are not interested in sales forecasts, even the models for partition A5 do a satisfactory job.

Dividing 25 product categories between two and five groups leads to worse statistical performance compared to the most complex model which treats all 25 categories as one group. Of course, one cannot rule out the possibility that other partitions than the ones investigated here (which are typical of those used by grocery retailers) could do better. Therefore, the next step of this work consists in determining post hoc

 ISBN: 978-1-61208-423-7

partitions with different numbers of category groups by a stochastic search algorithm drawing upon work of Hoeting *et al.* [11].

TABLE V. SELECTED ERROR CORRELATIONS

Category pair		Partition $A1$	Partition $A5$	Category pair		Partition $A1$	Partition $A5$
toitisu	paptowl	0.489	0	toitisu	factiss	0.336	0.285
mustketc	mayo	0.341	0.329	shamp	deod	0.330	0.176
laundet	hhclean	0.320	0.119	paptowl	laundet	0.311	0.150
toitisu	laundet	0.314	0	paptowl	factiss	0.310	0
saltsnck	carbbev	0.298	0	toitisu	shamp	0.279	0.167
paptowl	hhclean	0.255	-	yogurt	coldcer	0.274	0.092
toothpa	shamp	0.288	0.290	toothpa	deod	0.276	-
fzpizza	fzdin	0.311	0.243	toitisu	deod	0.238	-
shamp	paptowl	0.226	0	spagsauc	coldcer	0.218	0
toothpa	laundet	0.241	0	shamp	laundet	0.233	0
toitisu	hhclean	0.209	0	hhclean	factiss	0.223	0
toitisu	coffee	0.229	0	peanbutr	coldcer	0.219	0
spagsauc	soup	0.215	0	toothpa	toitisu	0.211	0.137
toitisu	margbutr	0.202	0	paptowl	deod	0.223	0
saltsnck	fzpizza	0.205	0.163	margbutr	hhclean	0.186	0
soup	margbutr	0.209	0	toitisu	saltsnck	0.199	0
paptowl	margbutr	0.200	0	mustketc	hotdog	0.191	0
paptowl	coffee	0.200	0	shamp	hhclean	0.227	0
spagsauc	mustketc	0.203	0.207	toothpa	paptowl	0.180	0

all significant correlations with absolute size $>= 0.200$ in $A1$ or $A5$;
- indicates insignificance, 0 that the error correlation is restricted to zero.

Such an approach would simultaneously estimate model parameters, assign categories to groups and households to clusters. Forming category partitions and clustering households would all be directly related to the overall statistical performance of the models. To our knowledge, such an integrated approach has not been attempted in a previous publication.

REFERENCES

[1] A. Ansari and C. F. Mela, "E-Customization," Journal of Marketing Research, vol. 40, 2003, pp. 131–145.

[2] J. Barnard, R. McCulloch, and X. L. Meng, "Modeling Covariance Matrices in Terms of Standard Deviations and Correlations with Application to Shrinkage," Statistica Sinica, vol. 10, 2000, pp. 1281-1311.

[3] Y. Boztuğ and L. Hildebrandt, "Modeling Joint Purchases with a Multivariate MNL Approach," Schmalenbach Business Review, vol. 60, 2008, pp. 400–422.

[4] Y. Boztuğ and T. Reutterer, "A Combined Approach for Segment-Specific Market Basket Analysis," European Journal of Operational Research, vol. 187, 2008, pp. 294–312.

[5] B. J. Bronnenberg, M. W. Kruger, and C. F. Mela, "The IRI Marketing Data Set," Marketing Science, vol. 27, 2008, pp. 745–748.

[6] S. Chib and E. Greenberg, "Bayesian Analysis of Multivariate Probit Models.," Biometrika, vol. 85, 1998, pp. 347–361.

[7] S. Chib, P. B. Seetharaman, and A. Strijnev, Analysis of Multi-Category Purchase Incidence Decisions Using IRI Market Basket Data. JAI, Amsterdam, 2002, pp. 57–92, in Franses, P. H., Montgomery, A. L., Econometric Models in Marketing.

[8] K. Dippold and H. Hruschka, "A Model of Heterogeneous Multi-category Choice for Market Basket Analysis," Review of Marketing Science, vol. 11, 2013, pp. 1–31.

[9] S. D. Duvvuri, V. Ansari, and S. Gupta, "Consumers' Price Sensitivities across Complementary Categories," Management Science, vol. 53, 2007, pp. 1933–1945.

[10] A. E. Gelfand, Model Determination Using Sampling-Based Methods. Chapman & Hall,Boca Raton,1996, pp. 145-161, in Gilks, W. R., Richardson, S., Spiegelhalter, D. J., Markov Chain Monte Carlo in Practice.

[11] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian Model Averaging: A Tutorial," Statistical Science, vol. 14, 1999, pp. 382–417.

[12] X. Liu and M. J. Daniels, "A New Algorithm for Simulating a Correlation Matrix Based on Parameter Expansion and Reparameterization," Journal of Computational and Graphical Statistics, vol. 15, 2006, 897–914.

[13] P. Manchanda, A. Ansari, and S. Gupta, "The Shopping Basket: A Model for Multi-Category Purchase Incidence Decisions," Marketing Science, vol. 18, 1999, pp. 95–114.

[14] R. M. Neal, "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," Journal of Computational and Graphical Statistics, vol. 9, 2000, pp. 249–65.

[15] G. J. Russell and A. Petersen, "Analysis of Cross Category Dependence in Market Basket Selection," Journal of Retailing, vol. 76, 2000, pp. 369–392.

[16] I. Song and P. K. Chintagunta, "A Discrete-Continuous Model for Multicategory Purchase Behavior of Households," Journal of Marketing Research, vol. 44, 2007, pp. 595–612.

 ISBN: 978-1-61208-423-7

Measuring the Effects of Moods and Heart Rate Variability when Playing Video Games

Subtitle: Video Games Impact on HRV and Moods

Fang-Yu Pai
Department of Medical Informatics
Chung Shan Medical University
Taichung, Taiwan
E-mail: jonan0607@gmail.com

Ching-Hsiang Lai
Department of Medical Informatics
Chung Shan Medical University
Taichung, Taiwan
E-mail: liay@csmu.edu.tw

***Abstract*—In this study, we recruited 20 college students who are video games enthusiasts and assigned them to two groups, namely, video game playing (as a testing group), and book-reading (as a control group). We use profile of mood state (POMS) scale and electrocardiography (ECG) patch physiological signal device as tools to measure moods and heart rate variability (HRV) before and after the experiment, and we assessed the effects of playing video games on physical and mental health. We expect that the low frequency (LF) and high frequency (HF) powers for HRV in testing group to be significantly higher than the value in the control one after the experiment and to have 30 minutes duration in the following sleep stage. We show that playing video games may excite the sympathetic and parasympathetic systems during and after playing. This will disturb sleep and increase the risk of chronic diseases. Based on these findings, we recommend avoiding playing video games for extended periods of time.**

Keywords—video game; mood; heart rate variability (HRV); sympathetic; parasympathetic.

I. BACKGROUND

As computers and the Internet have become more popular, teenagers and students often fall vulnerable to cyber risks [1]. Researchers [2][3][4] who studied junior and senior high school students have found that those who indulged in the cyber world are more likely to confront health, academic or even family problems. Since video games is one of the key reasons behind cyber time, it has played a more dominant role even against sleep. Other studies point out that devoted gamers lose their temper easier, and often exhibit signs of aggression, absentmindedness, lack of discipline, low self-esteem, plummeting social skills, and anxiety toward the society.

A. Motivation

Compared to other stages of personal development, the teenage stage is not only a turning point in one's life, it also lays down a foundation for a durable healthy lifestyle. With a motive to promote health, this research is dedicated to look into the physical and psychological experiences of teenagers playing video games, and how the video games impact their physical and psychological well-being, self-esteem, and personal relationships. With this research, we assess the video game impacts on players' physical and emotional conditions, and try to re-evaluate the recreational and entertaining merits of video games.

B. Objectives

Objectives of this research are as follows:

- To use profile of mood state (POMS) scale to measure the changes of emotions before and after playing video games.
- To find the effects on the sympathetic and parasympathetic systems before and after playing video games exhibited in heart rate variability (HRV).
- To find the time duration of effects after playing video games.

The flowchart of this research is displayed below in Figure 1:

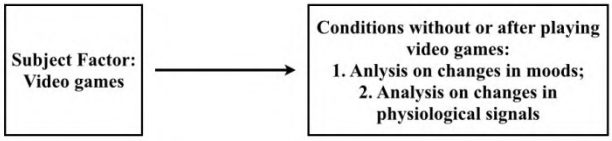


Figure 1. Research flowchart

C. Limitations

Given restrictions in time and manpower, this research is subject to limits and has a restricted scope, as follows:

- As subjects of this research are college students, and the desired trial period begins when the activity starts and ends when the person goes to sleep, only students residing in central Taiwan are selected for this research.
- The research was designed to take place when college students play video games, with the trial period between 9 and 12 pm, followed by a sleeping period of about five hours, for a total of around 9 hours.
- Given limitations in time, budget and manpower, this research consists of 20 subjects. Each subject is assigned to video game playing group once (as a testing group), and motionless book-reading group (as a control group), and for a total of sample collected data of 40.

 ISBN: 978-1-61208-423-7

II. RESEARCH STEPS

This research used POMS scale and electrocardiography (ECG) patch to measure moods and HRV for effects of playing video games on mental and physical health. The processes of the experiment related to the testing and control groups are shown in Figure 2 and Figure 3, respectively, with each group undertaking three steps, as follows:

A. Testing Group

- Subjects fill out a questionnaire on POMS, before they do the ECG device for the trial.
- After a ten-minute recess, subjects proceed into one-hour video game playing, and they repeat the entire process consisting of recess and game playing.
- When the second-round of game-playing is finished, the subjects take another ten-minute recess and then fill out the POMS for the second time. After another ten-minute recess, they sleep.

B. Control Group

- Subjects fill out two questionnaires: Questionnaire on Internet Addiction and POMS, before they do the ECG device for the trial.
- After a ten-minute recess, subjects proceed into a one-hour sedentary reading, and repeat the entire process consisting of recess and reading.
- When the second-round sedentary reading is finished, the subjects take another ten-minute recess and then fill out the POMS for the second time. After another ten-minute recess, they sleep.

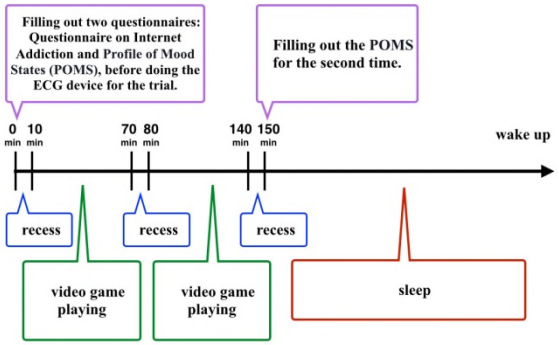


Figure 2. Steps undertaken by the testing group.

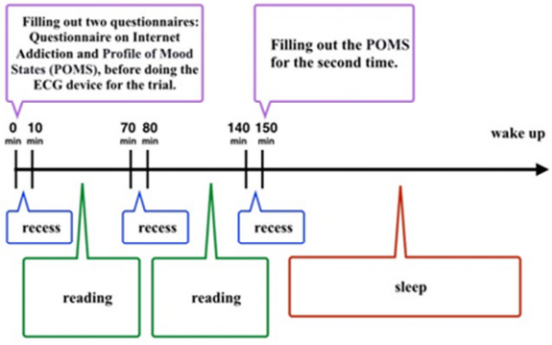


Figure 3. Steps undertaken by the control group.

III. RESEARCH METHODS

The physiological signals obtained from the ECG patch are digitalized into figures, before being analyzed by SPSS statistical software. Analysis are given on the physiological signals generated from the video game playing/book-reading of the testing and control groups, and including all the sleep stage.

The independent sample *t* test is used to compare the PMOS data and physiological signals before video game playing/book-reading of two group to ensure absence of discrepancy. However, paired-*t* test is used to test the data before and after experiment for each group to examine the effect of video game playing/book-reading.

IV. PROJECTED RESULTS

Our projected results are shown in Figure 4, Figure 5, Figure 6 and Figure 7:

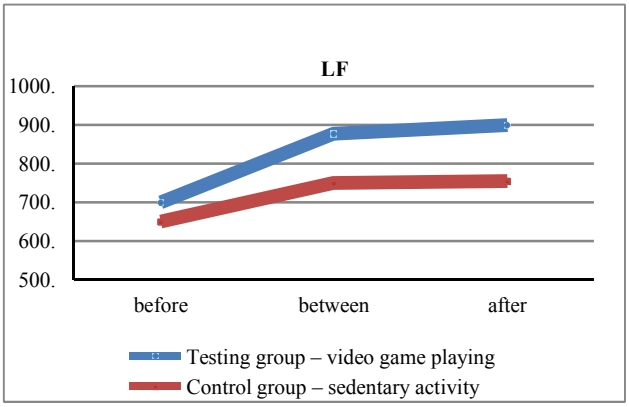


Figure 4. LF collected from the testing and control stages before, in between and after the video game playing and sedentary activity.

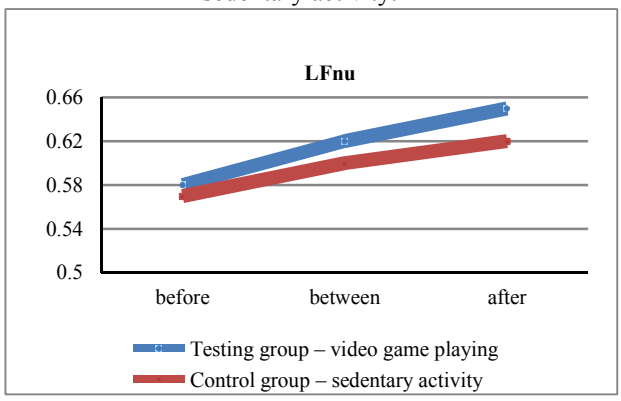


Figure 5. LFnu collected from the testing and control stages before, in between and after the video game playing and sedentary activity.

 ISBN: 978-1-61208-423-7

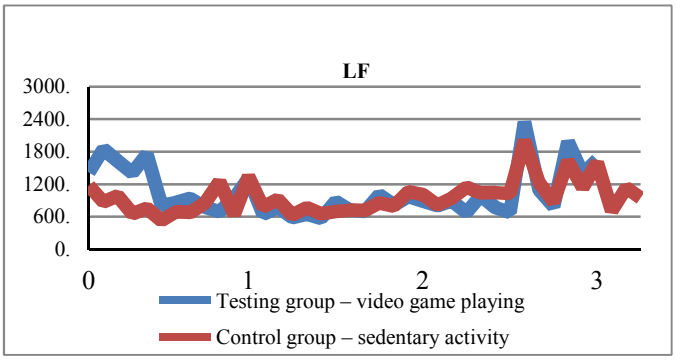


Figure 6. LF collected during the sleep of the testing and control stages.

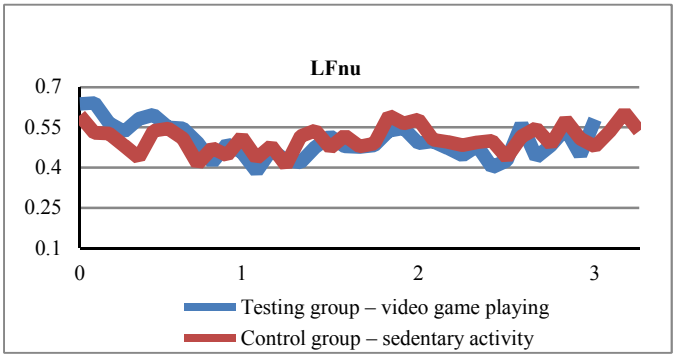


Figure 7. LFnu collected during the sleep of the testing and control stages.

V. CONCLUSION

The conclusion inferred from this research is the following:

- Based on the PMOS data collected before and after video game playing, the scores for fatigue of the testing stage increase notably, while the scores for vigor fall significantly. This dramatic change in scores is absent in the PMOS data of the control stage. It is concluded that after long-time video game playing, players start to feel higher level of fatigue and lower level of vigor.
- The physiological signals collected from the testing stage are not significantly different from that of the control stage. Still, the testing stage generates stronger physiological signals like LF and Lfnu, against the signals generated by the control stage. This suggests that, after long-time video game playing, players start to experience sympathetic and parasympathetic elevations.
- Within a thirty minutes period before sleeping, the testing stage generate higher LF, Lfnu and HF against the control stage. The assenting LFnu implies elevation of the sympathetic and parasympathetic energies, of which the former imposes greater impact on subjects. We thus can conclude that after playing the video games, players start to experience temporary sympathetic elevations that prevent sleeping or leads to poor sleeping quality, and consequentially impose detrimental impacts on the physical well-being of players. That is why video game players should avoid playing video games for a long time or in the night so that their physical and mental well-being can be better protected.

REFERENCES

[1] Kanwal Nalwa and Archana Preet Anand, "Internetaddictionin students: a cause of concern," in CyberPsychology & Behavior, vol. 6, 2003, pp. 653-656.

[2] Chih-Hung Ko, Ju-Yu Yen, Cheng-Fang Yen, Huang-Chi Lin, and Ming-Jen Yang, "Factors predictive for incidence and remission of internet addiction in young adolescents: a prospective study," in CyberPsychology & Behavior, vol. 10, 2007, pp. 545-551.

[3] Sunny S.J Lin, and Chin-Chung Tsai, "Sensation seeking and internet dependence of Taiwanese high school adolescents," in Computers in Human Behavior, vol. 18, 2002, pp. 411-426..

[4] Shu Ching Yang, and Chieh-Ju Tung, "Comparison of Internet addicts and non-addicts in Taiwanese high school," in Computers in Human Behavior, vol. 23, 2007, pp. 79-96.

 ISBN: 978-1-61208-423-7

A Modified Multi-objective Differential Evolution Algorithm with Application in Reinsurance Analytics

Omar Andres Carmona Cortes

Instituto Federal do Maranhão
Informatics Department
São Luis, MA, Brazil
Email: omar@ifma.edu.br

Andrew Rau-Chaplin

Dalhousie University
Faculty of Computer Science
Halifax, NS, Canada
Email: arc@cs.dal.ca

***Abstract*—In the reinsurance marketplace, the risk of financial loss in the event of natural catastrophes (such as earthquakes, hurricanes and floods) is exchanged between market participants for a premium. Here, prudent risk management takes the form of a hedge against the risk of a contingent uncertain loss in exchange for a payment. Reinsurance contracts that define the terms of the transfer are elaborated multi-layered financial treaties that represent complex trade-offs between expected return and risk. Formulating an effective risk transfer strategy depends on a careful multi-objective optimization process. In this paper, we study from the perspective of an insurance company the Reinsurance Contract Optimization problem in which, given the structure of a multi-layered reinsurance contract, we are required to discover specific contractual terms that capture the best trade-offs between expected return and risk for the insurer. Our approach is based on an adaptation of Multi-Objective Differential Evolution. In searching for the best mutation operators, we performed an experimental analysis on large-scale real problem instances using industrial datasets and evaluated five different mutation operators. Our experimental results indicate that those mutation operators based on selecting non-dominated individuals from the archive tend to produce better outcomes. Since speed is critical in this application, we also developed a parallel version achieving a speedup up to 9.3 on a 16 core machine.**

Keywords–Risk Analytics; Differential Evolution; Multi-objective; Parallel Computing.

I. INTRODUCTION

Many real world applications involve the optimization of two or more conflicting objectives, where the search for solutions generates a Pareto frontier [1], on which no particular solution is better than another. Solutions on the Pareto frontier represent trade-offs between multiple objectives. Given these trade-offs, human experts can select final solutions based on other, often more qualitative, criteria.

An interesting real world application in computational finance is the Reinsurance Contract Optimization (RCO) problem. RCO is a treaty optimization problem in which we are given as input the structure of a complex risk transfer treaty consisting of a fixed number of contractual layers, a simulated set of expected loss distributions (one per layer), and a model of reinsurance market costs [2][3]. The task is to generate as output the best set of shares, a key financial parameter, which balance the expected return and the risk. In this context, typical risk measures include variance, Value at Risk (VaR) or a Tail-Value at Risk (TVaR) [4].

An enumeration method can be used for solving the RCO problem; however this approach presents two main problems: 1) it has to be discretized, demanding some changes in numerical algorithms and 2) in practice, it is only applicable to small problems instances (i.e., 2 to 4 layers), whereas real instances of the RCO problem can have 7 or more layers. For instance, a 7 layered problem can take several week to be solved with a 5% level of discretization on the search space using the enumeration method as presented in [2]. As a consequence, it is important to explore alternative methods for addressing this type of problem.

In this context, evolutionary algorithms, such as differential evolution (DE) [5], seem to be a natural choice. DE is reasonably simple to implement and has been successfully used in many applications including Reservoir System Optimization [6], Communication Systems [7], and Speaker Recognition [8]. Risk and reinsurance problems have also been tackled using evolutionary methods, such as in [9][10][11]; however, the focus in these particular applications was on stop loss and ruin prediction, *i.e.*, a very different problem than the RCO problem studied in this paper. In the context of RCO, the first studies using evolutionary methods were [2] and [3].

While the techniques proposed in these papers performed significantly better than the enumeration approach, they suffered from a critical drawback. They were based on single-objective optimization methods in which the risk could be optimized only for a given expected return value in any one call to the optimizer. Consequently, creating a Pareto frontier that covers a range of expected return values was very time-consuming, making it unsuitable for many industrial scale problems. In [12], a faster vector evaluated differential evolution method for RCO was presented. While this approach was multi-objective, producing the whole Pareto frontier at once, it suffered from solution quality issues in that it often produced Pareto frontiers with holes or large gaps between solutions especially in the critical middle portion of the curve.

In this paper, we present a modified Differential Evolution for Multi-objective Optimization (DEMO) [13] algorithm - simpler than Multi-Objective Differential Evolution Algorithm (MODEA) [14] - which is both fast and solves the previously noted gaps problem, thus producing high quality Pareto frontiers. Our approach uses an archive of previously identified solutions in order to avoid losing non-dominated candidates as the optimization converges. Solutions were lost when the

 ISBN: 978-1-61208-423-7

number of non-dominated solutions is truncated by crowding distance when it is bigger than the population size in the original version. Additionally, we present a study of different mutation operators, and propose the use of a non-dominated solution in the mutation step, instead of any random individual. An experimental evaluation of our modified DEMO method applied to the RCO problem demonstrates that it can solve extremely large real-world RCO problems with between 7 and 15 layers (subcontracts) in under three minutes. Our experimental results indicate that the quality of solutions, when evaluated in terms of the average size and hyper volume of the generated Pareto frontiers, is high and the previously noted gaps problem, especially in the critical middle portion of the curve, is now largely absent.

The remainder of this paper is structured as follows: Section II presents fundamentals concepts of multi-objective problems and an introduction to the RCO problem; Section III shows how DEMO works and our proposal; Section IV introduces the metrics that were applied and the results, including some parallelization features; finally, Section V presents conclusion and future work.

II. MULTIOBJECTIVE PROBLEMS

A Multi-objective Optimization Problem (MOP) has to address two or more conflicting objective function [15] at the same time. The resulting solution is a Pareto frontier, *i.e.*, a set of points where no solution is better than another one. Otherwise, the global optima would be only one point in the search space [12]. Thus, assuming that a solution to a MOP is a vector in a search space X with m elements. A function $f : X \rightarrow Y$ evaluates the quality of a solutions mapping it into an objective space. Therefore, a multi-objective problem is defined as presented in (1), where f is a vector of objective functions, m is the dimension of the problem and n represents the number of objective functions.

$$Max\ y = f(x) = (f_1(x_1, ..., x_m), ..., f_n(x_1, ..., x_m)) \quad (1)$$

In order to determine whether a solution belongs to the Pareto frontier or not, it is necessary to use the concept of optimality (i.e., Pareto dominance), which states that given two vectors x, $x* \in \Re$ and $x \neq x*$, x dominates $x*$ (denoted by $x \succeq x*$) if $f_i(x)$ is not worse than $f_i(x*), \forall i$ and there exist at least one i where $f_i(x) > f_i(x*)$ in maximization cases and $f_i(x) < f_i(x*)$ otherwise. Hence, a solution x is said Pareto optimal if there is no solution that dominates x, in such case, x is called non-dominated solution. Mathematically, assuming a set of non-dominated solutions $\wp$, a Pareto frontier(pf) is represented as $pf = \{f_i(x) \in \mathbb{R} | x \in \wp\}$.

A. A Treaty Optimization Problem: RCO

Insurance organizations, with the help of the global reinsurance market, look to hedge their risk against potentially large claims, or losses [4]. This transfer of risk is done in a manner similar to how a consumer cedes part of the risk associated with their private holdings [2]. The claims received by the insurer in case of a natural catastrophe are also referred to as expected return.

The reinsurance contract optimization consists of a fixed number of contractual layers and a simulated set of expected loss distributions (one per layer), plus a model of reinsurance market costs [2]. Hence, the main task is to discover the best combination of shares, also known as placements, which leads to a set of trade-offs between expected return and risk. In other words, insurance companies aim to hedge their risk against potentially large claims, or losses [4], especially those ones resulting from natural catastrophes. When these trade-offs are set, the insurance companies are able to offer them to the reinsurance market.

Overall, the purpose is both to maximize the amount of return ($) received from the reinsurance company and maximize the risk transferred to it. Doing so, the insurance companies minimize the loss faced per year in case of a natural disaster. In this context, (2) represents the problem in terms of optimization, where VaR is a risk metric, $\mathbf{R}$ is a function in term of placements (π) and E is the Expected Value. In probability theory, the expected value, usually denoted by E[X], refers to the value of a random variable X that we would "expect" to find out if we could repeat the random variable process an infinite number of times and take the average of the values obtained. For further details about the problems, refer to [2] and [4].

$$\begin{array}{ll} maximize & f_1(x) = VaR_\alpha(\mathbf{R}(\pi)) \\ maximize & f_2(x) = E[\mathbf{R}(\pi)] \end{array} \quad (2)$$

III. DIFFERENTIAL EVOLUTION MULTI-OBJECTIVE (DEMO)

The DEMO algorithm is shown in Figure 1, where we can observe that it is similar to the canonical version of DE whose strategy is DE/Rand/1 [16]. The differences start in line 16 when the new population is selected for the next iteration. Thus, if a new individual ($indiv$) dominates the target one (Pop_i) then the new one is added into a new population; if the target individual dominates the new one then the target element is added into the new population; otherwise, both individuals go to the new population. The dominance process builds a new population whose size rages from pop_size to $2 \times pop\_size$. Finally, if the size of the new population is larger than pop_size then the new individuals which go to next iteration are selected by crowding distance ($select\_cdistance$ function).

A. Our Proposal

The main drawback of the original DEMO was not maintaining an archive, thereby loosing good solutions when the number of non-dominated points overcomes the size of the population. Taking this into account, we changed the original algorithm in two parts. Firstly, we introduce an archive in the algorithm (after line 31), which is maintained on each iteration in order to do not lose non-dominated solutions from one iteration to another due to the crowding distance algorithm in line 30. Secondly, we tested some mutation operators (line 6 in the Figure 1) as presented in (4), (5), (6), and (7). Unless the original mutation operator which uses three any random individuals in order to build the F vector, in (4), we uses a random individual from the set of non-dominated ones. Thus, it is necessary to compute the non-dominated set between lines 3 and 4, *i.e.*, before starting the loop which deals with the population. (5) is similar to the previous one; however, F is a random number between 0 and 1. Then, in (6) and (7), we

```
1  Pop ← generate_pop(n, d)
2  fit ← evaluate(Pop)
3  while (Stop Criteria is FALSE) do
4      for i = 1 to #pop_size do
5          idx ← select_indiv(3)
6          v ← Pop_idx3 + F * (Pop_idx1 − Pop_idx2)
7          for j = 1 to dimension do
8              nj = rand()
9              if (nj < CR) then
10                 indiv← v_j
11             else
12                 indiv← Pop_i j
13             end
14         end
15         fit′ ← evaluate(indiv)
16         if (fit′ dominates fit_i) then
17             pop′ ← indiv
18             nf ← fit′
19         else if (fit_i dominates fit′) then
20             pop′ ← Pop_i
21             nf ← fit_i
22         else
23             add indiv and Pop_i into pop′
24             add fit and fit′ into nf
25         end
26     end
27     if (size(pop′) == size(Pop_i))) then
28         Pop ← pop′
29     else
30         [Pop, fit] ← select_cdistance(pop′, nf)
31     end
32 end
```

Figure 1. DEMO Algorithm

randomly pick up the first individual from the archive which currently contains the best solutions found by the algorithms. The difference between the last two equations is the use of F which is randomly chosen in (7).

$$v \leftarrow non\_dom_{idx} + F * (Pop_{idx_1} - Pop_{idx_2}) \quad (3)$$
$$v \leftarrow non\_dom_{idx} + Rand() * (Pop_{idx_1} - Pop_{idx_2}) \quad (4)$$
$$v \leftarrow archive_{idx} + F * (Pop_{idx_1} - Pop_{idx_2}) \quad (5)$$
$$v \leftarrow archive_{idx} + Rand() * (Pop_{idx_1} - Pop_{idx_2}) \quad (6)$$
$$(7)$$

In the next section, the mutation operators will be referred to as M1 (canonical mutation), M2 (4), M3 (5), M4 (6), and M5 (7).

IV. Computational Experiments

All tests were conducted using R version 2.15.0 and RStudio on a Windows 7 64-bit Operating System running on an Intel i7 3.4 Ghz processor with 4 physical cores and hyper threading, with 16 GB of RAM. We executed the parallel version in an Intel Xeon comprising of two Xeon processors E5-2650 running at 2.0 Ghz with 8 cores and hyper threading and 256 GB of memory. The experiments used $F = 0.7$ or a random F, and $CR = 0.9$ considering 250, 500 and 1000 iterations with a population size equals to 50. Further, all averages are calculated in 30 trials. Our data set is composed by 7 layers of real anonymized data. The 15 layers data set was synthetically created based on the 7 layers one.

A. Metrics

In this section, we discuss the experimental evaluation of MODE algorithm. Firstly, the average number of non-dominated points (number of solutions) found in the Pareto frontier was determined. Secondly, the average hyper volume, which is the volume of the dominated portion of the objective space as presented in (8), was measured, where for each solution $i \in Q$ a hypercube v_i is constructed. The extreme points are those one belonging to the Pareto front. Having each v_i, we calculated the final hyper volume by the union of all v_i. The final number of solutions after all trials is showed as well.

$$hv = volume(\bigcup_{i=1}^{|Q|} v_i) \quad (8)$$

Thirdly, the dominance relationship between Pareto frontiers obtained with different mutation operators was calculated as depicted in (9). Roughly speaking, $C(A, B)$ is the percentage of the solutions in B that are dominated by at least 1 solution in A [17], therefore, if $C(A, B) = 1$ then all solutions in A dominate B, and $C(A, B) = 0$ means the opposite. It is important to notice that this metric is neither complementary by itself nor symmetric, *i.e*, $C(A, B) \neq 1 - C(B, A)$ and $C(A, B) \neq C(B, A)$ making important to compute it in both direction: $C(A, B)$ and $C(B, A)$. Finally, the resulting frontiers can be reviewed by experts for reasonability. For further details about the use of these metrics see [15].

$$C(A, B) = \frac{|\{b \in B | \exists a \in A : a \preceq b\}|}{|B|} \quad (9)$$

In terms of parallelism, we calculated the speedup according to $speedup = \frac{T_s}{T_p}$, where T_s is the execution time considering one thread and T_p represents the time in parallel using p threads. This kind of metric is called weak speedup and it was suggested in Alba [18] because the code is exactly the same regardless the number of threads, thus it is not necessary to guarantee that the serial version is the best one.

B. Results

Table I presents the average of number of solutions, the hypervolume, the elapsed time and the final number of solutions for 7 layers. As we can observe, using non-dominated individuals either from the population or the archive tend to produce more number of solutions; however, better hypervolumes are obtained by the version using the archive. Also, the difference in terms of time is not significant between mutation operators. Figure 2 shows the final Pareto frontier for 7 layers where we can noticed that visually it is not possible to identify which mutation operator is the best for this particular application.

Table II shows the coverage metric between the different mutation operators where it noticeable that the canonical mutation M1 dominates only between 5% and less than 1% of the other approaches. On the other hand, M5 dominates more

 ISBN: 978-1-61208-423-7

TABLE I. METRICS FOR 7 LAYERS AND 250 ITERATIONS

	#NS	Hypervolume	Time	Final #NS
M1	83.73	2.19E+15	105.18	226
Stdev	5.59	9.40E+13	3.40	
M2	151.30	1.96E+15	104.50	339
Stdev	12.91	3.52E+14	2.35	
M3	170.77	1.80E+15	103.41	330
Stdev	13.95	3.16E+14	2.60	
M4	203.67	2.23E+15	105.08	374
Stdev	10.80	2.60E+14	2.19	
M5	232.93	2.21E+15	105.63	383
Stdev	11.44	2.43E+14	2.13	

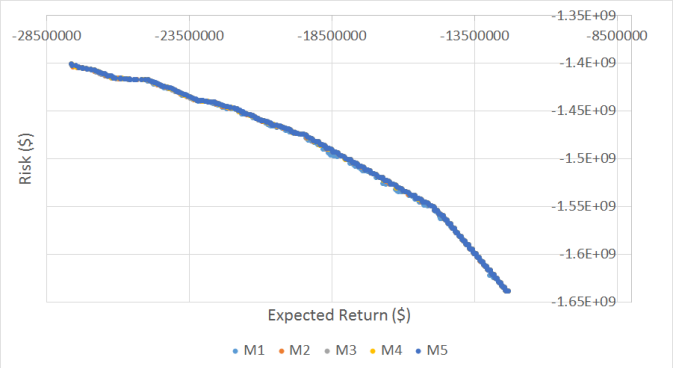


Figure 2. Final Pareto frontier for 7 layers and 250 iterations

TABLE III. METRICS FOR 15 LAYERS AND 250 ITERATIONS

	#NS	Hypervolume	Time	Final #NS
M1	55.23	2.96E+15	166.66	104
Stdev	4.70	2.87E+14	5.61	
M2	93.90	2.54E+15	165.20	220
Stdev	8.86	5.83E+14	5.81	
M3	139.90	2.39E+15	166.32	354
Stdev	15.20	6.38E+14	2.77	
M4	145.23	3.40E+15	166.43	340
Stdev	10.12	8.53E+14	4.41	
M5	188.67	3.05E+15	168.57	390
Stdev	12.71	6.81E+14	2.38	

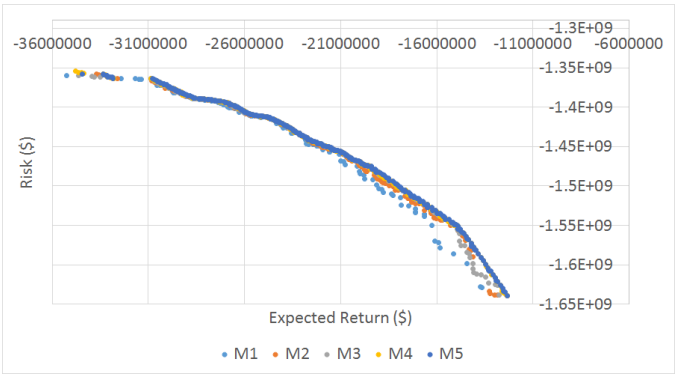


Figure 3. Final Pareto frontier for 15 layers and 250 iterations

solutions from the other operators. For example, M5 dominates 78% of solutions from M1 and 30% of solutions from M4. On the other hand, M4 dominates only 9% of M5 solutions, demonstrating that M5 is the most effective in this case.

TABLE II. COVERAGE FOR 7 LAYERS AND 250 ITERATIONS

	M1	M2	M3	M4	M5
M1	-	0.05	0.06	0.03	0.007
M2	0.66	-	0.19	0.098	0.034
M3	0.66	0.30	-	0.14	0.06
M4	0.73	0.39	0.28	-	0.09
M5	0.78	0.48	0.31	0.30	-

Table III illustrates the same metrics aforementioned for 250 iterations and 15 layers. In this case, the behavior is similar to the previous one in terms of number of solution and time; however, M4 presented the best hypervolume. Moreover, visually there are some differences between the performance of the operators as we can see in Figure 3, where M4 and M5 seem to be better than the other operators. The coverage metric in Table IV indicates that M5 dominates more solutions than M4, therefore, the bigger hypervolume might be caused by the non-dominated points in the beginning of the Pareto frontier curve.

The behavior for 7 and 15 layers using 500 iterations are presented in Tables V and VII. As we can see, the results are similar to the previous ones including the final Pareto frontier (Figures 4 and 5) and the coverage rates in Tables VI and VIII; but, now the differences are in a smaller scale. This result is expected because as we increase the number of iterations the differences tend to be smaller. On the other hand, this number of iterations is not sufficient for approximating the results using 15 layers because this last one is a much harder problem to solve.

TABLE IV. COVERAGE FOR 15 LAYERS AND 250 ITERATIONS

	M1	M2	M3	M4	M5
M1	-	0.036	0.025	0.006	0.000
M2	0.88	-	0.087	0.07	0.036
M3	0.89	0.69	-	0.30	0.10
M4	0.95	0.80	0.42	-	0.11
M5	0.98	0.89	0.70	0.61	-

TABLE V. METRICS FOR 7 LAYERS AND 500 ITERATIONS

	#NS	Hypervolume	Time	Final #NS
M1	101.40	2.26E+15	209.48	237
Stdev	9.19	5.40E+13	4.18	
M2	182.43	2.20E+15	208.09	373
Stdev	11.74	2.01E+14	4.96	
M3	191.57	1.65E+15	205.97	367
Stdev	16.82	4.68E+14	4.88	
M4	240.77	2.15E+15	206.66	399
Stdev	10.36	2.75E+14	4.87	
M5	290.17	2.20E+15	207.13	397
Stdev	12.02	2.33E+14	5.59	

TABLE VI. COVERAGE FOR 7 LAYERS AND 500 ITERATIONS

	M1	M2	M3	M4	M5
M1	-	0.083	0.087	0.01	0.002
M2	0.616	-	0.19	0.085	0.015
M3	0.60	0.22	-	0.102	0.025
M4	0.718	0.365	0.29	-	0.053
M5	0.747	0.437	0.34	0.235	-

The results for 1000 iterations for 7 and 15 layers are presented in Tables IX and XI. Visually, the results for 7 layers in Figure 6 are the same; however, Table X indicates that the differences are still there. Looking at the number of the Pareto frontier, we will see similar solutions between M1 and M5; nonetheless, the M5 solutions dominates the

 ISBN: 978-1-61208-423-7

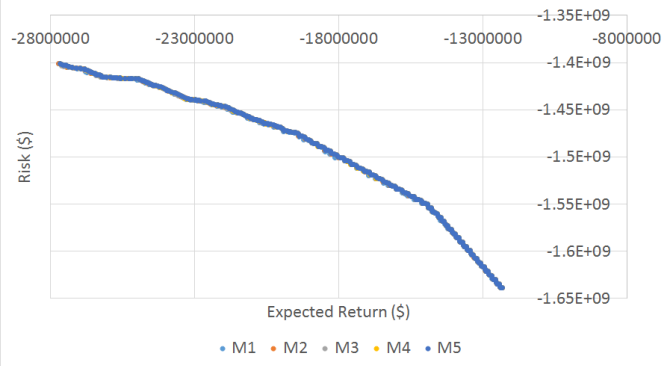


Figure 4. Final Pareto frontier for 7 layers and 500 iterations

TABLE VII. METRICS FOR 15 LAYERS AND 500 ITERATIONS

	#NS	Hypervolume	Time	Final #NS
M1	65.93	3.27E+15	332.18	139
Stdev	4.46	3.79E+14	13.08	
M2	116.67	2.87E+15	331.33	233
Stdev	11.43	6.70E+14	8.75	
M3	171.77	2.18E+15	328.35	358
Stdev	16.77	6.18E+14	8.24	
M4	191.33	3.71E+15	331.55	379
Stdev	12.99	6.90E+14	8.51	
M5	246.067	3.27E+15	329.20	460
Stdev	14.88	8.72E+14	10.02	

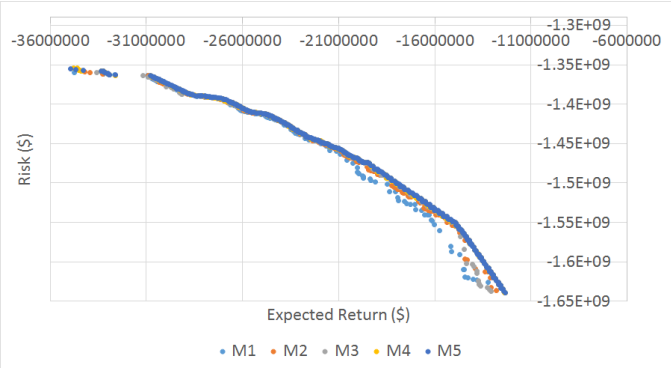


Figure 5. Final Pareto frontier for 15 layers and 500 iterations

TABLE VIII. COVERAGE FOR 15 LAYERS AND 500 ITERATIONS

	M1	M2	M3	M4	M5
M1	-	0.051	0.05	0.005	0.00
M2	0.777	-	0.154	0.0474	0.002
M3	0.77	0.70	-	0.184	0.046
M4	0.964	0.83	0.497	-	0.104
M5	0.99	0.897	0.706	0.547	-

solutions from M1 as shown by Table X. On the other hand, the differences between M4 and M5 are not so substantial because M5 dominates only 13% of solutions from M4 which represents that 87% of the solutions on both sets are either the same or non-dominated solutions.

When we move to 15 layers (Figure 7), a larger number of iterations do not create better solutions for M1. Also, more iterations do not significantly approximate M4 from M5 as we can see in Table XII where M5 dominates 49% of solutions from M4, whereas M4 dominates only 6.3% of solutions from M5. This behavior is a strong indication that M5 is the best operator for solving this problem.

TABLE IX. METRICS FOR 7 LAYERS AND 1000 ITERATIONS

	#NS	Hypervolume	Time	Final #NS
M1	122.83	2.32E+15	412.13	281
	6.80	2.05E+13	9.14	
M2	215.93	2.27E+15	411.85	400
	14.21	2.28E+14	11.16	
M3	199.73	1.88E+15	410.44	353
	20.62	3.75E+14	6.91	
M4	292.57	2.20E+15	413.74	404
	11.19	2.42E+14	8.87	
M5	333.27	2.19E+15	412.37	397
	14.25	2.59E+14	11.29	

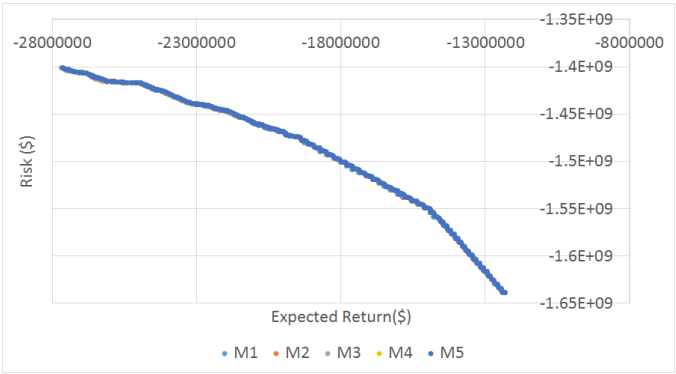


Figure 6. Final Pareto frontier for 7 layers and 1000 iterations

TABLE X. COVERAGE FOR 7 LAYERS AND 1000 ITERATIONS

	M1	M2	M3	M4	M5
M1	-	0.190	0.187	0.131	0.118
M2	0.665	-	0.136	0.029	0.005
M3	0.68	0.240	-	0.066	0.00
M4	0.79	0.305	0.21	-	0.00
M5	0.808	0.350	0.240	0.133	-

TABLE XI. METRICS FOR 15 LAYERS AND 1000 ITERATIONS

	#NS	Hypervolume	Time	Final #NS
M1	75.00	3.36E+15	655.36	147
	6.88	3.00E+14	26.82	
M2	143.43	3.22E+15	656.79	298
	12.18	6.30E+14	23.45	
M3	211.33	2.25E+15	662.09	361
	21.87	6.30E+14	10.68	
M4	242.53	3.67E+15	655.87	439
	13.17	7.48E+14	20.70	
M5	299.37	3.52E+15	679.67	507
	14.33	7.31E+14	29.95	

TABLE XII. COVERAGE FOR 15 LAYERS AND 1000 ITERATIONS

	M1	M2	M3	M4	M5
M1	-	0.067	0.017	0.006	0.006
M2	0.782	-	0.061	0.041	0.012
M3	0.857	0.775	-	0.273	0.065
M4	0.952	0.778	0.243	-	0.063
M5	0.972	0.88	0.46	0.49	-

 ISBN: 978-1-61208-423-7

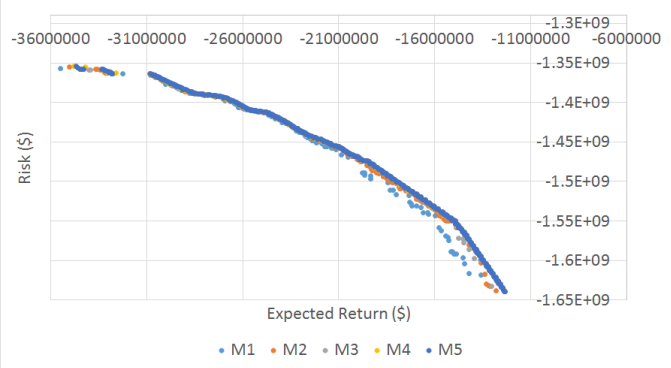


Figure 7. Final Pareto frontier for 15 layers and 1000 iterations

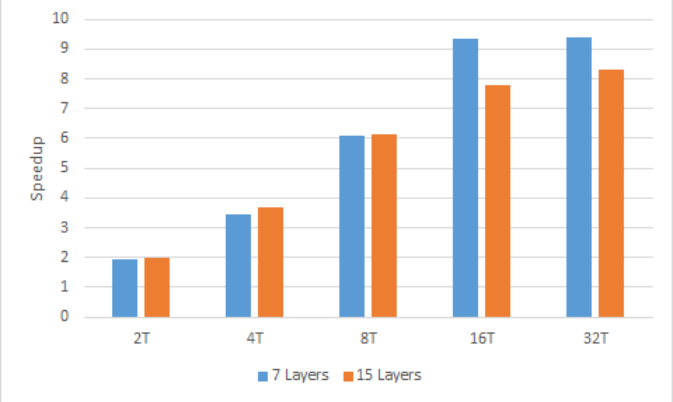


Figure 9. Speedup for 7 and 15 layers and 1000 iterations on Xeon

C. Parallel Version

In order to parallelized the code, we used the Snow package [19] from R which is a package for automatic parallelization. We parallelized the iteration loop using a *foreach* instruction associated with the parameter *%dopar%*. This parameter is responsible for dividing the iterations between threads. The main advantage of this approach is to maintain intact almost the entire code, being necessary to add instructions only for gathering results delivered by each thread. On the other hand, the main disadvantage lays in the fact that as we increase the number of threads the results tend to be worse in terms of quality.

The mutation operator we used is the M5 with 1000 iterations because it presented better results than the other ones. Figures 8 and 9 show the time and speedup reached in the Xeon architecture variating the thread count. Regardless the number of layers, the best efficiency is reached using 2 thread representing an efficiency of 96.7% and 98.2%, respectively. In terms of speedup, it is almost linear up to 4 threads. Then, the best one is reached using 32 threads representing 9.38 and 8.33 for 7 and 15 layers, respectively; however, the use of 32 threads represents an efficiency of 29.3% and 26% for 7 and 15 layers. Moreover, the best speedups are reached by 7 layers saturating in approximately 16 threads.

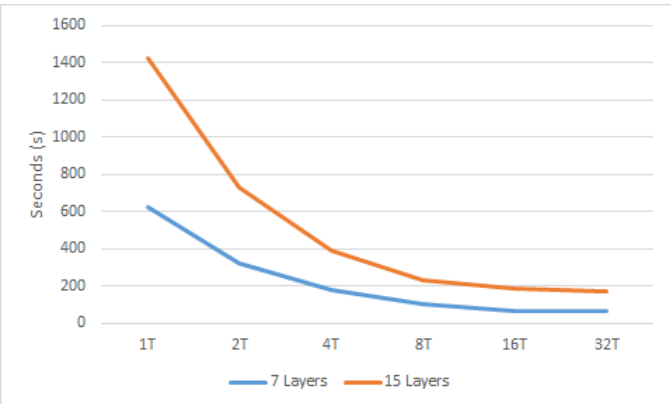


Figure 8. Time for 7 and 15 layers and 1000 iterations on Xeon

Figure 10 presents the Pareto frontier obtained by varying the thread count for 1000 iteration and 7 layers, where we can observe that visually all Pareto frontiers seem to be the same. Table XIII depicts the averages in term of metrics. Even though, the number of solutions decrease as we increase the number of threads, the final number of solutions is not affected. Moreover, the hypervolume is quite stable between threads; therefore, the faster the execution the better. In fact, the small numbers in Table XIV, which represent the coverage, mean that the Pareto frontiers are very similar regardless the number of threads.

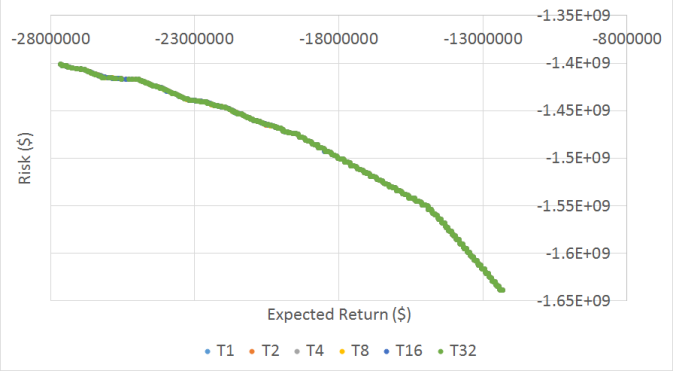


Figure 10. Pareto frontier varying thread count for 1000 iteration and 7 layers

TABLE XIII. METRICS FOR 7 LAYERS AND 1000 ITERATIONS

	#NS	Hypervolume	Time	#NS Final
1T	337.3666667	2.25E+15	626.3866667	403
	12.60673967	2.14E+14	3.61856901	
2T	336.6333333	2.30E+15	323.8888	398
	11.60999371	1.59E+14	1.890259066	
4T	329.6333333	2.34E+15	181.6333667	390
	10.49296425	1.05E+14	0.808078286	
8T	315.8666667	2.35E+15	103.0467333	406
	12.01359383	1.28E+13	0.340365719	
16T	288.9666667	2.35E+15	67.123	403
	9.86628999	3.00E+13	0.711079801	
32T	246.9	2.35E+15	66.7494	390
	13.47628824	1.45E+13	2.711422067	

Figure 11 shows the Pareto frontier obtained by varying the thread count for 1000 iteration and 15 layers, where we can observe that visually the difference between Pareto frontiers obtained by different counting of threads is not meaningful.

 ISBN: 978-1-61208-423-7

TABLE XIV. COVERAGE FOR 7 LAYERS AND 1000 ITERATIONS

	T1	T2	T4	T8	T16	T32
T1	-	0.028	0.03	0.08	0.16	0.20
T2	0.04	-	0.04	0.09	0.16	0.215
T4	0.03	0.03	-	0.07	0.14	0.20
T8	0.030	0.035	0.028	-	0.14	0.17
T16	0.019	0.015	0.015	0.057	-	0.16
T32	0.017	0.022	0.026	0.02	0.086	-

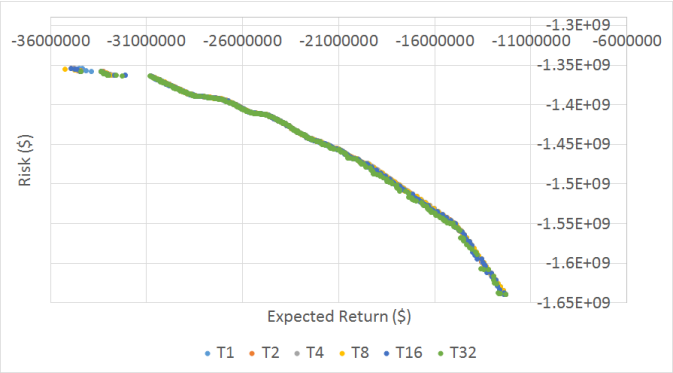


Figure 11. Pareto frontier varying thread count for 1000 iteration and 15 layers

TABLE XV. METRICS FOR 15 LAYERS AND 1000 ITERATIONS

	#NS	Hypervolume	Time	#NS Final
1T	290.73	3.39E+15	1426.90	517
	22.97	8.93E+14	6.33	
2T	296.03	3.82E+15	726.32	515
	15.83	7.22E+14	2.16	
4T	280.37	4.03E+15	387.39	470
	12.70	5.74E+14	1.13	
8T	237.40	4.22E+15	232.96	450
	13.63	3.72E+14	1.12	
16T	201.00	4.12E+15	182.87	378
	13.43	3.30E+14	9.25	
32T	164.00	3.99E+15	171.36	336
	12.71	4.05E+14	16.34	

TABLE XVI. COVERAGE FOR 15 LAYERS AND 1000 ITERATIONS

	T1	T2	T4	T8	T16	T32
T1	-	0.50	0.57	0.68	0.79	0.87
T2	0.40	-	0.32	0.54	0.65	0.77
T4	0.30	0.14	-	0.46	0.61	0.75
T8	0.20	0.09	0.15	-	0.52	0.70
T16	0.13	0.05	0.10	0.16	NA	0.55
T32	0.059	0.017	0.04	0.08	0.20	NA

Table XV presents data showing the relationship between the number of solutions and the number of threads used. As we increase the number of threads the number of solutions decreases. Overall, 8 threads seems to be a sweet spot due to the size of the associated hyper volume combined with the reduction in running time. Table XVI reinforces this view in that the 8 thread solution dominates 52% and 70% of solutions using 16 and 32 threads, respectively.

V. CONCLUSION

This paper presented a modified version of a DE algorithm for multi-objective algorithms with application in reinsurance analytics. Five mutations operators were tested. Results indicated that the best one is called M5, where the first element of the mutation operator is chosen from the archive. Parallel speedup experiments were performed on a Xeon based multi-core machines achieving a speedup of 9.38 using 32 threads.

ACKNOWLEDGMENT

The authors would like to thank NSERC, CNPq and IFMA for funding this research.

REFERENCES

[1] J. Branke, K. Deb, K. Miettinen, and R. Sowiski, "Introduction to Multiobjective Optimization: Interactive and Evolutionary Approaches", Lecture Notes in Computer Science (LNCS), vol. 5252, 2008.

[2] O. A. C. Cortes, A. Rau-Chaplin, D. Wilson, I. Cook, and J. Gaiser-Porter, "Efcient Optimization of Reinsurance Contracts using Discretized PBIL", DATA ANALYTICS, Porto-Portugal, 2013, pp. 18–24.

[3] O. A. C. Cortes, A. Rau-Chaplin, D. Wilson, and J. Gaiser-Porter, "On PBIL, DE and PSO for Optimization ofReinsurance Contracts", EvoStar, EvoFin, LNCS, Barcelona, 2014, pp. 227–238.

[4] J. Cai, K. N. Tan, C. Weng, and Y. Zhang, "Optimal reinsurance under VaR and CTE risk measures". Insurance: Mathematics and Economics, 43, 2007, pp. 185–196.

[5] R. Storn and K. Price, "Minimizing the real functions of the ICEC96 contest by differential evolution", Proc. of IEEE International Conference on Evolutionary Computation, Nagoya, Japan, 1996, pp. 842–844.

[6] M. J. Reddy and D. N. Kumar, "Multiobjective Differential Evolution with Application to Reservoir System Optimization", Journal of Computing on Civil Engineering, no. 21, 2007, pp. 136–146.

[7] S. P. Sotiroudis, S. K. Goudos, K. A. Gotsis, K. Siakavara, and J. N. Sahalos, "Application of a Composite Differential Evolution Algorithm in Optimal Neural Network Design for Propagation Path-Loss Prediction in Mobile Communication Systems", Antennas and Wireless Propagation Letters, IEEE, vol. 12, 2013, pp. 364–367.

[8] H. Zhou and J. Zhang, "Application of Differential Evolution Optimization Based Gaussian Mixture Models to Speaker Recognition", The 26th Chinese Control and Decision Conference (2014 CCDC), 2014, pp. 4297–4302.

[9] A. F. Shapiro and R. P. Gorman, "Implementing adaptive nonlinear models", Insurance: Mathematics and Economics, vol. 26, Issues 23, 2000, pp. 289–307.

[10] P. Posík, W. Huyer, and A. Pál, "A comparison of global search algorithms for continuous black box optimization", Evolutionary Computation, vol. 20, 2012, pp. 509–541.

[11] S. Salcedo-Sanz, L. C. Calvo, M. M. C. Bielsa, A. Castañer, and M. Marmol, "An Analysis of Black-Box Optimization Problems in Reinsurance: Evolutionary-Based Approache". Available at SSRN: http://ssrn.com/abstract=2260320 or http://dx.doi.org/10.2139/ssrn.2260320, 2013, [Retrieved: May-2015]

[12] O. A. C. Cortes, P. F. do Prado, and A. Rau-Chaplin, "On VEPSO and VEDE for Solving a Treaty Optimization Problem", Data Analytics, IEEE International Conference on System, Man, and Cybernetics, Sandiego-USA, 2014, pp. 2427–2432.

[13] T. Robič and B. Filipič,"DEMO: Differential Evolution for Multiobjective Optimization", 3rd International Conference on Evolutionary MultiCriterion Optimization, LNCS, 2005, pp. 520-533.

[14] M. Ali, P. Siarry, and M. Pant, "An efficient Differential Evolution based algorithm for solving multi-objective optimization problems", European Journal of Operational Research, Volume 217, Issue 2, 2012, pp. 404–416.

[15] Deb. K., "Multi-objective Optimization using Evolutionary Algorithms", John Wiley and Sons LTDA, 2001.

[16] A. K. Qin, V. L. Huang, and P. N. Suganthan, "Differential Evolution Algorithm With Strategy Adaptation for Global Numerical Optimization", IEEE Transactions on Evolutionary Computation, vol. 13, no.2, 2009, pp. 398–417.

[17] Q. Zhang and H. Li, "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition", IEEE Transactions on Evolutionary Computation, vol. 11, no. 6, 2007, pp.712–731.

[18] E. Alba, "Parallel Evolutionary Algorithms Can Achieve Super-Linear Performance", Information Processing Letters, vol. 82, 2002, pp. 7-13.

 ISBN: 978-1-61208-423-7

[19] Snow Package, http://cran.r-project.org/web/packages/snow/index.html, [Retrieved: May-2015]

[20] H. Wang, O. A. C. Cortes, A. Rau-Chaplin, “Dynamic Optimization of Multi-layered Reinsurance Treaties”, Symposium on Applied Computing, ACM, Salamanca-Spain, 2015, pp. 125–132.

 ISBN: 978-1-61208-423-7

League Adjusted Salary Model using Local Polynomial Regression

Shinwoo Kang
Development
Seattle Humane Society
Bellevue, USA
e-mail: chrisk@seattlehumane.org

***Abstract* - Since the 2012 National Hockey League (NHL) Lockout, there have been many economic trends in the league that one might argue inconsistent. While many players' salaries were significantly altered as results of buy-outs or extravagant contract signings, the salary cap has fluctuated dramatically in the following years due to these chaotic activities. To understand the seemingly contradicting NHL economic trends, in this paper, we discuss League Adjusted Salary Model (LASM) applying Local Polynomial Regression Modeling to properly gauge a player's monetary vs. production feasibility value. The League Adjusted Salary Model is a approach that is dependent on a player's League-Relative Salary Percentages and his Individual Production. The League Relativity is emphasized to account for the different payrolls of all 30 NHL teams and to understand the year-by-year economic trend. The Individual Production is a user flexible element of the individual level model that can be improved with utilizations of "Enhanced Statistics" such as Unblocked Shot Attempt Relative Percentage values. Combining these two data sets, we apply the Local Polynomial Regression Modeling to compute the feasibility of cost and production.**

Keywords-hockey; Local Polynomial Regression; Economics; Salary Cap.

I. INTRODUCTION

After the new Collective Bargaining Agreement (CBA) in 2012, National Hockey League (NHL) teams were granted opportunities to buy-out players under contract. A record number of 26 players were bought out since June 23, 2013. Of the 26 players, only 16 remain in NHL at reduced salary (with a notable exception of Christian Ehrhoff). Unfortunately, the rate of reduction in salary is seemingly random. The sudden decrease in salaries for these players impacts the overall economy of the game. The new cap space acquired by the decrease in salaries allows (1) teams to sign more players, or (2) teams to re-sign players with a bump in salary. These two scenarios present difficulties in projecting salaries of other players based on performance.

Once a player's decrease or increase in salary can be put into the context of whole league, then we may establish a regression model that projects a player's upcoming salary, which we'll call "League Adjusted Salary Model." League Adjusted Salary Model employs Local Polynomial Regression Modeling to account for random noises and possibly misunderstood NHL contracts. League Adjusted Salary Model is an improvement from the simple linear regression on salary vs. performance, which is the traditional school thought in hockey analytics community [1].

The rest of the paper is organized as follows: Section 2 explains the methodology behind League-Relative Salary Percentage, League-Relative Cap Percentage, and League-Adjusted Salary Model. Section 3 describes the application of the model on training set data from the 2010~2011 NHL season to 2013~2014 NHL season. Section 4 concludes the paper with final remarks on the potential of the proposed model and possible improvements to it.

II. THE LEAGUE ADJUSTED SALARY MODEL

The League Adjusted Salary Model is a two-part process, where the League-Relative Percentages (Salary and Cap) must be computed first. Then, a comparison of Linear Regression and Local Polynomial Regression Modeling is performed to provide a method that better fits the Cap and Salary. With the League Adjusted Salary Model, one may apply it for various purposes such as for determining the Expected Future Salary or possible statistical areas of improvement to maximize the salary potential. For this research, only the data for forwards and defensemen were considered, as goalies have independent valuation processes.

A. League-Relative Percentages

In order to compensate for the uncertainty level of buyouts, we introduce "League-Relative Salary Percentage" and "League-Relative Cap Percentage." The League-Relative Percentages ignore the unpredictability of contract buy-outs and re-signing, as one player moves from one team to another, the relative worth changes in respect to the particular team. The League-Relative Salary Percentage allows for low-market teams that are bounded by internal payroll amount**.** The League-Relative Salary Percentage is essentially a proportion of a player's True Salary/Cap to a sum of all NHL team's payrolls. We make a note that True Salary and Cap will be treated separately as explained further later.

If a player was bought-out, we create a rule to apply weighted average of salary/cap as the adjusted predictor in

 ISBN: 978-1-61208-423-7

respect to performance before and after the buy-out. Since it is difficult to measure if a player was initially overpaid and/or still overpaid after the buy-out. A striking example is of Scott Gomez who received the cap and salary of $7,357,143 and $7,500,00 in the 2011-2012 season, while receiving $700,00 for cap and salary, after the buy-out. He had .289 Points per Game (PPG) in 2011~2012 and .385 PPG in 2012~2013 season. By having the weighted average on production for the years a particular player was bough-out, it relaxes the noise it would be created in the ratio of "bought-out" cap/salary vs. production. The formulas for League Relative Cap and Salary (shortened for Sal) are,

$$LeagueRelCap\%_i = \frac{Cap_i}{\sum_{Team \in NHL} \sum_{player \in Team} Cap_{player}} \quad (1)$$

$$LeagueRelSal\%_i = \frac{Sal_i}{\sum_{Team \in NHL} \sum_{player \in Team} Sal_{player}} \quad (2)$$

where i indicates a particular player on a team.

The advantage of League-Relative Salary Percentage is that each NHL season is treated as an independent economy as a whole.

B. League Adjusted Salary Model

With League-Relative Percentages, we compare two methods: Linear Regression and Local Polynomial Regression Modeling on Production vs. League-Relative Percentages. The results of the comparison in Section III will show why linear regression is insufficient for modeling Salaries and Cap, and need a more flexible methods that is capable of modeling general nonlinear relationship [2].

For the predictors, we utilize Points Per 60 Minutes (P60), Offensive Zone Start Relative % (OZS%), Unblocked Shot Attempt Relative % (USAT Rel%), and Time on Ice (TOI), as they are the modern day go-to-metrics for evaluating a player's game, in addition to the two traditional statistics, Goals and Assists. The number of different metrics we compare may not be limited to these six. The general model for the linear regression may be represented as follows:

$$y = \alpha + \beta_i x_i + \varepsilon \quad (3)$$

where y is desired expected League-Relative Percentages. x_i is the training set of above predictors. For the Local Polynomial Regression, we use the traditional tri-cube kernel weights [3]:

$$\omega(x) = \begin{cases} (1 - |x^3|)^3 & for\ |x| < 1 \\ 0 & for\ |x| \geq 1 \end{cases} \quad (4)$$

III. CASE STUDY

In this section, we discuss the procedure of obtaining the proper NHL data, and correctly modeling it, by separating fixed and random effects.

A. Data Sources

The data sources for the two components of the proposed model are [4]–[9]. During the research, many data sources had to be aggregated and cross validated into a single database, since the industry leading [9] ceased its operation in 2014. For the League Relative Salary Percentage, we utilize statistics beginning with 2010~2011 season.

It must be noted that for the purpose of the analysis, we make a distinction between Cap and Salary, as they are indicators of their monetary compensation, but hold different meanings. These two numbers will be treated differently, as Cap Space, due to its nature, is uniform through the duration of the contract, while the true Salary usually changes from year to year and it may trend upwards or downwards, depending on age, and whether a player is entering his prime or not.

For production, we gather data exclusively from [4] and [5]. In addition to conventional statistics, such as Goals/60 and Assists/60, we utilize advanced shot metrics to compare across different linear regressions and Local Polynomial Regression. As previously stated, we utilize Points Per 60 Minutes (P60), Offensive Zone Start Relative % (OZS%), Shot Attempt Relative % (SAT%), and Time on Ice (TOI), as initial predictors because they give contextual clues to a player's game.

Combining the six data sources, we create one data frame to help compute League Adjusted Salary Model. The final data frame will include two extra columns of predictors in League Adjusted Cap Percentage and League Adjusted Salary Percentage. The initial plot of the two League Adjusted Percentages against USAT Rel% (Figures 1 and 2) shows that Salary and Cap have different spreads.

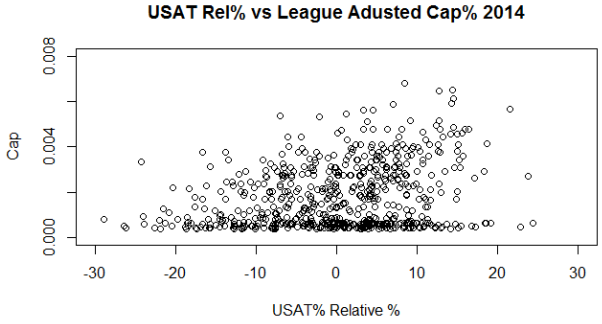


Figure 1. Disbrituion of League Adjusted Cap Percentage over USAT%

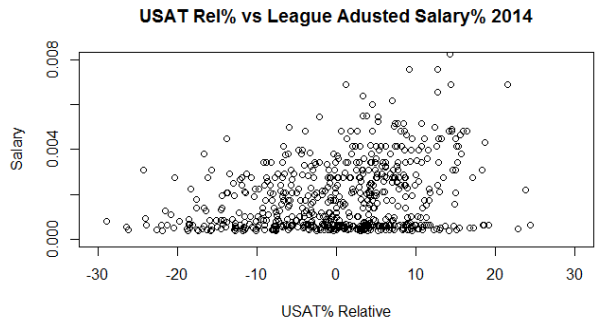


Figure 2. Disbrituion of League Adjusted Salary Percentage over USAT%

For this particular example in showing the difference in spreads, we use the 2014~2015 season data due to its availability in salary/cap information, but incompleteness in games played. The rest of the paper utilizes only 2010~2014 season data as training set for the model.

B. Applications of the Model and its Results

With the new data frame including League Adjusted Salary Percentage, we proceed with Linear Regression and Local Polynomial Regression on the proposed Enhanced Statistics. Utilizing R package, 'loess,' we compute the following results.

TABLE I. SCALED LEAGUE ADJUSTED CAP DISTRIBUTION

Min	*1Q*	*Median*	*3Q*	*Max*
.03752	.06099	.1365	.2739	.6823

TABLE II. SCALED LEAGUED ADJUSTED SALARY DISTRIBUTION

Min	*1Q*	*Median*	*3Q*	*Max*
.03785	.06194	.13760	.27530	.96350

TABLE III. LINEAR REGRESSION VS LOESS CAP

	LR Coeff Estimate	*LR Std. Error.*	LRR^2	*Loess Std Error*
G60	4.297e-04	8.680e-05	.03625	1.344e-04
A60	3.973e-04	6.033e-05	.06498	1.3e-04
P60	3.457e-04	4.486e-05	.08693	1.312e-04
OZS%	1.854e-05	3.600e-05	.04076	1.35e-04
USAT Rel%	4.232e-05	5.862e-06	.07707	1.335e-04
TOI	1.836e-04	1.029e-05	.3379	1.09e-04

TABLE IV. LINEAR REGRESSION VS LOESS SALARY

	LR Coeff Estimate	*LR Std. Error.*	LRR^2	*Loess Std Error*
G60	4.529e-04	9.450e-05	.03351	1.468e-04
A60	4.278e-04	6.565e-05	.06371	1.42e-04
P60	3.695e-04	4.886e-05	.08397	1.432e-04
OZS%	1.967e-05	3.918e-05	.03883	1.471e-04
USAT Rel%	4.477e-05	6.389e-06	.07294	1.454e-04
TOI	1.959e-04	1.129e-05	.3255	1.213e-04

Tables 1 and 2 display the feature scaled distribution of the League Adjusted Cap and Salary Models, respectively. Numbers suggest that the Salary Model has wider ranges of residuals than the Cap Model. This can be attributed to the fact that the cap numbers of a contract are uniform through out the duration of the contract, and salaries are often front or back-loaded by age, resulting in little changes despite a possible improvement or a decline in a player's performance. In accordance to the Residuals and Variances in the tables, the plots of the League Adjusted Cap Model (Figure 3) and League Adjusted Cap Model (Figure 4) display smooth lines with a concave dip in the center. The concavity of the plot is the result of players who possess large contracts with high variability in statistics across G60, A60, P60, OZS%, SAT, and TOI.

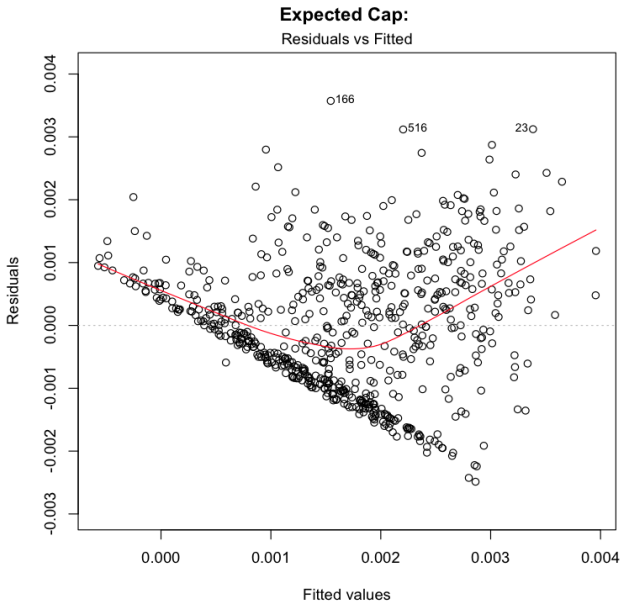


Figure 3. Plot of League Adjusted Cap Model

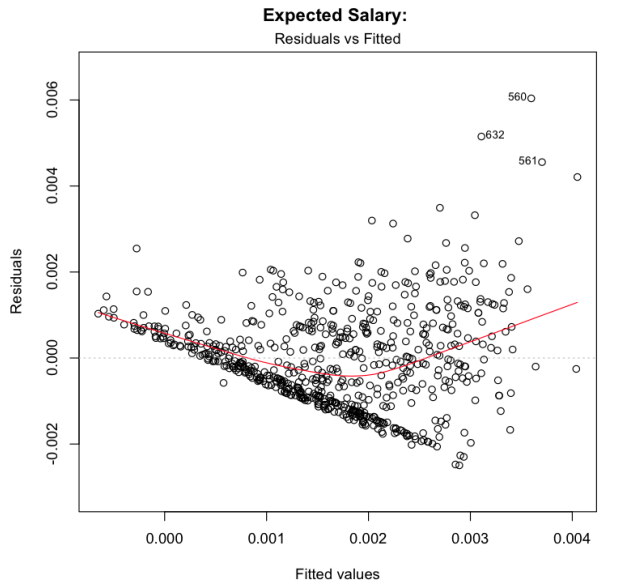


Figure 4. Plot of League Adjusted Salary Model

 ISBN: 978-1-61208-423-7

Tables 3 and 4 are the direct results (coefficient estimates, standard errors, R-squared) of the Linear Regression and Local Polynomial Regression on the Enhanced statistics vs. League Adjusted Cap and Salaries. Standard Errors of all the estimates are negligibly small. There exist many counterintuitive results from the League Adjusted Salary Model. As can be seen in Tables 5 and 6, Time on Ice has the strongest R-squared value at .3379 and .3325 for both Salary and Cap. This may suggest that despite any type of production, Time on Ice is the most likely determining factor contract signings. What may be surprising is that the next highest determining factor of salary is the P60. In modern Enhanced Statistics, USAT Rel % is generally accepted as better indication of a player's ability than P60. However, this result may shows that perhaps obvious numbers in production are more valued in contract signing than, possession numbers (USAT Rel %).

In addition, as evident by near-0 Residuals for the random effect (Teams), the team association has zero impact on the salary itself. In other words, given a set of on-ice production, you will not be paid higher or lower by playing on a certain team.

IV. CONCLUSION

The League Adjusted Salary Model proposed in this paper is not just a predictive model to gauge a player's potential salary. As discovered through this analysis, with the weighting the bought-out players, and by deriving the League-Relative Salary Percentage, we can create a meaningful training set for which a plethora of statistical models, not limited to Local Polynomial Regression Modeling, may be applied. While this model is at an early stage with comparisons of only six advanced statistics as dependent variables, with expanded parameters and caution, League Adjusted Salary Model has the potential to become a powerful tool in analyzing sports economics.

There are many possible areas of improvements to League Adjusted Salary Model. As is the case in most statistical analyses, it is possible to improve the underlying statistical model. While we incorporated Local Polynomial Regression Modeling to account for standard error in Linear Regression, a more advanced modeling technique could be applied to better fit the data and reduce errors. Another area of improvement could be within the data itself. There were assumptions made in the data and methodology that may be deemed unnecessary in the hockey analytics community. Incorporating more independent variables, such as age and nationality may result in a better training set for the League Adjusted Salary Model. Inclusion of goalies in a much more complicated model is due next. An examination of previous lockout years such as the 2004 NHL Lockout may be another relevant area of research. Finally, valuation of contract clauses, such as No Trade Clause (NTC) was ignored for this paper. The author believes that these issues could have a significant impact in salary models to come.

REFERENCES

[1] R. Vollman, I. Fyffe, and T. Awad, "Rob Vollman's Hockey Abstract," N.p.: CreateSpace Independent Platform, 2014.

[2] M. Kuhn, and K. Johnson. "Applied Predictive Modeling," Springer, 2013, pp. 464-465.

[3] T. Hastie, R. Tibshriani, and J. Friedman. "The Elements of Statistical Learning," Springer, 2013, pp. 197-198.

[4] NHL.com

[5] War-on-Ice.com

[6] HcokeyBuzz.com

[7] Spotrac.com

[8] NHLPA.com

[9] CapGeek.com

 ISBN: 978-1-61208-423-7

Context-Aware Data Analytics for Activity Recognition

Mohammad Pourhomayoun
Ebrahim Nemati, Majid Sarrafzadeh
Computer Science Department
University of California, Los Angeles (UCLA)
Los Angeles, USA
e-mails:{mpourhoma, ebrahim, majid}@cs.ucla.edu

Bobak Mortazavi
School of Medicine
Yale University
Connecticut, USA
e-mail: bobak.mortazavi@yale.edu

***Abstract*— Remote Health Monitoring Systems are gaining an important role in healthcare by collecting and transmitting patient information and providing data analytics techniques to analyze the collected data and extract knowledge. Physical activity recognition and indoor localization are two of the most important concepts in assistive healthcare, where tracking the positions, motions and reactions of a patient or elderly is required for medical observation or accident prevention. In this paper, we propose a novel context-aware data analytics framework to classify and recognize the physical activity based on the signals received from a worn SmartWatch, the location information of the human subject, and advanced machine learning algorithms. In this approach, we take into account the physical location of the human subject as contextual information to improve the accuracy of the activity classification. The hypothesis is that the location information can get involved in classifier decision making as a prior probability distribution to help improve the accuracy of activity recognition. The results demonstrate improvements in accuracy and performance of the activity classification when applying the proposed method compared to conventional classifications.**

Keywords-Activity Recognition; Indoor Localization.

I. INTRODUCTION AND BACKGROUND

As the number of elderly people grows rather quickly over the past few decades and continues to do so [1], it is essential to seek alternative and innovative ways to provide affordable healthcare to the aging population [2]. A compelling solution is to enable pervasive healthcare for the elderly or patients with chronic disease at their own homes, while reducing the use and dependency of healthcare facilities. New technologies, such as Body Sensor Networks (BSN) and Remote Health Monitoring Systems (RHMS) allow for collecting continuous data and monitoring the patients in their home environment. There have been a number of studies on end-to-end remote health monitoring and medical data analytics using wearable or environmental sensors known as Smart Environment or Smart Home [3]-[7]. RHMS has shown substantial potential in reducing healthcare costs and improving quality of care [3]-[10]. Rapid advances in many technological domains including electronics, wireless communications, Internet, and sensor design has led to the development of effective RHMS that can collect varying physiological information, vital signs, and physical activity from patients [3]-[7].

Although RHMS have shown promise in reducing healthcare costs and improving quality of care, effective analysis of the data collected by these systems and the potential benefits of utilizing such analysis is by large an open problem. One of the key demands in such an assistive environment is to promptly and accurately determine the state and activities of an inhabitant subject. The physical activity recognition and indoor localization provide effective means in tracking the positions, motions, and reactions of a patient, the elderly or any person with special needs for medical observation or accident prevention [11][12].

Physical activity recognition using wearable sensors or smartphones has been a long-standing problem. There have been a number of studies on utilizing machine learning algorithms to monitor the activities of daily living [24][25]. However, in this paper, we propose a novel context-aware data analytics framework to classify the physical activity based on the signals received from a wearable sensor (e.g., SmartWatch [28]), the position information of the human subject, and advanced machine learning algorithms. The location of a patient can provide important prior information that can be used to better classify the physical activity. We hypothesize that the location information of the human subject can get involved in classifier decision making as a prior probability distribution to improve the accuracy of activity recognition. In other word, we take into account the location of the subject as contextual information to improve the accuracy of the activity classification. The results demonstrate improvements in accuracy and performance of the classifier when applying the proposed method compared to typical classifications.

The rest of the paper is organized as follows: Section II describes the systems architecture and main modules for the proposed context-aware data analytics framework, Section III provides a brief overview of the indoor localization technique that we use to come up with the contextual information. This localization technique is a novel approach developed by the authors. However, since the focus of this paper is on data analytics, we just briefly review this technique, and use the results as contextual information in

 ISBN: 978-1-61208-423-7

our analytics framework. Section IV describes the details of the proposed context-aware analytics framework for activity recognition, including feature extraction, feature selection algorithms, classification, training/testing stages, and the context-awareness characteristics of the system. Finally, Section V describes the results and conclusion.

II. RELATED WORK

Physical activity monitoring and indoor localization are important problems in the areas of wireless health and assistive healthcare that have raised increasing attention recently [12]-[37][28][36]. Monitoring the activities of daily living with smartphones and devices with these phones have been well-studied [4][24]-[37]. In particular, Alshurafa, et al. [4] presents a comprehensive activity recognition process and particularly, looks at activity tracking for a clinical environment, and how to guarantee that patients are performing the desired activity. Gupta, et al. [37] presents an activity recognition system using a single waist-mounted accelerometer to classify gait events into six daily living activities. SmartWatches have also been used to provide activity tracking applications to date [28][35]. Mortazavi, et al. [28] provides visual feedback and interface for activity repetition counting using SmartWatch. Park, et al. [35] develops a watch sensor to track fall, walking, handrelated shocks, and general activity. Using a feature extraction and selection technique, results are presented in a 10-fold cross validation to determine the ability to track elderly patients. Park, et al. [35] uses a forward selection technique for feature selection and a support vector machine, to obtain accuracy results and recall results. In this study, we propose a new context-aware activity recognition system that utilizes the SmartWatch accelerometer and gyroscope signals, and takes into account the location of the subject as contextual information to improve the accuracy of the activity classification. The results demonstrate improvements in accuracy and performance of the classifier when applying the proposed method compared to typical classifications.

III. SYSTEM ARCHITECTURE

The proposed framework includes two main modules: a) Indoor Localization/Tracking Module and b) Context-Aware Activity Recognition Module. Indoor Localization and Tracking Module is responsible for estimating and tracking the position of a patient. We use a novel approach for localization based on spatial sparsity of target in x-y-z space and the Received-Signal-Strength (RSS) between a SmartWatch and RF beacons mounted in the building.

Context-Aware Activity Recognition Module is responsible for classifying and recognizing patients' physical activities using data analytics techniques based on the wearable embedded accelerometer and gyroscope signals. This module includes feature extraction, feature selection and dimensionality reduction, and context-aware classification submodules. In the proposed approach, we exploit the location information of the subject (received form patient tracking module) to achieve more accurate results for activity recognition. Details of these modules are described in next sections.

IV. INDOOR LOCALIZATION AND TRACKING

As mentioned before, the main focus of this paper is not on indoor localization; instead it is on context-aware data analytics for activity recognition knowing the indoor location of the individual. In other word, we take into account the position of the human subject as contextual information to improve the accuracy of the analytics engine for activity recognition. Thus, in this paper, we only provide a brief overview of the novel indoor localization techniques that we have developed in our other studies, and then apply these techniques to estimate individual's location that will be later used in our analytics framework. For more details about our developed localization techniques please refer to [11]-[17].

Indoor localization has been a long-standing and important problem in the areas of signal processing and sensor networks that has raised increasing attention recently [11]-[23]. One of the key demands in assistive environment is to promptly and accurately determine the state and activities of an inhabitant subject. Indoor localization provides an effective means in tracking the positions, motions, and reactions of a patient, the elderly or any person with special needs for medical observation or accident prevention.

The classic approach for localization is to first estimate one or more location-dependent signal parameters, such as Time-Of-Arrival (TOA), Angle-Of-Arrival (AOA) or RSS. Then in a second step, the collection of estimated parameters is used to determine an estimate of the subject's location. The TOA-based methods are usually more accurate than RSS or AOA techniques. However, the accuracy of the classic TOA based methods often suffer from massive multipath conditions for indoor localization, which is caused by the reflection and diffraction of the RF signals from objects (e.g., interior walls, doors or furniture) in the environment [23]. Moreover, it usually necessitates using synchronized emitters/sensors to be able to estimate accurate time-of arrival or time-difference-of-arrival.

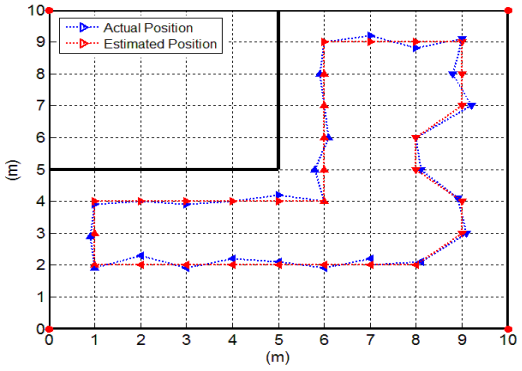
In [11]-[15], we introduced a novel accurate localization method based on the spatial sparsity in the x-y-z space. In this approach, we directly estimate the location of the emitter without going through the intermediate stage of TOA or RSS estimation. To this end, we utilize the spatial sparsity of the target (SmartWatch worn by a human subject) in the X-Y-Z space, and use the convex optimization theory to estimate the location of the subject. Assume that we divide the X-Y-Z space into fine enough grids. By assigning a positive number to each grid that contains the target and zeros to all the rest of grid cells, we will have a very sparse 3-dimensional grid matrix that can be reformed as a *sparse vector*. Since each element of this

 ISBN: 978-1-61208-423-7

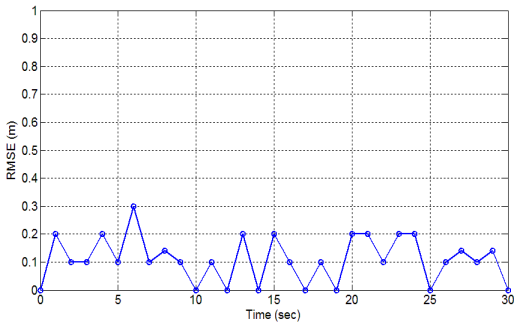
grid vector corresponds to one grid point in the X-Y-Z space, we can estimate the location of emitters by extracting the position of non-zero element (or non-zero elements when we have more than one subject to be determined) in the sparse vector. To this end, we have to estimate the sparsest vector that minimizes the cost between the predicted received signal and the actual observed signal with respect to the signal model and distance between the transmitted signals and the received signals (for details and problem formulation please refer to [11]-[15]).

The results demonstrate that the proposed method has very good performance even with small number of sensors. The results also indicate that, in contrary to the classic methods, the proposed approach is a very effective and robust tool to overcome multipath issues, which is a very serious problem in indoor localization. Furthermore, the system works well in noisy environments with low SNRs. It implies that, even with low transmitted power (to keep the devices small with long battery life), we can still achieve a high localization accuracy.

Figure 1 shows some of the results for patient localization and tracking in a sample building using only 4 RF sensors mounted at the corners of a building. Figure 1-(a) shows the actual trajectory (blue line) of the patient walking around in the room, and the estimated path (red line) by the proposed system. Figure 1-(b) shows the error defined as the root-mean-square (RMS) errors for positioning in the X, Y and Z dimensions.



(a)



(b)

Figure 1. (a) True position of the patient (in blue) and the estimated position (in red), (b) Error in positioning for each location in part (a).

V. CONTEXT-AWARE ANALYTICS FRAMEWORK FOR ACTIVITY RECOGNITION

Context-Aware Activity Recognition Module is responsible for recognizing the physical activities based on the accelerometer and gyroscope signals. This work will investigate the ability of the SmartWatch to recognize and track the necessary activities of human subjects in order to better assess their health status. In particular, by identifying the transitions between sitting, standing, and lying, this work approaches the classification of patient status. Monitoring the Activities of Daily Living (ADL) through wearable body sensors has attracted extensive attention recently [24]-[28]. In this study, we propose a context-aware activity recognition system based on the signals received from embedded accelerometer and gyroscope of a SmartWatch, a real-time machine learning based analytics engine, and the position information received from the indoor localization module.

Our preliminary results [28] show that the watch can provide accurate activity tracking results similar to custom sensing environment. However, in this work, we propose a context-aware technique by taking into account the indoor position of the individual as prior contextual information that can modify the classifier model, and consequently provide more accurate results for activity recognition. The activity recognition module includes feature extraction, feature selection, and context-aware classification submodules as described in the following.

A. *Feature Extraction and Feature Selection*

The first step is to gathering the patient's activity signals from the SmartWatch embedded accelerometer and gyroscope. After receiving the signals, the next step is to data preprocessing and feature extraction. We use a moving average window as a low-complexity low-pass filter for the purpose of denoising. Then, a total number of 150 features are extracted from accelerometer and gyroscope signals. Statistical and morphological features are the most common features used for data analytics. Theses feature are extracted for each one of the three axes of the accelerometer and gyroscope. Some of the extracted features include Mean, Standard Deviation, Kurtosis, Skewness, Energy, Variance, Median, RMS, Minimum, Maximum, Sum, Average Difference, Eigenvalues of Dominant Directions, CAGH, Average Mean Intensity, Dominant Freq., Peak Diff., Peak RMS, Root Sum of Squares, First Peak, Second Peak. In this study, the Samsung Galaxy Gear SmartWatch is used for experimentation. It employs a $\pm 2g$ triaxial accelerometer and ±300 degree per second gyroscope sensors.

Once the features are extracted, a dimensionality reduction algorithm is applied to select the most prominent features and reduce the redundancy. The conventional feature selection algorithms usually focus on specific metrics to quantify the relevance and redundancy of each feature with the goal of finding the smallest subset of

 ISBN: 978-1-61208-423-7

features that provides the maximum amount of useful information for prediction. Thus, the main goal of feature selection algorithms is to eliminate redundant or irrelevant features in a given feature set. Applying an effective feature selection algorithm not only decreases the computational complexity of the system by reducing the dimensionality and eliminating the redundancy, but also increases the performance of the classifier by removing irrelevant features. In this paper, we tried both wrapper and filter methods; the two well-known feature selection categories. Wrapper methods usually utilize a classifier to evaluate feature subsets in an iterative manner according to their predictive power. A new feature subset is used to train a predictive model that will later be evaluated on a testing dataset to assess the relative usefulness of subsets of features [39]. Figure 2-(a) provides an illustration of the wrapper feature selection method.

Filter methods use a specific metric to score each individual feature (or a subset of features together). The most popular metrics used in filter methods include correlation coefficient, mutual information, Fisher score, chi-square parameters, entropy and consistency. Filter methods are very popular (especially for large datasets) since they are usually very fast and much less computationally intensive than wrapper methods. Figure 2-(b) illustrates the steps involved in the filter feature selection method.

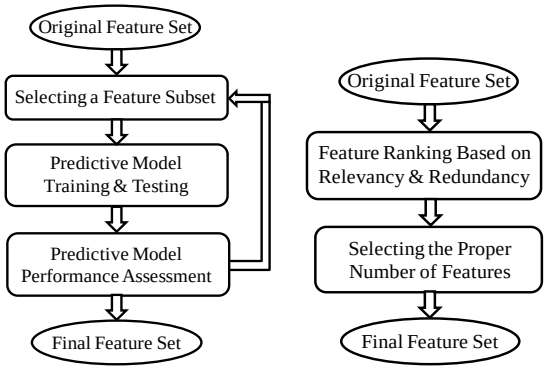


Figure 2. Feature Selection: (a) Wrapper method, (b) Filter method.

In this study, after trying several filter and wrapper methods, we finally chose only 5 features to keep the computational complexity low on the device. The selected features includes: minimum of acceleration axis x (min ax), average acceleration axis z (avg az), eigenvalue acceleration axis z (eigen az), correlation between acceleration axis x and y (cor axy), sum gyro axis z (sum gz).

B. *Classification: Training and Testing*

Once the subset of features is selected, a machine learning based classifier is applied to classify the motions. In this research, we tried various classification algorithms such as SVM, Random Forest, BayesNet, and Artificial Neural Net (ANN) as the predictor. According to our results, a Random Forest classifier with 100 trees provided fast and accurate prediction results for our dataset. Random Forest is an ensemble learning classification method comprising of a collection of decision tree predictors operating based on i.i.d random vectors. In this process, each tree casts a unit vote for the most popular class [40]. The classifier was supplied with training data labeled with 6 labels being the six transition movements (sit_to_lie, sit_to_stand, stand_to_sit, stand_to_lie, lie_to_sit, lie_to_stand). The recognition algorithm must then be validated to ensure the proper development of a system to accurately track the state of subjects. Figure 3 indicates the Training and Testing stages. The next section describes the context-awareness approach and how we take into account the location information to improve the classifier accuracy.

C. *Context Awareness*

The indoor position of a patient (received from indoor localization and tracking module) can provide significant prior information about the possible physical activity. For example, when we know that the patient is in the kitchen, the probability of standing is much higher than lying, consequently, the labels are not uniformly distributed anymore. Thus, by knowing the approximate position of the patient, we will have better understanding about the possible activities that the patient can have.

We hypothesize that the location information can get involved in classifier decision making as a prior probability distribution to help improve the accuracy of activity

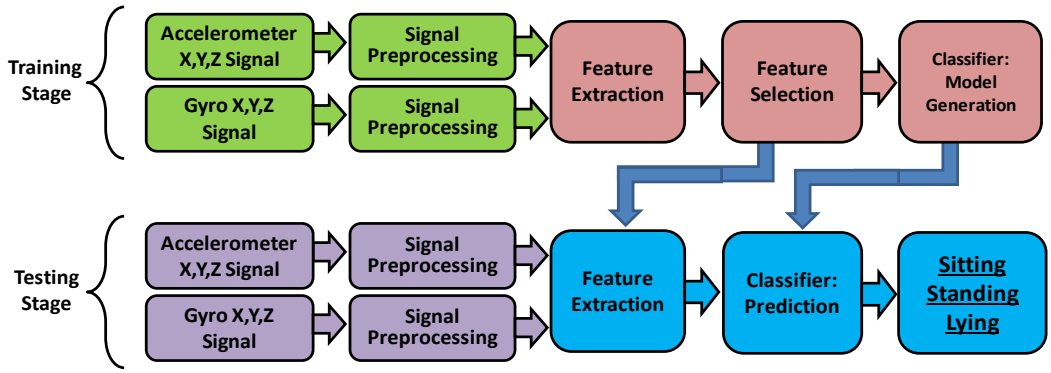


Figure 3. The regular Physical Activity Classification.

 ISBN: 978-1-61208-423-7

recognition module.

Assume that $F_1, \ldots, F_N$ are the classifier input features and C represents the classifier labels. Then, the classifier probability model can be expressed as a conditional probability $p(C \mid F_1, \ldots, F_N)$ (known as *Posterior Probability*) that can be formulated using the Bayes' Theorem as following [41]:

$$p(C \mid F_1, \ldots, F_N) = \frac{p(C, F_1, \ldots, F_N)}{p(F_1, \ldots, F_N)} \quad (1)$$

The joint probability in the numerator can be reformulated as:

$$\begin{aligned} & p(C, F_1, \ldots, F_N) = p(C)\, p(F_1, \ldots, F_N \mid C) \\ & = p(C) p(F_1 \mid C) p(F_2, \ldots, F_N \mid C, F_1) \\ & = p(C) p(F_1 \mid C)\, p(F_2 \mid C, F_1)\ \ldots\ p(F_N \mid C, F_1, \ldots, F_{N-1}) \end{aligned} \quad (2)$$

A "*Maximum A Posteriori*" (MAP) decision making rule can be applied as following to pick the most probable class label:

$$\begin{aligned} & calssify(f_1, \ldots, f_N) \\ & = \arg\max_c p(C = c)\ p(f_1, \ldots, f_N \mid C = c) \end{aligned} \quad (3)$$

The term $p(F_1, \ldots, F_N \mid C)$ (called *likelihood*) is usually determined in the training stage. For the case of simplicity (e.g., in Naive Bayes classifier [41]), the features can be assumed to be conditionally independent. In this case, the equation (3) can be simplified to:

$$\begin{aligned} & calssify(f_1, \ldots, f_N) \\ & = \arg\max_c p(C = c) \prod_{i=1}^{N} p(F_i = f_i \mid C = c) \end{aligned} \quad (4)$$

In traditional classification, a uniform distribution is used for *Prior Probability* $p(C)$. However, in our approach, we hypothesize that the patient's position can provide some information about the distribution of the *prior probability* $p(C)$. Thus, we can write $p(C)$ as:

$$\begin{aligned} p(C = c) &= \sum_i p(C = c, L = l_i) \\ &= \sum_i p(L = l_i)\, p(C = c \mid L = l_i) \end{aligned} \quad (5)$$

where $p(C, L)$ is the joint probability distribution of location and activity label. Thus, when the location is known, the uniformly distributed *Prior Probability* $p(C)$ will be replaced by the conditional probability $p(C \mid L = l_i)$ and consequently, the equation (4) provides more accurate model for activity recognition.

VI. RESULTS AND CONCLUSION

A pilot trial has been conducted to collect the data. The dataset contains 1200 data samples collected from 20 subjects. Table I shows the F-Score results for the activity recognition using only 5 features in two different cases: a) Using conventional classification without considering the location information, b) Context-aware activity recognition knowing and taking into account the location information. As we see, for example in the kitchen, we achieve 7% improvement (using 5 features) since knowing the location of the subject provides significant information about the activity. However, in the living room, we achieve 3% improvement, and it totally makes sense, because the likelihoods of sitting, lying, and standing in the living room are almost similar, and consequently the prior probability distribution is closer to the uniform distribution which is the pre-assumption for conventional activity recognition too.

TABLE I. F-SCORE FOR REGULAR AND CONTEXT-AWARE ANALYTICS USING ONLY 5 FEATURES

Location	F-Score for conventional classification	F-Score for context-aware classification
Kitchen	0.81	0.88
Living room	0.82	0.85
Bedroom	0.80	0.84

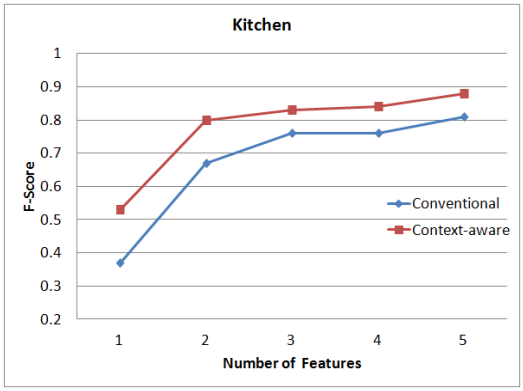


Figure 4. F-Score versus the number of selected features for conventional and context-aware activity recognition in kitchen.

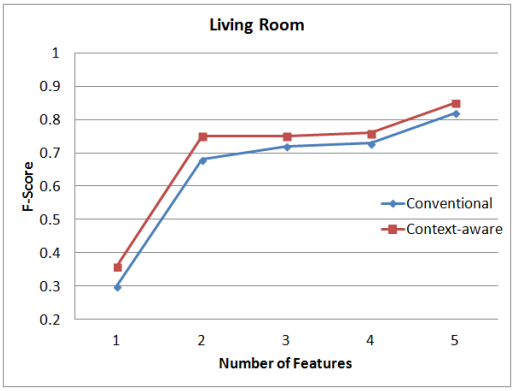


Figure 5. F-Score versus the number of selected features for conventional and context-aware activity recognition in the living room.

Figures 4 and 5 show the F-Score [41] versus the number of selected features for conventional and context-aware analytics in the kitchen and living room. F-Score is a well-known measure for classification accuracy, and it can be interpreted as the harmonic mean of precision (the fraction of retrieved instances that are relevant) and recall (the fraction of relevant instances that are retrieved). Thus, F-

 ISBN: 978-1-61208-423-7

score is an indication of how well the system can identify the activity and how strong it is at not mis-predicting.

For example, for kitchen, we achieved 43% improvement using 1 feature and 9% improvement using 5 features in activity recognition accuracy, which is a significant improvement. Our work in [42] investigates the impact of improvement in classification accuracy on cost.

Again, as we expected, the improvement by using context-aware approach is higher in the kitchen compared to living room because the probability distribution of various activities in the living room is closer to uniform distribution.

REFERENCES

[1] W. He, M. Sengupta, V. Velkoff, and K. DeBarros, 65+ in the United States, Current Population Reports, U.S. Census Bureau, 2005.

[2] United Nations, "World Population Prospects: The 2008 Revision, Highlights," Department of Economic and Social Affairs, Population Division, Working Paper No. ESA/P/WP.210, 2009.

[3] M. Lan, et al., "WANDA: An End-to-End Remote Health Monitoring and Analytics System for Heart Failure Patients," Wireless Health Conf., 2012, pp. 68–74.

[4] N. Alshurafa, et al. "Anti-Cheating: Detecting Self-Inflicted and Impersonator Cheaters for Remote Health Monitoring Systems with Wearable Sensors," BSN 2014, 2014, pp. 92-97.

[5] N. Alshurafa, et al. "Battery optimization in smartphones for remote health monitoring systems to enhance user adherence," Int. Conf. on PErvasive Technologies Related to Assistive Environments, 2014.

[6] N. Alshurafa, et al. "Improving Compliance in a Remote Health Monitoring System through Smartphone Battery Optimization," IEEE J Biomed Health Inform, 2014, pp. 57-63.

[7] M. Pourhomayoun, et. al., "Multiple Model Analytics for Adverse Event Prediction in Remote Health Monitoring Systems," IEEE EMBS Conference on Healthcare Innovation & Point-of-Care Technologies, 2014, pp. 106-110.

[8] R. Cebul, T. Love, A. Jain, and C. Herbert, "Electronic Health Records and Quality of Diabetes Care," J. Med., 2011, pp. 825-833.

[9] R. Hillestad, et al., "Can Electronic Medical Record Systems Transform Health Care? Potential Health Benefits, Savings, and Costs," Health Affairs, vol. 24, no. 5, 2005, pp. 1103-1117.

[10] R. W. Jang, et al., "Simple prognostic model for patients with advanced cancer based on performance status," JOP, 2014,pp. 10-15.

[11] M. Pourhomayoun, Z. Jin, and M.L. Fowler, "Indoor Localization, Tracking and Fall Detection for Assistive Healthcare Based on Spatial Sparsity and Wireless Sensor Network," Journal of Monitoring and Surveillance Tech. Research, 2013, pp. 72-83.

[12] M. Pourhomayoun, Z. Jin and M.L. Fowler, "Spatial Sparsity Based Indoor Localization in Wireless Sensor Network for Assistive Healthcare Systems," 34th IEEE Int. Conference of the Engineering in Medicine and Biology (EMBC2012), 2012, pp. 3696-3699.

[13] M. Pourhomayoun and M. L. Fowler, "Spatial Sparsity Based Emitter Localization," Conf. on Information Sciences and Sys., 2012, pp. 1-4.

[14] M. Pourhomayoun, Z. Jin and M.L. Fowler, "Accurate Localization of In-Body Medical Implants Based on Spatial Sparsity," IEEE Transactions on Biomedical Engineering, 2013, pp. 590-597.

[15] M. Pourhomayoun, M. Fowler, and Z. Jin, "A Novel Method for Medical Implant In-Body Localization," Conf. of IEEE Engineering in Medicine & Biology Society (EMBC), 2012, pp. 5757-5760.

[16] M. Pourhomayoun and M. Fowler, "Sensor network distributed computation for Direct Position Determination," IEEE In Sensor Array and Multichannel Signal Processing, 2012, pp. 125-128.

[17] M. Pourhomayoun and M. Fowler, "An SVD approach for data compression in emitter location systems," IEEE Signals, Systems and Computers (ASILOMAR) Conf., 2011, pp. 257-261.

[18] K. Pahlavan, P. Krishnamurthy, and J. Beneat, "Wideband radio propagation modeling for indoor geolocation applications," IEEE Commun. Mag., vol. 36, 1998, pp. 60–65.

[19] K. Pahlavan, X. Li, J. Makela, "Indoor geolocation science and technology," IEEE Commun. Mag., vol. 40, pp. 112–118, Feb. 2002.

[20] X. Li, and K. Pahlavan, "Super-Resolution TOA Estimation With Diversity for Indoor Geolocation", IEEE Transactions on Wireless Communications, vol 3, January 2004, pp. 224-234.

[21] Y. Chen and H. Kobayashi, "Signal Strength Based Indoor Geolocation," Int. Conf. on Communications, 2002, pp. 436-439.

[22] G. Zàruba, M. Huber, F. Kamangar, and I. Chlamtac, "Indoor location tracking using RSSI readings from a single Wi-Fi access point," Wireless Networks, 2007, pp. 221-235.

[23] A. Hatami, B. Alavi, K. Pahlavan, and M. Kanaan, "A comparative performance evaluation of indoor geolocation technologies," Interdisciplinary Information Sciences, 2006, 133-146.

[24] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," AAAI, 2005, pp. 1541-1546.

[25] N. Alshurafa, et. al., "Robust human intensity-varying activity recognition using stochastic approximation in wearable sensors," BSN, 2013, pp. 1–6.

[26] Z. Yan, D. Chakraborty, A. Misra, H. Jeung, and K. Aberer, "Sammple: Detecting semantic indoor activities in practical settings using locomotive signatures," in Wearable Computers (ISWC), 16th International Symposium on. Ieee, 2012, pp. 37–40.

[27] J. Fontecha, F. Navarro, R. Herv´as, and J. Bravo, "Elderly frailty detection by using accelerometer-enabled smartphones and clinical information records," Personal and ubiquitous computing, 2013, pp. 1073-1083.

[28] B. Mortazavi, et al., "Determining the single best axis for exercise repetition recognition and counting with SmartWatches," 11th IEEE Body Sensor Networks Conference (BSN), 2014, pp. 33-38.

[29] B. Mortazavi, M. Pourhomayoun, S. Nyamathi, B. Wu, S. Lee, M. Sarrafzadeh, "Multiple Model Recognition for Near-Realistic Exergaming," IEEE International Conference on Pervasive Computing and Communications (PerCom), 2015.

[30] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in Pervasive computing. Springer, 2004, pp. 1-17.

[31] B. J. Jefferis, et al., "Trajectories of objectively measured physical activity in free-living older men." Med. & Sci. in sports, 2014.

[32] H. L. Brooke, K. Corder, A. J. Atkin, and E. M. van Sluijs, "A systematic literature review with meta-analyses of within-and betweenday differences in objectively measured physical activity in school-aged children," Sports Medicine, 2014, pp. 1–12.

[33] B. Najafi, D. G. Armstrong, and J. Mohler, "Novel wearable technology for assessing spontaneous daily physical activity and risk of falling in older adults with diabetes," Journal of diabetes science and technology, vol. 7, no. 5, 2013, pp. 1147–1160.

[34] G. Bieber, M. Haescher, and M. Vahl, "Sensor requirements for activity recognition on smart watches," in Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments. ACM, 2013, p. 67.

[35] C. Park, J. Kim, and H.-J. Choi, "A watch-type human activity detector for the aged care," in Advanced Communication Technology (ICACT), International Conference on. IEEE, 2012, pp. 648–652.

[36] M. Zhang and A. A. Sawchuk, "A feature selection-based framework for human activity recognition using wearable multimodal sensors," 6th Int. Conference on Body Area Networks, 2011, pp. 92–98.

[37] P. Gupta, P. and T. Dallas, "Feature Selection and Activity Recognition System Using a Single Triaxial Accelerometer," Biomedical Eng., IEEE Trans., vol. 61, 2014, pp. 1780 - 1786.

[38] L. Palmerini, S. Mellone, L. Rocchi, and L. Chiari, "Dimensionality reduction for the quantitative evaluation of a smartphone-based timed up and go test," EMBC 2011, 2011, pp. 7179 - 7182.

[39] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection", J. of Machine Learning Research, 2003, pp. 1157-1182.

[40] L. Breiman, "Random Forests". Machine Learning, 2001.

[41] M. N. Murty and V. Susheela Devi, "Pattern Recognition: An Algorithmic Approach," Springer Science & Business Media, 2011.

[42] S. I. Lee, et al., "Remote Patient Monitoring Systems: What Impact Can Data Analytics Have on Cost?," Wireless Health Conf., 2013.

 ISBN: 978-1-61208-423-7

Monitoring Service Adaptation and Customer Churn in the Beginning Phase of a New Service

Teemu Mutanen
VTT Technical Research Center of Finland
Espoo, Finland
Email: teemu.mutanen@vtt.fi

Ville Österlund
Interquest Oy
Helsinki, Finland
ville.osterlund@interquest.com

Risto Kinnunen
Interquest Oy
Helsinki, Finland
risto.kinnunen@interquest.com

***Abstract*—This work focuses on the analytics and metrics of a service adaptation. Adaptation is analysed based on activity and churn behaviour. In a non-binding registration service, churn may not be a permanent phenomenon, but partial churn prediction could be more useful as a metric. Based on the findings service, adaptation remains similar as the customer base broadens. The burst type usage data are challenging for churn predictions, but combinations and merged variables benefit the analysis. As in the future, the findings may be integrated as part of a retention campaign or applied in the development of the service.**

Keywords–Data mining; Service adaptation; Customer churn; Customer lifetime; Logistic regression; Cross validation

I. INTRODUCTION

This work focuses on the analytics and metrics of a service adaptation. Data about the temporal adaptation of a service will provide knowledge and decision support for many causes. These are for example customer retention and churn, customer value prediction, and service development actions. Customer churn is not a measure of success itself, but understanding causes for it and taking actions to minimize the amount of churners may produce additional value.

In this work, we have formulated few research questions, while assuming the service in question is not under development during the time period. The hypothesis are based on the visual interpretation of the Figure 1. In Figure 1, is presented customer lifetime distribution of four groups, grouped by registration date. Each customer lifetime is measured from date of registration to the date of last visit. Number of trial customers are restricted to 100 for visual clarity. It seems that the share of active customers remains same and the trial users and churners could be visible in the data already after two months (60 days). In this study, we ask:

- Does the adaptation of the service change as customers base increases to include more users than early adopters?
- What would be a reasonable time horizon to follow customers after the registration?
- How much would additional data add value to the churn prediction accuracy?

The empirical study conducted in this work is based on the usage data of a browser based service received from a Finnish publishing company. The service in question is a publishing platform with various articles, stories and other content offered via browser. The study is based solely on back-end data about article views, browser specifics and reading times. In service adaptation, a fundamental concept is customer churn. This work address the issue by implementing predictions of possible churners and evaluate them over time.

The number of page views and click-through rates are widely used measures of success of web browser based services. There exists several ready made tools, both freemium and commercial-of-the-self type solution, for the analysis of page views and usage data statistics. These are outside the scope of this paper as the focus is on the service adaptation over time and the empirical study is based on the user specific data and not aggregated values.

The customer lifetime and service usage characteristics has been studied before. For example, how innovations spread and take hold [1] and about the nature of e-services and the e-service experience [2]. For this work, it is sufficient to recognize that in any new service implementation and roll out users may behave differently in early stages than later on as the so called early adopters may place higher value on different aspects than other users.

The customer event history has been studied from time window perspective before, *c.f.* [3]. The work by Poel *et al.* [3] focus on an already existing service and the aim is to predict churners with customer lifetime spanning for years. This work focus on new service and how the service adaptation develops over time. Jahromi *et al.* have studied the churn behaviour in business-to-business (B2B) context [4] and found out that similar methodology is suitable as in business-to-consumer (B2C) context.

This paper makes the following contributions: (1) problem description on a practical research topic, (2) empirical experiments of tools and methods for the analysis of service adaptation, (3) presents key findings how analytics benefits the selected domain, and (4) discuss possible future research aspects and implementations.

The paper is organized as follows: in Section I we discuss related research. In Section II we presents tools and methods in more detail which are used in this study. The empirical study is presented in Section III. Section III presents source data, pre-processing and results. In Section IV we state our conclusions and discuss about future research topics.

 ISBN: 978-1-61208-423-7

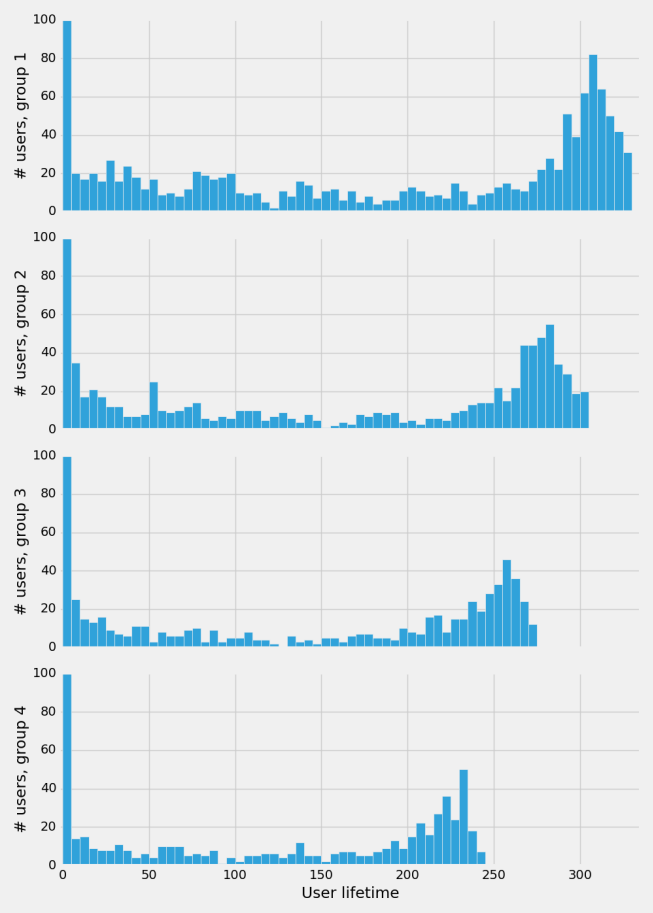


Figure 1. Customer lifetime distribution of four groups, grouped by registration date.

II. METHODOLOGY

A. Modelling customer churn

Customer churn has been studied and modelled before in various context. Understanding churners and possible causes for it is a way to analyse service adoption over time. Table I presents examples of customer churn studies. In the table, there are business sector and the amount of empirical data presented. As can be seen from the Table I, customer churn is highly interesting in sectors such as banking, insurance and telecommunication. Why interest is high in general on these sectors? Although few companies and services may have individual drawbacks, high churn rates are present in services and sectors where competing products are very similar and complement products are available, switching costs are low and customers need only one of them, for more see *c.f.* [5]. Current study focuses on churn from service adoption point of view in publishing sector.

In short, churners are predicted to increase overall satisfaction of a service. The process can be described in two phases, *c.f* [4] or [15]. In first phase the customer base is analysed and possible churners are identified. Then the churners are targeted with retention campaigns. Two widely used criteria for detection are either the most probable churners with the highest probability of defection or to weight future profit from lifetime value with the probability of defection.

In sectors such as publishing churn modelling and customer relationship management (CRM) are challenging in part because of rare event data. Rare event data may include individual events or numerous events, but which happen in bursts and may have long time periods in between. The case with rare event data has been studied by Ali *et al.* in [6], where rarity is addressed by sampling and observations from different time periods. Imbalanced data where churners are few in number may also be called as rare event data [6].

In publishing sector, the share of attention time is important concept. Electronic articles as well as all reading is in part entertainment and the amount of time customers spend in various channels is a zero-sum game. This implies that customers may be partial churners. Partial churners have been studied previously by Miguéis *et al.* [16] in grocery retail sector. Miguéis *et al.* [16] have analysed the importance of first impression as a measure of future customer loyalty. Instead of already established business or service, this work focus on the beginning phase of a service and how adaptation develops over time.

B. Predicting customer churn

In Table I, there are also presented methods applied to churn analysis or churn prediction. In literature more often than not the churn is either predicted by probability or classified in binary classes. Two of the most widely applied techniques are logistic regression and decision trees [15]. Other techniques, such as bayesian inference, partial least square, or support vector machines may produce additional value (*c.f.* [17]). In addition, dimension reduction and feature selection could also be applied if the amount of data or frequency is large (*c.f.* [18]).

In customer activity analysis, a complete churn is usually a rare event. This implies that class imbalance may create problems. Weiss *et al.* [19] names several challenges which may arise from class imbalance. These are for example, improper evaluation metrics, lack of data and noise. Class imbalance in churn prediction has been studied by Poel *et al.* [20]. Boosting techniques could be used in case of lack of data, *c.f.* [17].

C. Definition of customer churn

Defining customer churn in a non-binding registration service is complex and challenging. As described above, the publishing industry and its services compete with the other forms of entertainment, and with partial churners it is difficult to detect the exact time for churn. More importantly the partial churn maybe even more important from business perspective than the actual churn.

ID	m 1	m 2	m 3	m 4	m 5	m 6	Total	
#####	15	35	21	0	0	7	78	Churner
#####	10	12	3	7	15	24	71	Non -churner
#####	35	5	0	0	0	0	78	Churner
#####	76	32	41	92	102	90	435	Non -churner

Figure 2. Examples of derivation of partial churning variable and reading habits.

 ISBN: 978-1-61208-423-7

TABLE I. EXAMPLES OF DATA SETS APPLIED TO THE CHURN PREDICTION IN THE LITERATURE.

Authors	Year	Market sector	Source data	Methods used	Temporal coverage
Ali *et al.* [6]	2014	Banking	7204 customers	survival analysis	2008-2009
Amin *et al.* [7]	2014	Telecommunications	public data sets	rough set	n/a
Coussement *et al.* [8]	2008	Newspaper	90000 subscriptions	svm, random forests	Jan 2002 - Sep 2005
Gunther *et al.* [9]	2011	Insurance	160000 customers	logistic regression	Nov 2007 - May 2009
Jahromi *et al.* [4]	2014	B&B	11021 customers	decision tree	Sep 2011 - Sep 2012
Karahoca *et al.* [10]	2011	Telecommunications	24900 customers	fuzzy clustering	n/a
Lee *et al.* [11]	2012	Telecommunications	114000 customers	nn-classification	n/a
Mutanen *et al.* [12]	2006	Banking	151000 customers	logistic regression	2002 -2005
Xia *et al.* [13]	2008	Telecommunications	two public data sets	factor analysis, svm	n/a
Xie *et al.* [14]	2009	Banking	20000 users	Random forests	n/a

In this study churn is defined as a form of inactivity. There are numerous other ways to define churn, for example based on monetary value [16] or cancellation of an account [9]. We choose first six months of 2014 as the data set for the churn analysis. In Figure 2, examples of the derivation of partial churn are presented. We define churn based on article views. Leaving bounce visits unnoticed, the churners are customer who are inactive during the month in focus.

III. EMPIRICAL ANALYSIS

As mentioned in the introduction, our research questions were formulated based on the visual interpretation of customer lifetimes and frequency of visits. In Figure 1 lifetime distribution of four customer groups were presented. It seems that the churners could be detected already early on the duration of the customer relationship. However, before we can predict churners we will find out if the customer population changes during the time period.

A. Source data

The service in question was launched in the second half of 2013. For this study we selected customers who registered in the first half of 2014. In Figure 3 is shown the amount of customers registered on weekly basis during time period. The selection of customers from early 2014 allowed us to monitor the status of each customer for at least six months since the registration.

The data set in this study included 5956 customers in total.

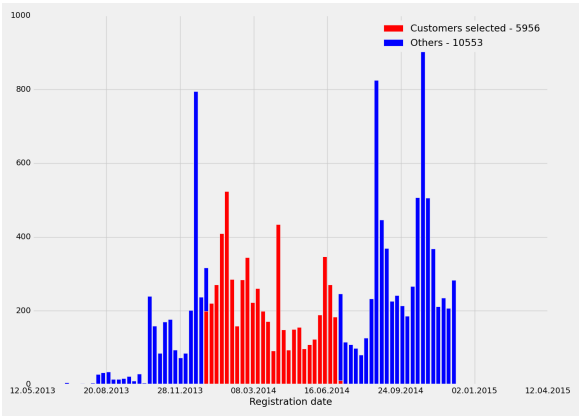


Figure 3. Selected set of customers based on the registration date.

It is worth noting that we selected our data set based on customer IDs and sorted them based on the registration date. However, we used different data sets for reading habit and activity analysis and churn prediction.

For reading habits and activity, we collected and analysed data from each customer of the first six months of their customer lifetime. Examples of this type aggregated values are presented in Figure 2. In the figure data from each customer has been aggregated on monthly basis from six months but for example different months (m1-m6) might not refer exactly to the same period in the calendar between customers.

For customer churn prediction, we selected customers who registered between January and April. Each group consisted of customers registered in a given month. From these customer groups we formulated seven set of customers for churn prediction. Illustration of the source data and churn prediction time period is presented in Figure 4. The division of sets aims to provide information about the reasonable time horizon after registration and would the additional data provide value in customer churn context.

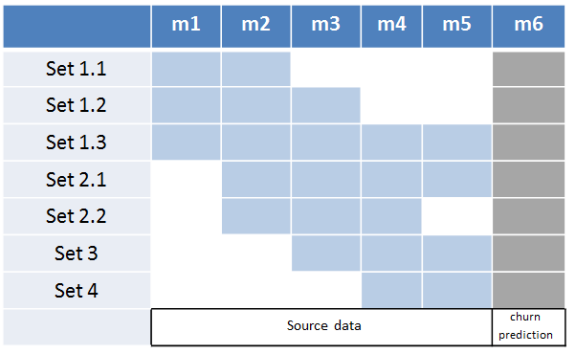


Figure 4. Illustration of the collection period of source data from customers for the prediction of customer churn.

B. Reading habits

One of our research questions was how adaptation of the service change during the beginning of phase of the service. For this we calculated information about frequency of visits for each customer. We also calculated monthly data table for each customer about the articles viewed. In the service for each article view information about the channel was saved. The channel is unique for each piece and describes source

of the content. In Figure 5 examples of the data collected by channels are presented; the figure presents data from one customer. There were 14 different channels in total.

month	'c1'	'c2	'c3'	'c4'	...	'c14'
1	15	0	0	0	0	0
2	5	0	0	0	18	2
3	0	0	30	0	0	0
4	3	0	0	2	0	0
5	1	11	5	2	1	0
6	1	3	7	0	3	0

Figure 5. Examples of the derivation of reading habits variables in various content channels (c1-c14) of one customer.

Based on the reading data and frequency of visits, we classified customers to five activity segments. The five segments were formed to give information about the share of customers registered in a given month which are active after six months from registration. Out of the five segments one included trial users and one passive users. The share of each of these segments are presented in Figure 6. In Figure 6 is visible that the share of trial, passive and different types of active customers remain closely aligned. The customer base can be assumed to be similar among the customers who registered in the beginning of 2014.

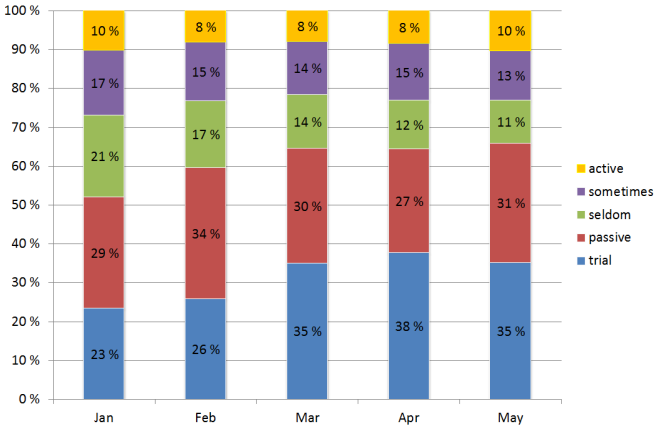


Figure 6. Customers classified to five activity segments based on their reading habits and visit frequency. The segments describe how active a set of customers are after six months from the registration.

C. Customer churn

1) Variables in training and test data: To predict customer churn we selected 24 variables for each customer. The aggregation of source data was done according to the principle presented in Figure 4. Table II presents description of the variables selected. Variables were formulated based on expert knowledge and available data. We selected only customers with lifetime over five days.

For the division of training and test set data, we selected all active customers and twice as many churners. We assigned

TABLE II. SELECTED VARIABLES FOR CHURN PREDICTION.

	Description
1	visit score (average)
2	visit total pages (average per month)
3	number of visit (average per month)
4	number of bounce visits (average per month)
5	visit duration (average)
6	days between visits (average)
7	page view time, hour of the day before noon (average)
8	page view time, hour of the day after noon (average)
9	number of morning pages views
10	number of evening page views
11-24	number of views per channel / active months

customers randomly to either training or test set. In Table III is presented size of each data set. Reason for selecting only part of the churning customers was simply to avoid over estimation of one class.

In addition to the crisp division between training and test sets, we applied five-fold cross validation [21] to the churn prediction model fit.

TABLE III. SIZE OF TRAINING AND TEST SETS IN CHURN PREDICTION. NOTE, THAT AS THE CLASSES WERE NOT BALANCED, ONLY PART OF THE CHURNERS WERE SELECTED.

	Total	Training set *(non-churn/churn)*	Test set *(non-churn/churn)*
Set 1.1-1.3	1166	(143 / 300)	(53 / 94)
Set 2.1-2.2	799	(83 / 169)	(29 / 55)
Set 3	574	(60 / 151)	(34 / 37)
Set 4	483	(83 / 164)	(27 / 56)

2) Prediction: For customer churn we applied logistic regression, *c.f* [17]. Each of the data sets described in Tables II and III were trained separately. Results are presented in Table IV. In Table IV average model fit from five-fold cross validation is presented along with total values for precision, recall and area under the curve (AUC) score. As the churners were the majority population prediction could have been targeted as activity prediction as well. Thus, in Table IV total precision and recall are presented. Two examples of the resulting confusion matrix is presented in Table V. Confusion matrix present the actual values of classification results.

TABLE IV. CLASSIFICATION RESULTS FOR EACH SET. PRECISION AND RECALL ARE TOTAL VALUES FOR THE CLASSIFICATION RESULT.

	Area Under the Curve (AUC) score	Precision total	Recall total	model fit avg, 5-fold cv
Set 1.1	0.66	0.69	0.68	0.70
Set 1.2	0.67	0.70	0.70	0.72
Set 1.3	0.77	0.69	0.70	0.74
Set 2.1	0.62	0.66	0.68	0.71
Set 2.2	0.64	0.66	0.68	0.71
Set 3	0.49	0.56	0.55	0.69
Set 4	0.76	0.78	0.77	0.66

TABLE V. CONFUSION MATRIX OF THE SET 1.3 AND SET 4.

	predicted churn	set 1.3 active	predicted churn	Set 4 active
churn	82	12	54	2
active	32	22	17	10

 ISBN: 978-1-61208-423-7

IV. CONCLUSION

In this study, service adaptation and customer churn were analysed. We asked three questions in the study. The initial hypothesis were made based on visual interpretation of the data, see Figure 1. For each of these questions, we formulated suitable metrics and data based support. Classificaiton and logistic regression techniques were applied in the analysis.

Based on the findings, it seems that the user characteristics and behaviour does not change significantly. The so called early adopters do not have notable differences in usage behaviour. Thus, we can compare customer behaviour from different time periods to those of another. The adaptation overall is tilted towards short customer lifetimes. Majority of the customers have tried the service and thus we classified them as trial users.

As the service was novel at the time of data collection and the service has non-binding registration, we analysed the time period for how long it is reasonable to follow inactive customers. We are aware that all customers are important and no service provider will cancel any account on their part. However, as the trial and test users were significant population, for the service provider it is beneficial to have understanding of the lifetimes of the recently registered customers.

From the prediction results can be seen that the best results are found in either from the very recent data set or from the most extensive data set. For our research question, it is reasonable to conclude that more data produces higher accuracy for predictions.

A. Future research directions

In this work, churn was defined based on customer activity. Future research could focus on other systematic methods to formalize churning customers and active customers. The non-binding registration is challenging for the service provider as the exact moment of churn is difficult to detect. However, as registration already exists, the service development could benefit from a closer user studies in the future. Another beneficial direction for the future research could be to include and research usefulness of other variables.

It is unclear for us how these results can be generalized over a wider range of services. The churn studies are already extensive in academic literature. It would be beneficial to extent and validate these findings with other databases in other services. As the churn prediction is usually combined with customer retention campaign, it would be beneficial to conduct comprehensive studies with retention campaign and a follow-up. In [22] is presented directions and ideas how analytics could benefit the CRM overall.

From the methodology point of view, this study applied only logistic regression. Future studies may utilize other methods and scoring variations. For example, decision trees might provide value and additional information about the source data and variables. In case the amount of usage data and possibly the number of users will increase, feature selection methods could be applied before the classification and prediction results are computed.

REFERENCES

[1] E. M. Rogers, Diffusion of innovations. Simon and Schuster, 2010.

[2] J. Rowley, "An analysis of the e-service literature: towards a research agenda," Internet Research, vol. 16, no. 3, 2006, pp. 339–359.

[3] M. Ballings and D. V. den Poel, "Customer event history for churn prediction: How long is long enough?" Expert Systems with Applications, vol. 39, no. 18, 2012, pp. 13 517 – 13 522.

[4] A. T. Jahromi, S. Stakhovych, and M. Ewing, "Managing {B2B} customer churn, retention and profitability," Industrial Marketing Management, vol. 43, no. 7, 2014, pp. 1258 – 1268.

[5] J. Lee, J. Lee, and L. Feick, "The impact of switching costs on the customer satisfaction-loyalty link: mobile phone service in france," Journal of services marketing, vol. 15, no. 1, 2001, pp. 35–48.

[6] Özden Gür Ali and U. Aritürk, "Dynamic churn prediction framework with more effective use of rare event data: The case of private banking," Expert Systems with Applications, vol. 41, no. 17, 2014, pp. 7889 – 7903.

[7] A. Amin, C. Khan, I. Ali, and S. Anwar, "Customer churn prediction in telecommunication industry: With and without counter-example," in Nature-Inspired Computation and Machine Learning, ser. Lecture Notes in Computer Science, A. Gelbukh, F. Espinoza, and S. Galicia-Haro, Eds. Springer International Publishing, 2014, vol. 8857, pp. 206–218.

[8] K. Coussement and D. V. den Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," Expert Systems with Applications, vol. 34, no. 1, 2008, pp. 313 – 327.

[9] C.-C. Günther, I. F. Tvete, K. Aas, G. I. Sandnes, and O. Borgan, "Modelling and predicting customer churn from an insurance company," Scandinavian Actuarial Journal, vol. 2014, no. 1, 2014, pp. 58–71.

[10] A. Karahoca and D. Karahoca, "{GSM} churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system," Expert Systems with Applications, vol. 38, no. 3, 2011, pp. 1814 – 1822.

[11] Y.-H. Lee, C.-P. Wei, T.-H. Cheng, and C.-T. Yang, "Nearest-neighbor-based approach to time-series classification," Decision Support Systems, vol. 53, no. 1, 2012, pp. 207 – 217.

[12] T. Mutanen, J. Ahola, and S. Nousiainen, "Customer churn prediction - a case study in retail banking," in ECML/PKDD 2006 Workshop on Practical Data Mining: Applications, Experiences and Challenges, 2006.

[13] G. en XIA and W. dong JIN, "Model of customer churn prediction on support vector machine," Systems Engineering - Theory & Practice, vol. 28, no. 1, 2008, pp. 71 – 77.

[14] Y. Xie, X. Li, E. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," Expert Systems with Applications, vol. 36, no. 3, Part 1, 2009, pp. 5445 – 5449.

[15] S. A. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. H. Mason, "Defection detection: Measuring and understanding the predictive accuracy of customer churn models," Journal of Marketing Research, vol. 43, no. 2, 2006, pp. 204–211.

[16] V. Miguéis, D. V. den Poel, A. Camanho, and J. F. e Cunha, "Modeling partial customer churn: On the value of first product-category purchase sequences," Expert Systems with Applications, vol. 39, no. 12, 2012, pp. 11 250 – 11 256.

[17] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[18] A. Rajaraman and J. D. Ullman, Mining of Massive Datasets. New York, NY, USA: Cambridge University Press, 2011.

[19] G. M. Weiss, "Mining with rarity: a unifying framework," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, 2004, pp. 7–19.

[20] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," Expert Systems with Applications, vol. 36, no. 3, 2009, pp. 4626–4636.

[21] R. Kohavi et al., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Ijcai, vol. 14, no. 2, 1995, pp. 1137–1145.

[22] A. Tuzhilin, "Customer relationship management and web mining: the next frontier," Data Mining and Knowledge Discovery, vol. 24, no. 3, 2012, pp. 584–612.

 ISBN: 978-1-61208-423-7

An Approach to Highly Available and Extensible Data Management Systems for Large Scale Factory Floor Data

Jaehui Park and Su-young Chi
Knowledge Convergence Service Research Team
Electronics and Telecommunications Research Institute
Daejeon, Korea
email:{jaehui,chisy}@etri.re.kr

***Abstract*—In recent years, Big data prevailed in various domains such as marketing, manufacturing and finance. Although many analytical models and practical systems have been studied, it is not a typical task to adopt them to domain problems. There exists a gap between understanding the problems in the field and adapting the Big data techniques. This work proposes a data-driven framework bridging the gap between existing data management issues and problems on the shop floors in manufacturing industries. In this paper, we propose a highly available and extensible data management system to integrate manufacturing data with regard to four factors: man, machine, material and method. This data management system supports a large scale data in terms of reliability and efficiency for data collectors and analytical systems. Furthermore, as an ongoing study, we have investigated a set of requirements and practical problems from field-workers rather than executive managers. We formulated a comprehensive system structure towards a predictive manufacturing system that can capture, in advance, potential risk factors, such as, machine worn out progress and production time loss tendency.**

Keywords; Big data; data-driven framework; high availibility; extensibility; predictive manufacturing.

I. INTRODUCTION

With the rapid growth of sensors and communications, various data from the real world comes into the digital world with large volume and heterogeneity and at high speed. The words 'Big data' have been attracting much interest from various domains such as marketing, manufacturing and finance. Big data concerns many aspects of recent data issues, for example, gathering data from scattered sources, storing data in persistent storages with efficiency and reliability, and analyzing data to bring business insights. Although many analytical models and practical systems in Big data environment have been studied, it is not a typical task to adopt them directly to field problems. There exists a gap between understanding problems in the field and adapting the Big data techniques. For example, industrial enterprises, which are in relatively conservative domains, are seeking business insights within their ERP/SCM information while they lack practices of associating web data, social networks and even with the their shop floor data.

Manufacturing execution systems (MES) manage product life-cycle, resource schedules, order execution and dispatch and so on. MES is known as a production information system for manufacturing decision makers. A MES provides useful insights for the managers in industrial enterprises to make decisions to optimize the production performance. Many MES vendors, for example, SAP or Rockwell Automation, try to adopt recent big data techniques to improve the performance of their solutions. However, we note that utilizing MES capabilities does not always improve manufacturing performance. This is due to the nature that MES is a system for the executive managers, and is not coping with the requirements of field workers. However, field workers concerns on more specific problems on the shop floors, such as, optimal settings of machine equipment. These are a variety of practical problems issued by technical workers on the factory floor, which cannot be solved by existing solutions, such as MES/ERP/SCM. This work aims at proposing a data-driven framework bridging the gap between recent data management issues and actual problems on the shop floors in manufacturing industries.

In this paper, we propose a highly available and extensible data management system for manufacturing data. This paper is a work-in-progress report of the ongoing project "Development of a predictive manufacturing system using data analytics in small and medium enterprises (SME)". This report mainly describes a system architecture for industrial data management systems and integrated data structure design. Our contributions summarize as follows: 1) a highly available data management server architecture and 2) an extensible manufacturing database model for integrating data with regard to four factors: *man*, *machine*, *material* and *method*. Furthermore, we have investigated a set of requirements and practical problems from workers on the shop floors rather than executive managers. We try to formulate a prototype for predictive manufacturing system that can capture, in advance, potential risk factors, such as, machine worn out progress and production time loss tendency. The system is expected to enhance the existing manufacturing enterprises complementing the vision of smart factories [1], factory of the future [2], industry 4.0 [3]. We plan to build a test-bed factory to evaluate our proposed

 ISBN: 978-1-61208-423-7

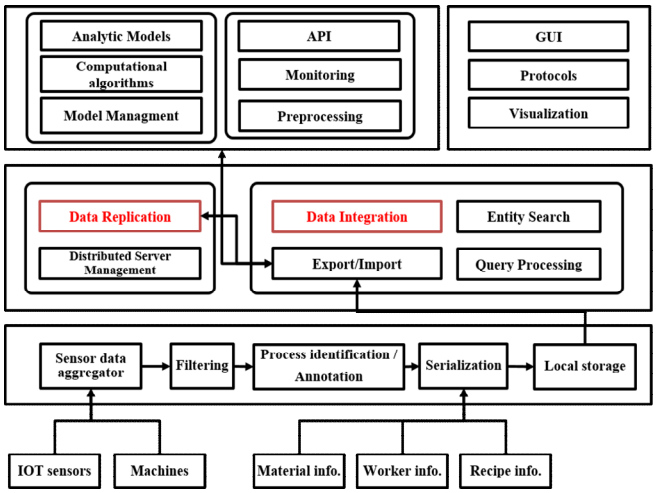


Figure 1. An architecture of a predictive manufacturing system.

system in the domain of manufacturing automobile components, especially, motor shafts. Discussions including the current issues and plans will be presented in the last sections.

II. ARCHITECTURE

To introduce the integrated data model, we first introduce an architecture of *predictive manufacturing systems* (PMS). PMS provide four sub-modules to 1) gather data from different sources and serialize them, 2) integrate the data and replicate them in the distributed server systems, 3) build analytics models learned from the data, and compute future events 4) visualize data. Figure 1 shows the sub-modules. In the bottom level, available data sources are identified, which can be extended. *Data Collector* module aggregates the raw data and annotates/serializes into meaningful inputs. This module can be regarded as supervisory control and data acquisition (SCADA) system. The difference of our system include ad-hoc sensors, so called, internet-of-things (IOT) sensors to acquire the data that cannot be obtained from machine internals. Collectors can be easily extended by simply adding IOT sensors based on the problem solving requirements gathered directly from the field workers. For a trial, we plan to set up a video system to acquire vision data of material form changes in milling machines. *Analytics System* leverages the data gathered from the factory floor. It builds analytical models using machine learning algorithms. Based on the model, it predicts events to provide the workers get an insights to solve the problems. For example, based on the vision data gathered from the sensors, the system learns the patterns of anomalies to machine faults or product defectives. In our test-bed, we try to detect the occurrence of rolled chips, which are residue of milling process, in milling machines. Although the problems are well-known in the fields of milling processes, there is not a systematic solutions. The solution has been relied on experience of the field workers. The benefit of our systems is that the faced-problems raised from the factory shop floors can be solved. Existing solutions, i.e., MES solutions, cannot be suitable to

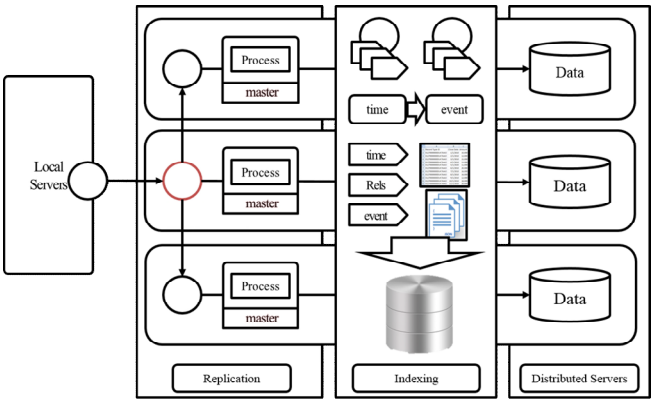


Figure 2. A processing framework.

this kind of field problem solving because they focus on the upper levels of application and the lower levels of data. Hence, we believe our systems complements the existing systems.

In this section, we report the sub-modules for 1) Data Replication and 2) Data Integration (colored in red in Figure 1.) First, we build replicated server system to achieve the availability of aggregated industrial data. Every data element is a type of codes and numbers with temporal information. Therefore, it can be easily maintained in a distributed system. Figure 2 shows the conceptual front-end architecture of data replication module. This module gets input from local servers and replicate them. In our implementation, we use wsrep APIs for synchronous write-set replication. Because industrial data should be process in a transactional manner, we should guarantee that each data nodes in the replication cluster process input in the same order and uninterrupted. Replicated data can be indexed in distributed manner. Distributed server maintains the data using uniform structures we defined the next section. High availability is a distributed server system characteristic to maximize up time to running time. To tolerate down time, the system show eliminate of single points of failure. This means adding redundancy to the system both at data level and at system level. Also, failure detection should be provided. Maintenance activities should be provided. However, server maintenance is out of scope. Manufacturing environment always has scheduled downtime, for example, equipment maintenance. In our implementation, we avoid the single point of failure by introducing multi master replication. This enables high availability with no slave lag and no lost transactions. Data Replication server implementation is illustrated in Figure 3. We construct a physical system using three nodes with Intel Xeon E5 processors, 16G DDR3 memories, and 10TB HDDs. The physical system can be virtualized in clouds. Mater-master replication and distributed database management systems are organized using MariaDB galera clusters. We developed user interface using MariaDB connectors. To communicate with external modules, we define protocols using XML messages and SQL parameters. This implementation is still in progress.

Second, we define a database structure for the Data Integration module. This sub-module identifies the semantics

 ISBN: 978-1-61208-423-7

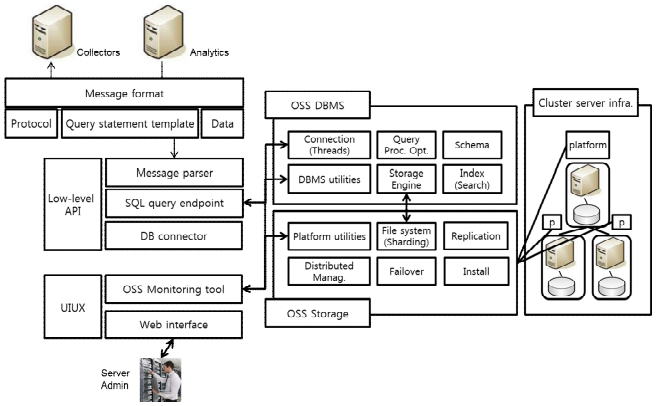


Figure 3. Data Replication server implementation.

from the shop floor data, for example, control codes in the numerical control machines and its continuous changes. The control codes implement the behaviors of the machines to obtain products from computer-aided design programs. Therefore, to obtain a predictive model by understanding the underlying data characteristics, the semantics should be designed in the database structure. Data integration task involves dealing with flexible adaptation of a variety of data. We note manufacturing domains share some common factors to identify shop floor data. We extract to factors in four categories: *man, machine, material and method*. Man denotes the category for personnel information, shift information, performance, and so on. Human factors in the manufacturing process can be utilized. Machine denotes the category for machine internal information, such as machine status, its logs, equipment detail, axis of machining tools, fault/alarm history and so on. Material denotes the category for the specifications of input materials and the intermediate results from previous process. Method denotes the category for machine parameters to yield an end product. We refer to the parameter settings as *recipe*. This information can be used to improve the production performance of the factory if we can produce a gold recipe by optimizing the parameter settings. The theoretical study parts of our project try to optimize the milling machine parameter to reduce initial setting time. To support the study, we provide integrated data structure to gather the features to build effective models. Based on our data model, all of those factors can be visible and utilized in a uniform way.

III. DATA MODEL OF DATA INTEGRATION MODULE

To understand the data with four factors in shop floor, we first analyze the data sheets, which are created manually. This data is maintained with sloppy identifies and different types of documents. We refine and categorize the semantics to integrate them with man, machine, material and methods. Figure 4 illustrates the conceptual image to categorize the manual data. Identifier parts comprise id, location, time, and man. Based on the identifier information we can break data into several pieces of similar characteristics into a single information unit. First, quality control data contains inspection item, allowance, inspection frequency, defective

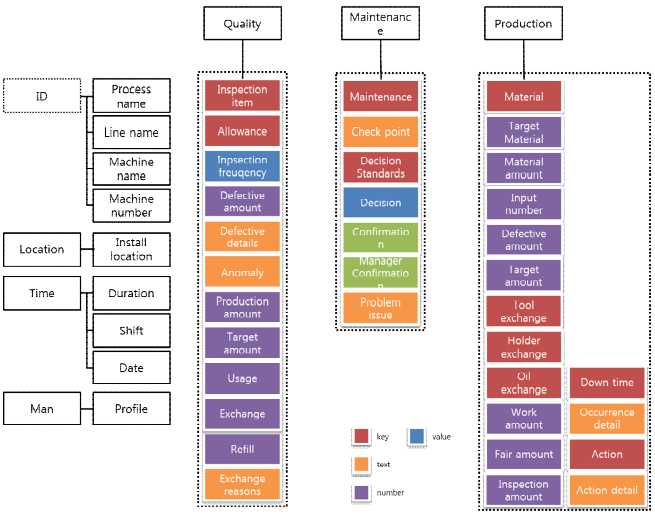


Figure 4. Manual data (production documents) mapping to the integrated database structure.

amount, tool usage/exchange/refills and so on. This information can be utilized to labels for machine learning algorithms in the analytics system. Also, production data contains amount data of material, fair/defective products, up/down time, occurrence/action time. This information explains the status of entire production processes. Maintenance information some temporal data for maintenance event for machine. This can be used for relationships to machine data. Figure 5 illustrate our first draft of integrated data structure. There are two main identifier p_code and m_code, which represents process identifier and machine identifier, respectively. Every event associated to machine, such as inspection, maintenance, machine errors are pre-defined by technical workers. Based on the definition, which is denoted by 'standards', the system records the tuple of 1) items in the standards, 2) temporal information and 3) values. This structure is flexible because every event can be extended by adding standard/log tables to the four factors. Equipment, material, process and others associate above three information. Our data structure can log every transactional and non-transactional information. In the context of server system (presented in Section 2), the vertical row-based data structure can be appropriate to multi-master replication setting. Our data model design cannot be a single solution to maintain the manufacturing data. However, our model is flexible enough to extend to various data sources.

IV. CONCLUSION

We are currently working on surveying a set of requirements and practical problems from field-workers rather than executive managers. Traditionally, in manufacturing information systems (MIS), data of production, costs, labor, warehousing and so on has been important because entrepreneurs and managers are more interested in it rather than the shop floor environment. As the number of machines and facilities increases and they are automated, field workers can be faced to real world problems

 ISBN: 978-1-61208-423-7

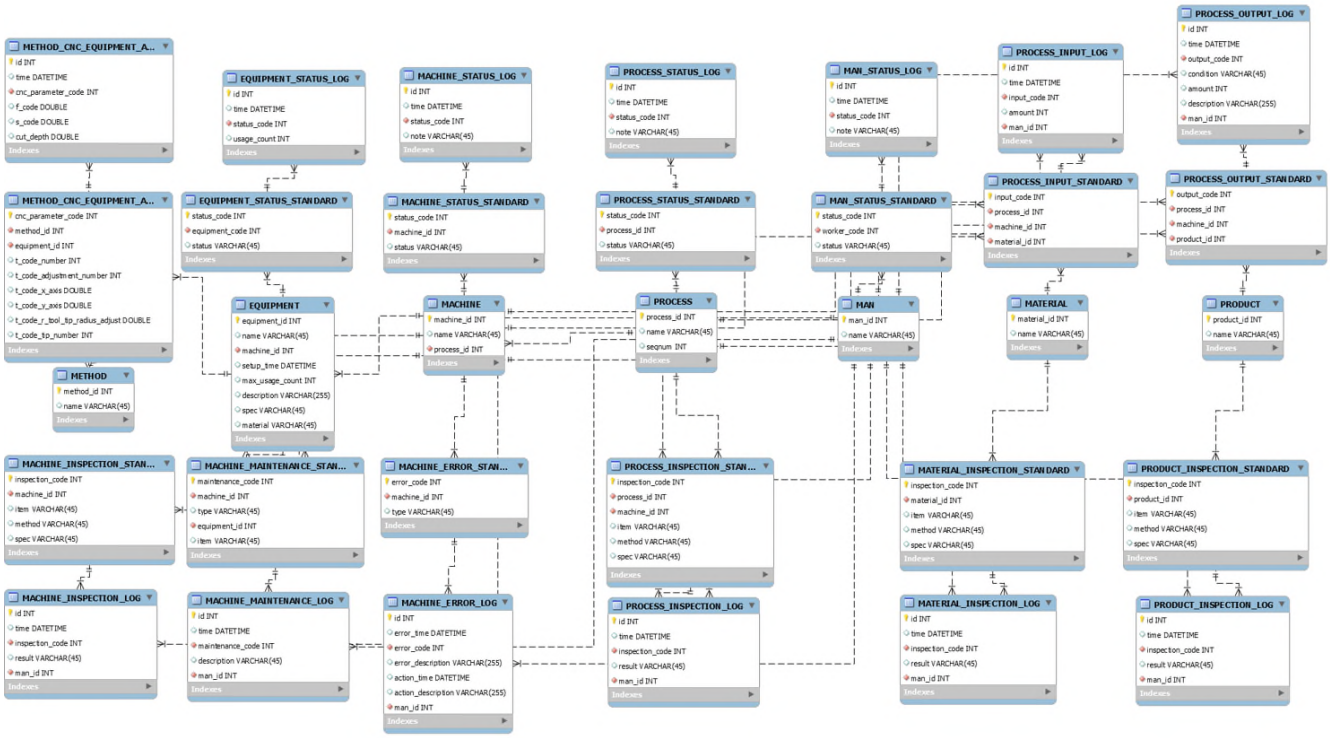


Figure 5. A data model for integrating manufacturing data.

not stretching to the upper levels. Therefore, there should be know-hows and experience that should be maintained by human intervention. However, we are currently formulating a comprehensive system structure towards a predictive manufacturing system that can capture, in advance, potential risk factors, such as, machine worn out progress and production time loss tendency. Our system reduces the human intervention in the process of manufacturing. Also, prediction results may reduce the down time of production to improve performance.

As a further work, we attach streaming data processing part at the input port of the replication modules. Such parts could be useful for computing descriptive statistics in real time manner, which can reduce the burden of analytics system. There are several well-known tools, for example, Apache Storms and Spark. We plan to utilize those tools for processing streaming data.

ACKNOWLEDGMENT

This work was supported by the Knowledge Services Industry Core Technology Development Program (10051028, Development of Predictive Manufacturing System using Data Analysis of 4M Data in Small and Medium Enterprises) funded By the Ministry of Trade, Industry & Energy (MI, Korea).

REFERENCES

[1] F. Shrouf, J. Ordieres, and G. Miragliotta. “Smart factories in Industry 4.0: A review of the concept and of energy management approached in production based on the Internet of Things paradigm,” Industrial IEEE International Conference on Engineering and Engineering Management (IEEM), 9-12., Dec. 2014. pp. 697-701

[2] http://www.economist.com/blogs/schumpeter/2013/10/manufacturing [retrieved: May, 2015]

[3] H. Kagermann, W. Wahlster, and J. Helbig. Recommen-dations for implementing the strategic initiative Industrie 4.0: Final report of the Industrie 4.0 Working Group. 2013

 ISBN: 978-1-61208-423-7

Fuzzy Clustering Based Approach to Network Traffic Classification and Anomaly Detection

Julija Asmuss, Gunars Lauks
Institute of Telecommunication
Riga Technical University
Riga, Latvia
e-mail: julija.asmuss@rtu.lv, gunars.lauks@rtu.lv

***Abstract*— In this work, we develop network traffic classification and anomaly detection methods based on traffic time series analysis using fuzzy clustering. We compare four fuzzy clustering techniques using different dimensionality reduction methods and validity indices to work out an effective anomaly detection algorithm. The effectiveness of the proposed classification system is evaluated on traffic data with and without traffic attack components.**

Keywords- fuzzy clustering; fuzzy transform; traffic classification; anomaly detection.

I. INTRODUCTION

The ability to classify and identify network traffic is the main area of interest for many network operation and research topics such as traffic engineering, monitoring, pricing, security, anomaly detecting. Anomalous traffic or unwanted traffic definition is still very fuzzy and immensely varies among networks. But it is clear that anomalies (such as Distributed Denial-of-Service (DDoS) attacks [1], for example) may cause significant variances in a network traffic level, and as a result, legitimate user requests can not get through the network. Our work mainly focuses on flood attacks. The most common DDoS flood attacks target the computer networks bandwidth or connectivity. In this context, traffic volume analysis is considered to be a sensitive tool for anomaly detection.

Many monitoring schemes against DDoS attacks have been reported in the literature (see, e.g., [2][3][7][9][12] [13][16]), but only a few of them have been applied in a real network environment. In the context of balance between computation speed and classification success, the task of network traffic anomaly detection is still very important [1].

Our research is devoted to anomaly detection methods based on traffic time series analysis using clustering technique. We follow the idea, which is based on anomalous traffic profile deviation from normal traffic profile, defined empirically on the basis of previously collected information on the properties of normal traffic conditions. The traditional approach to clustering can not be effectively used in this case due to the dynamic behaviour of network traffic in its development over time. This dynamic behaviour should be taken into account when solving traffic classification problems. Traffic time series may belong to one cluster during a certain period; afterwards, its profile may be closer to another cluster. This switch from one state to another can be naturally modelled using a fuzzy approach. We show that fuzzy logic based techniques allow us to deal effectively with the vague and imprecise boundaries between normal traffic and different levels of attacks.

The remainder of the paper is structured as follows. Section II introduces objectives and tasks of this research. Section III contains traffic data representation tools. Sections IV and V are devoted to the clustering and classification stages, respectively. Section VI contains the description of experiments and results.

II. OUTLINE OF RESEARCH OBJECTIVES

Our aim in this research is to use the advantages of the fuzzy logic based approach for analysis of network traffic dynamics and to suggest traffic classification and anomaly detection mechanisms based on fuzzy transforms, fuzzy clustering and classification with good computation speed and classification success characteristics. The tasks of the research follow directly from its objectives:

- To develop a fuzzy transform technique for representation of network traffic and to investigate its advantages in traffic data compression;
- To evaluate the performance of different fuzzy clustering methods for solving the traffic classification problem and to identify the best one;
- To develop the mechanism of traffic classification and anomaly detection using traffic fuzzy clusters and self-similarity characteristics;
- To evaluate the performance of the suggested method and to compare it with the existing solutions.

III. TRAFFIC DATA PRE-PROCESSING

We work with time series that represent aggregated traffic and we compare two dimensionality reduction methods. Usually, the dimension of time series representation is reduced using Piecewise Aggregate Approximation (PAA) [17]. In this research, we also apply time series representation based on Fuzzy Transform (for short, F-transform) introduced in [8].

Generally, F-transform depends on a chosen fuzzy partition, which consists of fuzzy sets given by membership functions. We apply uniform fuzzy partition with the triangular form of membership functions and use discrete formulas for F-transform components.

 ISBN: 978-1-61208-423-7

The idea of using fuzzy transforms in analysis of time series is not new (see, e.g., [4][5]). Our research develops the F-transform technique as a special tool for traffic data aggregation and investigates its role in reduction of traffic classification computational resources.

IV. FUZZY CLUSTERING TECHNIQUE

We consider four fuzzy clustering algorithms and evaluate their performance for our purposes: Fuzzy C-Means (FCM); Possibilistic C-Means (PCM); Possibilistic Clustering Algorithm (PCA); Unsupervised Possibilistic Fuzzy Clustering (UPFC) [6][11][14][15]. For each of them we consider also Gustafson-Kessel modification GK (see, e.g., [6]).

We apply four validity indices for determining the number of clusters: modified partition coefficient, Fukuyama and Sugeno index, Xie and Beni index, separation and compactness index (see, e.g., [10]). Taking into account that the result obtained by using each index can be interpreted as evaluation done by an expert, we apply the technique of aggregation of expert opinions.

Thus, at the clustering stage we obtain fuzzy clusters and cluster centroids given by their F-transform components or PAA components, correspondingly.

V. FUZZY CLASSIFICATION AND ANOMALY DETECTION

For the traffic classification merit we use the prototypes obtained at the previous stage. Decision making on classification of a new traffic time series is done in the following way. We suppose that we have new infinite time series, therefore the algorithm runs infinite time too. At each moment in time, the classification is done considering a finite number of time series components and their F-transform (or PAA) components, correspondingly.

In each next step, we start with the computation of F-transform (or PAA) components (as compared with the previous step, the first component is removed and one new component is added to the end). Then, we compute the membership degrees with respect to all clusters of normal (in some cases, of anomalous also) traffic. Next, we evaluate self-similarity parameter changing rate. Finally, decision making on the risk of anomalies is done on the basis of the above mentioned membership degrees and Hurst parameter changing rate by using the fuzzy rule based technique.

VI. EXPERIMENTS AND RESULTS

When studying a traffic classification technique with real traces, it is important to have a baseline for traffic classification that will be used as a reference. Because it is very difficult to obtain a dataset that is representative of real network activities and contains both normal and anomalous traffic, attack traffic for numerical experiments was generated and added as additional component of traffic data.

To evaluate the effectiveness of the proposed technique, we consider all major steps:

- Traffic data pre-processing, the main merit of which is to reduce the amount of traffic data and to allow a more effective use of data analysis techniques, both in time and space;
- Extracting the relationship between traffic data by using fuzzy clustering methods to characterize patterns, which identify normal network traffic;
- Detecting traffic anomalies by using fuzzy rule based prototypical classification methods.

The algorithm consists of two stages: fuzzy clustering, which can be executed only once, and real-time classification. When evaluating the computation time per classification, the clustering stage is not taken into account.

The results obtained by comparing different techniques show that the best compromise between computation speed and classification success is achieved using F-transforms for traffic time series representation and applying UPFC-GK clustering algorithm. The detection rate (DR) in our experiments with this technique was greater than 99%; the false alarm rate (FAR) was about 1% in the worst cases. A comparison with methods based on statistical analysis (D-WARD, DCD, T-test model) is done using DR and FAR evaluation results shown in [16]. For the exact comparison and performance evaluation, it is necessary to obtain results for real traffic classification in the same testing conditions.

ACKNOWLEDGMENT

This work has been supported by the European Social Fund within the project 2013/0024/1DP/1.1.1.2.0/13/APIA/VIAA/045 „Applications of mathematical structures based on fuzzy logic principles in the development of telecommunication network design and resource control technologies”.

REFERENCES

[1] Computer Emergency Response Team (CERT-EU), “DDoS overview and incident response guide,” 2014. [Online]. Available from: http://cert.europa.eu/static/WhitePapers/ [retrieved: May, 2015]

[2] K. Lee, J. Kim, K. H. Kwon, J. Han, and S. Kim, “DDoS attack detection method using cluster analysis,” Expert Systems with Applications, vol. 34, 2008, pp. 1659–1665.

[3] M. Li, “Change trend of averaged Hurst parameter of traffic under DDOS flood attacks,” Computers & Security, vol. 25, 2006, pp. 213–220.

[4] V. Novák, I. Perfilieva, M. Holčapek, and V. Kreinovich, “Filtering out high frequencies in time series using F-transform,” Information Sciences, vol. 274, 2014, pp. 192–209.

[5] V. Novák, M. Štepnička, V. Dvorák, I. Perfilieva, V. Pavliska, and I. Vavričková, “Analysis of seasonal time series using fuzzy approach.,” International Journal of General Systems, vol. 39, 2010, pp. 305–328.

[6] J. V. de Oliveira and W. Pedrycz, “Advances in fuzzy clustering and its applications,” John Wiley and Sons, 2007.

[7] T. T. Oo and T. Phyu, “A statistical approach to classify and identify DDoS attacks using UCLA Dataset,” Int. J. of Advanced Research in Computer Engineering & Technology, vol. 2, No. 5, 2013, pp. 1766–1770.

[8] I. Perfilieva, “Fuzzy transforms: Theory and applications,” Fuzzy Sets and Systems, vol. 157, 2006, pp. 993–1004.

[9] S. N. Shiaeles, V. Katos, A. S. Karakos, and B. K. Papadopoulos, “Real time DDoS detection using fuzzy

 ISBN: 978-1-61208-423-7

estimators", Computers & Security, vol. 31, 2012, pp. 782–790.

[10] W. Wang and Y. Zhang, "On fuzzy cluster validity indices," Fuzzy Sets and Systems, vol. 158, 2007, pp. 2095 –2117.

[11] X. Wu, B. Wu, J. Sun, and H. Fu, "Unsupervised possibilistic fuzzy clustering," Journal of Information & Computational Science, vol. 7(5), 2010, pp. 1075–1080.

[12] Z. Xia, S. Lu, J. Li, and J. Tang, "Enhancing DDoS flood attack detection via intelligent fuzzy logic," Informatica, vol. 34, 2010, pp. 497–507.

[13] Z. Xiong, Y. Wang, and X. F. Wang, "Distributed collaborative DDoS detection method based on traffic classification features", Proc. of International Conference on Computer Science and Electronics Engineering ICCSEE 2013, Atlantic Press, Paris, 2013, pp. 93–96.

[14] M.-S. Yang and K.-L. Wub, "Unsupervised possibilistic clustering," Pattern Recpgnition, vol. 39, 2006, pp. 5–21.

[15] R. J. Almeida and J. M. C. Sousa, "Comparison of fuzzy clustering algorithms for classification," Proc. of International Symposium on Evolving Fuzzy Systems EFS'06, Lake District, United Kingdom, 2006, pp. 112–117.

[16] M. H. Bhuyan, H. J. Kashyap, D. K. Bhattacharyya, and J. K. Kalita, "Detecting distributed denial of service attacks: methods, tools and future directions," Computer Journal, vol. 57, 2014, 537-556.

[17] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality reduction for fast similarity search in large time series databases," Knowledge and Information Systems," vol. 3, 2001, pp. 263–286.

 ISBN: 978-1-61208-423-7

Data-Driven Approach for Analysis of Performance Indices in Mobile Work Machines

Teemu Väyrynen, Suvi Peltokangas, Eero Anttila, and Matti Vilkko
Department of Automation Science and Engineering
Tampere University of Technology,
Korkeakoulunkatu 10, Tampere, Finland
e-mails: firstname.lastname@tut.fi

***Abstract*—This paper presents a data-driven approach for the analysis of performance indices in mobile work machines. Performance analysis and optimisation of mobile work machines has become increasingly important in recent years. The mobile work machine optimisation is performed based on performance measurements. One of the most interesting and potential approach for improving the quality of the performance analysis is the utilisation of Big Data and data-driven analysis methods, such as machine learning. This study utilises a machine learning algorithm, Classification and Regression Trees (CART), in the performance analysis of the mobile work machines. The most significant benefit of the presented method is that it provides a statistical reference of the machine performance for the operators. The method enables operators to compare performance against reference fleet of machines working in similar operating conditions. This feature can lead to more informative and reliable interpretations and analysis of the performance values. The results of this paper demonstrate how the presented method was used to analyse the performance of a mobile work machine fleet.**

Keywords–performance; mobile work machine; regression tree; CART.

I. INTRODUCTION

Performance analysis and optimisation of mobile work machines has become an increasingly important trend within the industry in the recent years [1], [2]. Both, the mobile work machine manufacturers as well as the operators have started to pay more attention to the performance optimisation of the machines. Optimising the performance of the mobile work machine results in increased productivity and efficiency. However, the optimisation of the mobile work machine is difficult if the performance of the machine cannot be measured and analysed accurately. The importance of the performance analysis is the main motivation for this work.

The objective of this work is to present a data-driven approach that utilises machine learning to assist the operators in the performance analysis of the mobile work machines. The approach is a combination of data preprocessing and Classification and Regression Trees (CART). CART is a supervised machine learning algorithm, that constructs classification and regression trees to model systems [3]. In this work, CART is used to model the relation between the different operating conditions and the performance of the machines based on the data of a mobile work machine fleet. The predictions of the model enable operators to compare the performance against reference fleet of machines working in similar operating conditions. This feature can provide more informative and reliable interpretations and analysis of the performance values.

The performance analysis of a mobile work machine is a challenging task due to the various factors affecting the performance. These factors are, e.g. objectives of the work, operating conditions, skill level of the operator, work load, technical properties of the mobile work machine, and control parameters. Figure 1 describes the factors affecting the performance of the machines. This work focuses on analysing the relation between the operating conditions and the performance of the mobile work machine.

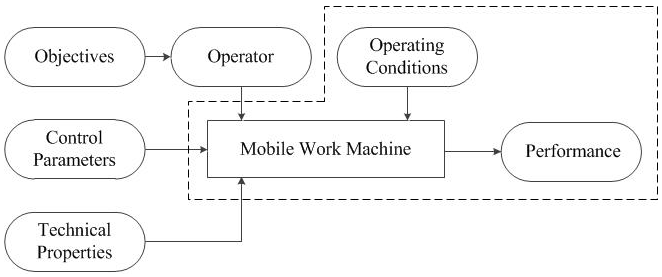


Figure 1. Factors affecting the performance of a mobile work machine. The main focus of this work is delimited by the dotted lines in the figure.

Improved performance analysis enables the operators to optimise the operations of the mobile work machine by tuning the control parameters of the automation system. Depending on the complexity of the machine, the automation system can allow operators to customise hundreds of parameters based on their personal preferences and requirements of the operating conditions. These parameters have a major impact on the operational performance of the mobile work machine in terms of efficiency and productivity.

Conventionally, parameter optimisation has been performed based on the rules of thumb developed by skilled instructors and machine operators. The proper tuning of a mobile work machine is an extremely difficult and time-consuming task especially for an inexperienced operator [4]. Various measurement and performance values can be presented to the operators via the graphical user interfaces of the machines. However, the interpretation of the performance values is most often left to the operators.

These interpretations are often made without proper understanding about the relation between the operating conditions and the performance of the machine. Also, due to the restricted performance analysing capabilities of the human operators, the results of the analysis might be incorrect. However, by utilising reference data and advanced data analysis methods, the

 ISBN: 978-1-61208-423-7

operators gain valuable information to support their analysis of the machine performance.

Originating from the described situation, the research problem of this work focuses on improving the analysis of the performance values in mobile work machines. Derived from the identified problem, the research question of this work is: How can the analysis of performance values be improved in mobile work machines?

The rest of this paper is organized as follows. Section II addresses the state of the art in analysis of performance values and describes the requirements set for the solution. Section III introduces data preprocessing and the CART algorithm. Section IV describes the design of experiment and the results. Section V sums up the work and proposes future research topics.

II. State of the art and requirements

This section introduces the state of the art in the analysis of the performance values in mobile work machines. The requirements set for the solution method are also introduced in this section.

A. State of the art

A wide range of methods have been applied to analyse the performance values of mobile work machines. These methods vary from simple monitoring of measurement values to more sophisticated and holistic analysis. The requirements of the data analysis and the application specific features of the data determine which method is most suited to the given application.

Data-driven performance analysis methods, which require domain expert knowledge, have been presented for mobile work machines and industrial processes [1], [2]. Various other research papers have addressed the problem of analysis and optimisation of performance values in mobile work machines [5]–[9]. These studies have used such methods as statistical data analysis, modelling, root-cause analysis, and optimisation to improve the operational performance of the mobile work machines. Previous research with machine learning algorithms has provided promising results also in the fields of agriculture and industry [10]–[12]. Due to privacy policy of the mobile work machine industry, it is difficult to find up-to-date information about data analysis methods utilised in the analysis of performance values in mobile work machines.

B. Requirements set for the data analysis method

The requirements set for the data analysis method is derived from the objectives of this work. The identified requirements are:

- The method should be able to predict typical performance values for machines in different operating conditions.
- The method should enable easy updating of the model as more measurement data is acquired.
- The model structure should be easy to interpret and utilise in the performance analysis and optimisation.
- The method should select the most relevant input variables for modelling, without extensive prior knowledge about the data.

In the present study, we selected a combination of data preprocessing and CART algorithm to analyse the performance values. CART was selected for this study because it meets the described requirements and also provides features such as nonparametric modelling, robust handling of outliers, computational speed, etc. [3]

III. Methods

This section introduces the data preprocessing and CART method utilised in this work. We also adress the regression tree complexity selection.

A. Data preprocessing

Among the most important factors affecting the performance of machine learning algorithms are the quality and the quantity of the data. In order to create accurate and reliable models from the data, the amount of irrelevant, erroneous, and redundant data should be low. The main goal of data preprocessing phase in this work is to provide a high-quality data set for the machine learning algorithm. [13] There is no standard method for data preprocessing; instead, a set of general guidelines and procedures have been proposed. The requirements for data preprocessing are set by the characteristics of the data and the objectives of the data analysis. [13]

Factors that need to be considered while performing data preprocessing include variable selection, detection and removal of outliers, missing value handling, discretization, resampling, data normalisation, and dimension reduction. Application-specific knowledge of the data preprocessing requirements is usually required. This knowledge can be acquired from machine operators, mathematical models, or by examining the characteristics of the data. [13]

The quality of the data used with the machine learning algorithms is important [13]. The original data is divided into two subsets: training data and validation data. The training data is used for creating the model and validation data is used to validate the prediction accuracy of the model. In order to model the system comprehensively, the variables in the training data need to have a sufficient amount of variation and scale. The correct and incorrect selections of training data is presented in Figure 2.

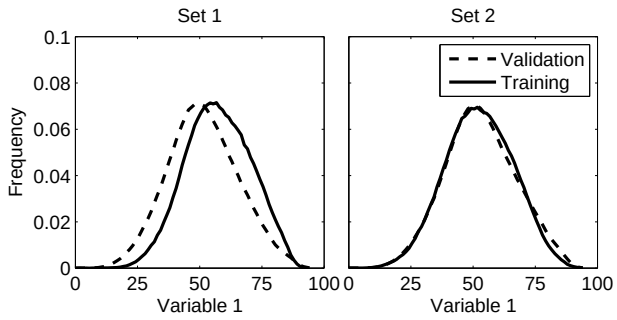


Figure 2. Incorrect (left) and correct (right) selection of training data sets.

The distributions in Figure 2 describe an example of an operating condition measurement in the data. On the right side of Figure 2, variable in the training data covers the whole scale of the validation data. The incorrect selection of training data is presented on the left side of Figure 2. Variable in the

training data does not cover the same scale as the validation data. Therefore, the model, which is generated by a machine learning algorithm, is expected to lack prediction accuracy as some parts of the validation data are not included in the model.

B. Classification and Regression Trees (CART)

CART is a supervised machine learning method that was originally presented by Breiman et al. [3]. CART constructs a model called a decision tree between input and output variables of the data. Two decision tree types are classification trees and regression trees. If the output variable has discrete and predetermined values (that is, classification problem), CART constructs a classification tree. However, if the output variable has continuous values (that is, the regression problem), CART constructs a regression tree. [3] In this work, CART is used to model the relation between the measurement variables describing the operating conditions and the performance of a mobile work machine.

The first stage of utilising regression tree in modelling is the selection of a data set. The data set consists of input and output variables, where inputs are parts of measurement space and outputs are real-valued numbers. The input variables are also known as predictor or independent variables. The outputs are called response or dependent variables. Regression tree creates a real-valued prediction function between the predictor and the response variables. The prediction function can be utilised in two different purposes: to predict the responses based on new predictor measurements, and to understand the relations between the response and predictor variables. [3]

A decision tree is constructed by splitting the data into subsets that are also known as nodes. The building of the decision tree starts from a root node that contains all of the data. A binary split is performed for the root node in a way that the split minimises the fitting error between response values and the predictions of the model in the two child nodes. The splitting variable (that is one of the predictor variables) and its value are the ones that minimise the fitting error. The splitting is then performed recursively for each child node. [3] An example structure of a decision tree is illustrated in Figure 3.

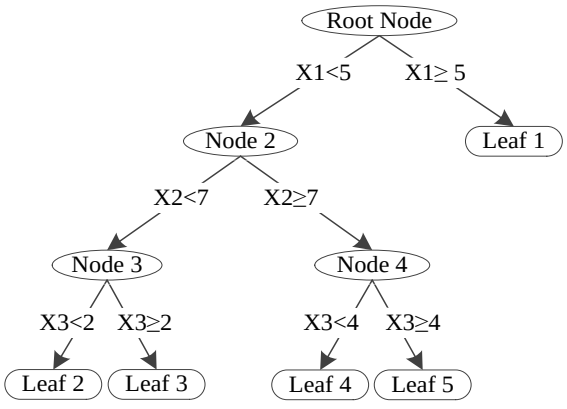


Figure 3. An example of a decision tree structure.

The splitting is continued until the proposed child nodes no longer decrease the fitting error of the regression tree or one of the user-specified stopping criteria is reached. Various stopping criteria can be applied to the splitting, e.g. maximum number of terminal nodes and minimum number of measurement values in each terminal node. The terminal nodes of the regression tree are called leaves. In each leaf, the predicted response value is the average of the response values in the leaf. The following pseudocode describes basic procedure for creating the prediction function with regression tree. [3]

1) SELECT the data used for modelling
2) INSERT the data into the root node of the regression tree
3) SPLIT the data of the node into two child nodes in a way that the fitting error is minimized
4) END IF one of the stopping criteria is met or every node holds only identical response values
5) SELECT the node that has the greatest potential for fitting error reduction
6) CONTINUE from step 3

As presented in the pseudocode, the regression tree continues the splitting of the data until every leaf holds only identical response values or one of the user-defined stopping criteria is met. If the stopping criteria are not used, the result of the modelling is a highly complex and over-fitted regression tree structure. The increased complexity of the regression tree does not necessarily result in improved prediction accuracy with new data. Therefore, while utilising a regression tree in practical applications, a compromise between tree complexity and prediction accuracy is often desired. The required balance between these features is considered to be application-specific. [3]

The selection of tree complexity can be performed with previously described stopping criteria and with pruning methods. Pruning methods can be used to simplify complex regression trees. The basic principle of the regression tree pruning is to decrease the number of leaves in the regression trees. The number of splits in the regression tree is pruned starting from the split that has the least effect on the fitting error. The level of pruning is selected subject to the desired complexity of the regression tree. [3]

The prediction accuracy of the regression tree can be estimated with the following procedure. First, the regression tree is created with a training data set and it is pruned to a desired level. Regression tree enables the estimation of prediction accuracy with resubstitution error and cross-validation error. The prediction accuracy of the regression tree is also validated with a validation data set. The validation data is measured from the same system as the training data, but it is not used in the creation of the regression tree. Comparing the original responce values of the validation data and the predictions of the regression tree, one can estimate the prediction accuracy of the model. [3]

IV. RESULTS

This section is divided into two subsections. The first subsection describes the design of the experiment and the utilised data. The second subsection introduces the results of the experiment.

A. Design of experiment

The purpose of the experiment is to test how the performance of a mobile work machine fleet can be analysed with

 ISBN: 978-1-61208-423-7

TABLE I. THE METRICS OF THE EXPERIMENT DATA

Type of metrics	*Amount*	*Description*
Machines	17	Preclassified machines, 10 training and 7 validation machines
Training data	2,254,901	Approximately 66 per cent of data for training and 34 per cent for validation
Validation data	1,178,485	
Predictor variable	8	Operating condition measurements
Response variable	1	Performance measurement

the regression tree. The scope of the experiment is focused on modelling the relation between the operating conditions and the performance values of the mobile work machines, as presented in Figure 1.

In this work, the regression tree is used to perform three consecutive actions. The following steps demonstrate an example of an approach that could be used in the performance analysis of a mobile work machines.

1) Model the relation between the operating conditions and the performance based on the data of a mobile work machine fleet.
2) Assign typical performance values for each operating condition.
3) Utilise the typical performance values in the performance analysis of an individual mobile work machine.

In order to test the proposed method, a mobile work machine data base was collected. The data for the experiment was acquired from a global mobile work machine manufacturer. A data set including 17 mobile work machines was selected for this work. The machines were selected based on preclassification criteria which were the same machine model and same country of operations. This kind of preclassification was performed to decrease the undesired variations in the data. These variations are caused by the different performance standards and operating conditions between the countries. Table I presents the metrics of the data set used in the experiment.

The data set used in the experiment was generated by combining data from measurement data bases. Additional data preprocessing methods applied to the data set were resampling, missing value handling, and data normalization. The data was then divided into training and validation sets as presented in Section III. Approximately 66 per cent of the data was used as training data and 34 per cent as validation data. Due to the privacy policy of the company that provided the data, all of the variable names are changed and values are normalised in this work.

B. Evaluation of results

The regression tree was applied to the preprocessed data and the model between the different operational conditions and the performance of the mobile work machines was created. Data preprocessing and analysis were performed with MATLAB software. Figure 4 presents the structure of the regression tree after the pruning procedure. The pruning of the tree was performed as presented in Section III. The original tree was constructed of 22,697 nodes, and then pruned to 41 nodes

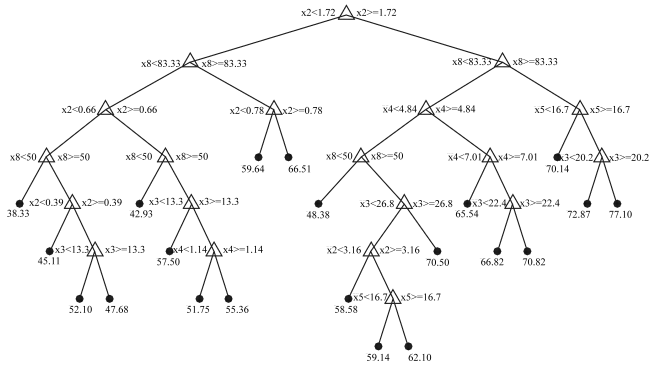


Figure 4. The regression tree constructed with CART.

– this was a compromise between prediction accuracy and complexity.

By looking at the structure of the regression tree, the most significant predictor variables in terms of modelling capability can be found in the upper nodes of the tree. In this work the variables $x2$, $x4$, $x5$, and $x8$ were identified to be the most important predictor variables. Additional predictor variables can be found in the lower nodes of the tree. The predictor variable significance in the regression tree structure was very well in line with the knowledge of the experienced mobile work machine operators. Also, the predictor variables with minor significance on the performance are not used for splitting in the pruned regression tree.

The prediction capability of the regression tree was first analysed by comparing how well the model is fitted to the training data. The resubstitution error of the tree is 19.83 and the 10-fold cross validation error of the tree is 16.29. Figure 5 presents the performance values of the mobile work machines in the training data and the predictions of the regression tree model. As Figure 5 illustrates, the predictions of the model and the original performance values correlate well on machines 1, 4, 6, 7, 8, and 9. The differences between the measurements and the predictions of the other machines are most likely caused by the pruning of the tree and the affects of non-operating condition related factors, such as the tuning of control parameters of the machines. Especially, the machine number 5 outperforms the typical performances of the machines mainly due to the advanced tuning of control parameters.

The model is then used to predict the reference performance values for the machines in the validation data. Figure 6 presents the performance values of the mobile work machines of the validation data and the predictions of the model. Based on the measurements of the operating conditions, the model predicts different performance values for the machines. These predictions are regarded as typical performance values for specific operating conditions. If the performance value of an individual mobile work machine is greater than the prediction, the machine has outperformed same types of machines working in the similar operating condition, and vice versa.

The following information can be observed from Figure 6: The measured performance values of the machines 12, 15, 16, and 17 are mostly similar to the predictions of the regression tree, which indicates average performance in given operating

 ISBN: 978-1-61208-423-7

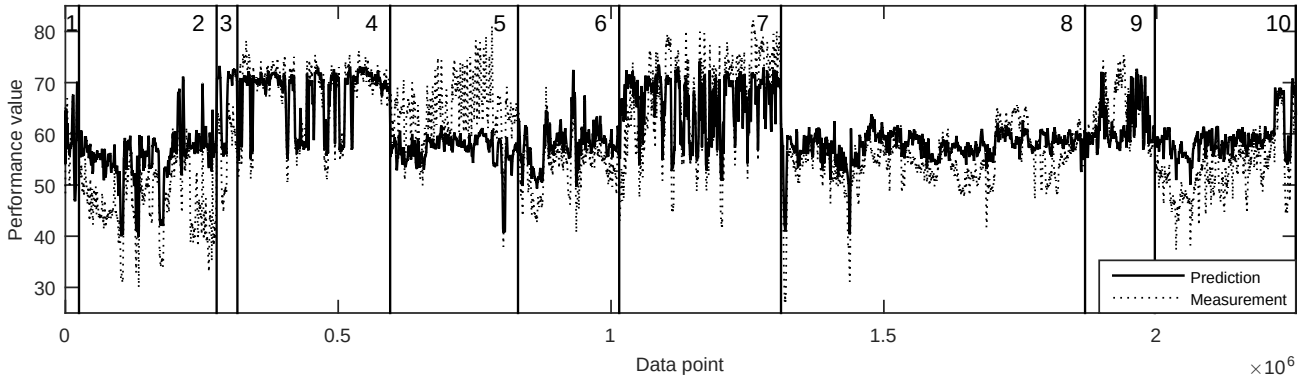


Figure 5. Performance predictions and the measured performance values of the training machines. Data is filtered for visualization.

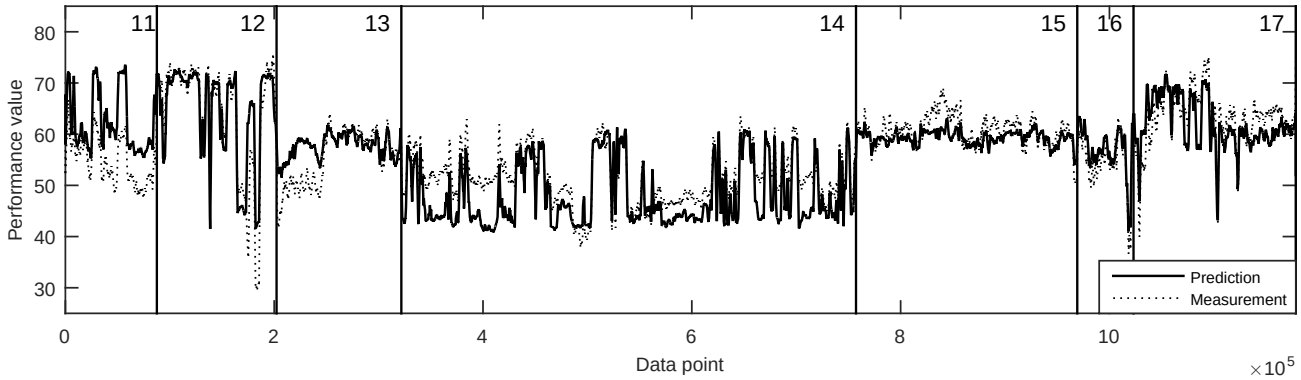


Figure 6. Performance predictions and the measured performance values of the validation machines. Data is filtered for visualization.

conditions. During the measurement periods, there are also better and worse performances compared to the predictions with the previously mentioned machines. These occasional variations can be considered normal, due to various factors, such as unmeasured variations in the operating conditions, temporary decrease in the technical condition of the machine, performance and skill level variations of the operators, etc.

As presented in the Figure 6, the machine number 11 has worse performance compared to the prediction during most of the measurement period. If comparison information about the performance would have been available for the operator, the declined performance could have been spotted and actions taken to improve the performance of the machine. Also the machine number 13 has at first worse performance than the predictions, but then due to actions taken by the operator, the machine reached an average performance in given conditions.

On the other hand, the machine number 14 has on average better performance compared to the references. However, the machine number 14 operates alternately in two different operating conditions. This can be noticed since the value of the performance prediction changes between two main levels. While operating in the environment that typically results in higher performance values, the measurement and the prediction are similar. However, while operating in the other operating condition, the measured performance is higher than the prediction. This is caused by the better control parameter selection.

The utilisation of regression tree enables more detailed analysis of the performance values. Figure 7 presents the performance distributions of the training data and machine number 13, for a specific operating condition. For demonstrative purposes, the selected operating condition is the one where the machine number 13 lacks performance compared to the training data. The distributions illustrate that the performance values of the machine number 13 are concentrated on the leftmost section of the original training data distribution.

Comparison information such as that presented in Figure 7 can be used to assist the machine operators to analyse the performance of the mobile work machine in different operating conditions. With the reference information about similar machines, the operator can analyse the performance of the machine with increased accuracy and reliability. There can be various reasons for the similarities and differences in the performance values between the machines. Depending on the work objectives of the machines, the differences can be explained with logical reasons, e.g. efficiency and productivity priorities set by the machine operators. However, the decreased performance is often a result of declined technical condition of the machine, low skill-level of the operator, or improper control parameter tuning.

 ISBN: 978-1-61208-423-7

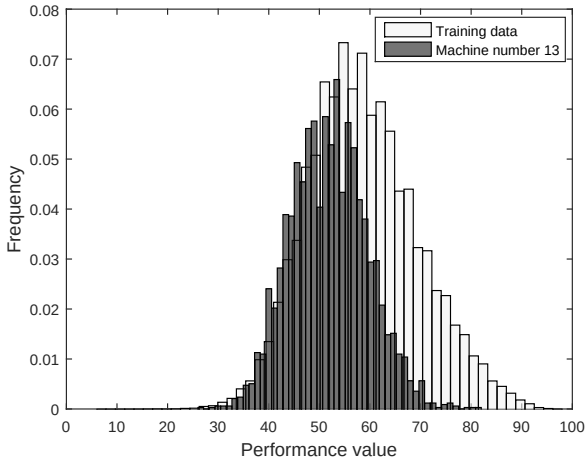


Figure 7. Performance distributions of the training data and the machine number 13, for a specific operating condition.

V. Conclusion and future work

The objective of this work was to research how to improve the performance analysis of mobile work machines. The most significant contribution of this study is the data-driven approach for analysing performances of mobile work machines. The presented approach is a combination of data preprocessing methods and the CART algorithm. The analysis of the performance is executed in three phases: modelling the relation between the operating conditions and performance values, predicting the typical performance values in specific operating conditions, and utilising the performance predictions as a reference values in the performance analysis of an individual mobile work machine.

The proposed method utilises a data-driven modelling approach. All of the operating conditions and performance values in the model are measured from machines working in various operating conditions. As more measurement data is collected, the regression tree can be updated to include new operating conditions and to increase the reliability of the performance predictions. One of the most important benefits of the presented method is the ability to model systems without extensive knowledge of the input-output variable relations in the data.

The results of this study indicate the potential of the presented method in the performance analysis of mobile work machines. However, further research is still required in data preprocessing, modelling, and analysis phases of the topic. Interesting topics related to data preprocessing are dimension and redundancy reduction. Further research topics concerning the modelling phase, include adding new input variables to regression tree modelling and testing new data analysis methods. The analysis phase is the most interesting part, since it enables the development of many practical applications designed for optimisation and root-cause analysis of the mobile work machine. The next step of the research is to increase the volume of the mobile work machine data and to evaluate the performance analysis in practice.

Acknowledgment

The research work was funded by Tekes (D2I – Data to Intelligence) and Academy of Finland (HOPE – Human Operator Modelling And Performance Evaluation In Human-machine Interaction) research programs.

References

[1] V. Hölttä, M. Repo, L. Palmroth, and A. Putkonen, "Index-based performance assessment and condition monitoring of a mobile working machine," in Proceedings of the 2005 ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Long Beach, California, USA: ASME, IEEE, 2006, pp. 615–622, the 2005 ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Long Beach, California, USA, September 24-28, 2005.

[2] V. Hölttä, "Plant performance evaluation in complex industrial applications," Ph.D. dissertation, Helsinki University of Technology, Espoo, Finland, 2009. [Online]. Available: http://lib.tkk.fi/Diss/2009/isbn9789522480927/isbn9789522480927.pdf

[3] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Monterey, CA: Wadsworth and Brooks, 1984.

[4] L. Palmroth, K. Tervo, and A. Putkonen, "Intelligent coaching of mobile working machine operators," in Proceedings of the IEEE 13th International Conference on Intelligent Engineering Systems. Barbados: IEEE, 2009, pp. 149–154, iEEE 13th International Conference on Intelligent Engineering Systems 2009 - INES 2009, Barbados, April 16-18, 2009.

[5] R. H. Macmillan, "The mechanics of tractor-implement performance: theory and worked examples: a textbook for students and engineers," 2002, p. 166.

[6] T. Jokiniemi, H. Rossner, J. Ahokas et al., "Simple and cost effective method for fuel consumption measurements of agricultural machinery," in Agronomy Research, vol. 10, no. Special Issue I. Estonian Research Institute of Agriculture, 2012, pp. 97–107.

[7] S. Park, Y. Kim, D. Im, and C. Kim, "An assessment of eco driving system for agricultural tractor," Journal of Agricultural Science and Technology, 2011, pp. 906–912.

[8] F. Inns, Selection, Testing and Evaluation of Agricultural Machines and Equipment: Theory, ser. FAO agricultural services bulletin. Food and Agriculture Organization of the United Nations, 1995, no. 115.

[9] Z. Ismail and A. Abdel-Mageed, "Workability and machinery performance for wheat harvesting," Misr J. Ag. Eng, vol. 27, no. 1, 2010, pp. 90–103.

[10] J. Lu, Y. Liu, and X. Li, "The decision tree application in agricultural development," in Artificial Intelligence and Computational Intelligence. Springer, 2011, pp. 372–379.

[11] T. Waheed, R. Bonnell, S. O. Prasher, and E. Paulet, "Measuring performance in precision agriculture: CART – A decision tree approach," Agricultural water management, vol. 84, no. 1, 2006, pp. 173–185.

[12] M. Li, S. Feng, I. K. Sethi, J. Luciow, and K. Wagner, "Mining production data with neural network & CART," in Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003, pp. 731–734.

[13] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Data preprocessing for supervised leaning," International Journal of Computer Science, vol. 1, no. 2, 2006, pp. 111–117.

 ISBN: 978-1-61208-423-7

Determing the Business Value of Business Intelligence with Data Mining Methods

Karin Hartl, Olaf Jacob
Department of Information Management
University of Applied Sciences Neu-Ulm (HNU)
Neu-Ulm, Germany
karin.hartl@hs-neu-ulm.de, olaf.jacob@hs-neu-ulm.de

***Abstract*—This paper presents a research project which aims to determine the value of Business Intelligence (BI) and Corporate Performance Management (CPM) with the help of Data Mining methods. The starting point of the research is the hypothesis that the value proposition of BI can be measured on the success of CPM. Previous empirical studies try to define the impact of BI on CPM with research methods like explanatory factor analysis or structural equation modeling. This paper discusses the use and benefit of Data Mining methods for exploring the value of BI. It clarifies why specific Data Mining methods are seen as a beneficial tool to determine the relation between BI and CPM.**

Keywords-Business Intelligence; Corporate Performance Management; Data Mining; Business Value.

I. INTRODUCTION

The challenge companies have to face nowadays for success and existence proves to be increasingly difficult. Globalization intensifies the competition and digitalization leaves enterprises with an immense amount of mainly unstructured data. These data and the contained information, however, are assumed to be the key to ensure the survival of an enterprise in the rapidly changing business environment. BI as a method of analyzing data and the business environment promises companies to support their decision making process [2]. The support is achieved by acquiring, analyzing and disseminating information from data significant to the business activities [1]. Accordingly, BI is seen as a source for quality data and actionable information. This implies that the appropriate use of BI systems supports the success of organizations [3].

As BI projects are not exempt from the increasing pressure in companies to justify the return on IT Investment, the business value of BI needs to be measured [5]. Due to the abstract nature of BI capturing its value, it is a strategic challenge [6][4]. Generally, BI systems don't pay for themselves strictly by cost reduction. Most BI benefits are intangible and hard to measure [5]. Williams and Williams [6] point out that the business value of BI lies in its use within the management processes. Therefore, the concept of CPM evolved, which is understood as the appropriate context to prove the value proposition of BI [4]. It is defined by Gartner as "*an umbrella term that describes all processes, methodologies, metrics and systems needed to measure and manage the performance of an organization*" [9]. CPM presents the strategic deployment of BI solutions and is born out of a company need to proactively manage business performance [8][10]. Inferentially, CPM needs BI to work effectively on accurate, timely and high quality data and BI needs CPM for a purposeful commitment [9]. Consequently, it is expected that the effectiveness of CPM increases with the effectiveness of the BI solution and therefore company success improves as well [11]. A hypothesis can be put in place, which states that the value proposition of BI can be measured on the success of CPM. This research aims to define the link between CPM and BI with the use of Data Mining methods and is based on the findings of Jacob and Lien Mbep [3]. In the research field of BI, exploratory factor analysis and structural equation modeling are the dominant research methods. As addressed in this paper, it is assumed that Data Mining techniques are able to answer different kinds of research questions than the above mentioned approaches on the subjects of BI and CPM. Data Mining could be suitable to gain more detailed information and could be a more appropriate approach to examine the business value of BI on the effectiveness of the CPM processes.

In Section 2 of this paper an overview of subject related research and its importance for this research is given. Section 3 highlights the motivators for using Data Mining techniques to discover the relationship between BI and CPM. In Section 4, the aspired research approach is explained including the Data Mining methods which will be applied. Section 5 closes with a short conclusion on why Data Mining is seen as a beneficial approach to determine the business value of BI.

II. SUBJECT RELATED RESEARCH

In the last couple of years, various studies regarding the business value of BI emerged. An early approach has been made by the Viva Business Intelligence Company [4] in discussing general principles on how the business value of BI could be measured. The article underlines that the direct monetary benefits are hard to calculate and that the significance of BI programs lies in the production and analyzation of information [4]. Even though the standardization of the BI-output for measurement purposes is suggested and possible measures are stated, no detailed

 ISBN: 978-1-61208-423-7

examination has been accomplished. Williams and Williams [6] expose the necessity to determine how BI is used in a company for the quest of defining the value proposition of BI. It is shown that the business value of BI lies especially in its use within the management processes of a company that impact the operational processes. The return on BI investment is assumed to be measurable on the increased revenues and reduced costs. But BI is more than monetary benefits. It is a complex process that makes an in depth evaluation of the impact it has on the management processes necessary. In 2004, Miranda [12] brought BI into context with CPM by summarizing CPM as a business management approach that supports companies in the way they operate by using business analysis. CPM is a management process based on BI systems. Therefore, CPM is identified as a suitable framework for determining the business value of BI. This conclusion provides the foundation for more detailed research in the field, including the following.

Empirical studies on the investigation of the business value of BI have mainly been realized just recently. Yogev et al. [13] addresses the question of the business value gained by implementing a BI system in an enterprise through using a process oriented approach. The research model formulated is built on the resource based view of a firm. It identifies key BI resources and capabilities as possible explanatory factors of the value creation that can be accomplished with the implementation and application of a BI system. Hypothesis are formulated that can be summarized to state that BI has a positive effect on the operational and strategic business processes. Data have been collected using a survey consisting of seven-point Likert scale items, anchored at the ends by "strongly agree" and "strongly disagree". The research method applied is structural equation modelling and the confirmatory factor analysis is identified as showing a satisfactory model fit. The results illustrate that BI has a positive effect on both the operational and the strategic level of the company. Nevertheless, no further details are given on the intensity of the positive effect BI has on performance management, on how this positive effect can be measured nor on the BI related resources which create these positive effects. Richard et al. [11] are the first to investigate the impact of BI on CPM. The aim of the study is the examination of the impact of commonly used BI technologies on the CPM related management practices which include planning, measurement and analysis. The role BI plays in supporting CPM related managements practices is to be identified to enable IT practitioners to better understand the influence of BI technologies across the CPM cycle. The main research hypothesis states that *"the more effective the BI implementation, the more effective the CPM-related management practices"* [11]. The research model as shown in Figure 1 supposes that BI directly influences and supports measurement, planning and analytics. The effectiveness of planning, measurement and analytics, again, influences the effectiveness of the company processes. To answer the research hypothesis, sample data have been collected by using a questionnaire which is based on items in the Likert scale format [11].

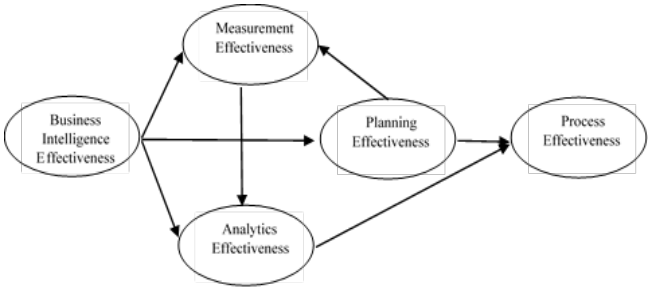


Figure 1. Research Framework [11]

After collecting the data, Richard et al. [11] used an exploratory factor analysis to reduce the number of variables complied in the questionnaire followed by the partial least square (PLS) analysis. The findings suggest that BI positively influences planning effectiveness, analytics effectiveness and through these, indirectly influences process effectiveness as well. Even though the research identifies a direction of how BI influences CPM, the specific BI mechanisms who do so are not defined. Therefore, Richard et al. [11] suggests further studies on this subject, which will be done by this research.

The subject related work is complemented by this study as the previous findings have been used as the initial point. Miranda [12] identifies the importance CPM has in determining the business value of BI and Richard et al. [11] applies this knowledge. But besides discovering and proving a positive connection between BI and CPM, both researches lack detail. It is still not clarified which BI related tools and processes influence the CPM of a company and which CPM processes are effected. Therefore, the CPM process has to be identified in more detail. Based on a process model defined by Lien Mbep et al. [10] and an empirical analysis Jacob and Lien Mbep [3] identified a set of both CPM and BI related items. These items are understood as suitable to measure the appropriate use of BI on one site and CPM on the other. By bringing them together it is aimed to identify the BI related resources and factors which influence the CPM cycle positively. Furthermore, the strength on how the resources and factors positively influence the effectiveness of a company CPM is to be identified.

III. MOTIVES FOR THE DATA MINING APPROACH

The common approach in the research field of determining the value of BI is using exploratory factor analysis first and then confirmatory factor analysis second. With the exploratory factor analysis correlating items are organized together in groups and summed up as a factor. Data can be structured and reduced this way. This structured and reduced data are then analyzed with the PLS method by seeking the optimal predictive linear relationship to assess the previous defined causal relationship. The creation of

 ISBN: 978-1-61208-423-7

factors for compacting information might be the right approach for many research subjects, but is it the only correct approach for defining the business value of BI? It is assumed that Data Mining can highly contribute to the subject. It presents the opportunity to answer different as well as more specific research questions then the commonly used approaches. Instead only testing assumed hypothesis with Data Mining otherwise undiscovered data attributes, trends and patterns can be explored [14]. This can be done with predictive and explanatory Data Mining methods. With explanatory Data Mining data can be interpreted and a better understanding of connections in the dataset is to be achieved [18]. Predictive Data Mining procedures aim to define prediction models from existing data, which allows forecasting unknown values in new data [18]. Although Data Mining is only seen as most suitable for large datasets, Natek and Zwilling [17] disclose that the use of small datasets in specific Data Mining analysis are not limiting the use of the tool. Data Mining can be understood as an extension of statistical data analysis and statistical approaches [16]. Both approaches aim to discover structure in data, but Data Mining methods are generally robust to non-linear data, complex relationships and non-normal distributions [15]. These differences between Data Mining and the commonly used statistical approaches are assumed to supply more detailed and surprising results for the research field of BI and CPM.

IV. RESEARCH APPROACH

This research bases on the findings of Jacob and Lien Mbep [3], where a set of criteria that is seen as most suitable to represent CPM on one and BI on the other side has been identified. In total a set of 20 CPM related items and 28 BI related items have been selected. Now a supplementary study is needed to bring the criteria of both fields together and to explain the relationship between BI and CPM [3]. Therefore, the identified criteria have been transformed into questionnaire items which are to be answered on a five-point Likert scale. The anchor points at the ends of the scale are “does not apply” and “does apply” and an additional definition “applies half and half” for the mid stage has been defined. The data collection is taking place at the moment by using an online questionnaire. Subjects are German companies who use BI for supporting their performance management. Hence decision makers from management, controlling and IT are addressed. With the survey results inter alia the following research questions are to be answered which can be done by using various Data Mining techniques.

R1: Which relations persist between the CPM criteria and the BI criteria?
R2: Which patterns exist between the interviewed companies?
R3: Is the CPM development of a company predictable through the occurrence of the BI characteristics of a company?

To answer the above questions a procedure model has to be first defined. In the Data Mining literature, a broad variety of those can be found, like the Cross Industry Standard Process for Data Mining (CRISP-DM) process model or the overall procedure model Knowledge Discovery in Data Bases (KDD) [18]. They all have the main steps in common. The ones shown in Figure 2 will be followed in the examination of the defined research questions [18]. As the data are especially generated for the research purpose, no selection of the appropriate dataset is necessary. Therefore, the starting point for the data analysis will be preprocessing of the data. Data will be cleansed and missing, as well as conflicting values corrected. The main issue this research assumedly has to deal with are missing values. Cleve and Lämmel [18] suggest alternatives for dealing with missing values depending on the data structure. The important items of the questionnaire are formatted as Likert scale items and can be interpreted as metric data. Metric data can be preprocessed by replacing the missing values in the sample by the mean value of all item-based compiled answers. The mean values also can be stated by contemplating the datasets closest to the dataset with the missing value. This idea follows the k-nearest neighbours (kNN) approach and will most likely be applied to the collected dataset. After preprocessing, the data will be transformed in the required format for the applicable Data Mining technique. Data Mining algorithms demand specific data types. For example, the kNN prefers metric data and the Apriory Algorithm needs binary data [18]. Before operating, the below discussed Data Mining techniques it has to be ensured, that the correct data types for each method are provided. In the third step the data will be mined by applying the algorithm identified as most suitable to answer the above mentioned research questions. Afterwards, in step 4, the outcomes will be interpreted and evaluated.

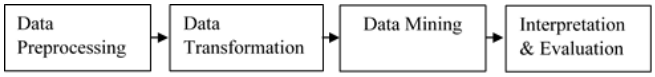


Figure 2. Research Procedure Model

As the study aims to determine the business value of BI on the processes of the CPM of a company, firstly, the relations between CPM and BI are to be explored. The common approach would be to correlate the collected data. This way, the linear connections between BI and CPM can be explored. To get a detailed result on criteria basis that would mean to correlate the 20 CPM items of the questionnaire with the 28 BI items, if no hypothesis are put in place beforehand. Accordingly, researchers would apply an exploratory factor analysis to reduce the data. This may mean more structure and a better overview, but also is associated with data loss and loss of accuracy plus detail.

 ISBN: 978-1-61208-423-7

Data Mining as a technique to discover new and unexpected patterns and relationships in data is assumed to be a second approach for determining connections and associations. In comparison with correlation or regression analysis many Data Mining techniques do not imply connections in advance but discover them automatically. It is assumed that the above mentioned research questions can be answered by the following Data Mining techniques, as shown in Table 1.

To answer research question one R1, Association Rule Discovery will be applied. With association rules, co-occurrence relationships between data items can be discovered, taking into account as many research items as needed and available [18]. This, indeed, can lead on the upside to more detailed results and on the downside to an enormous amount of discovered association rules. Unmanageable amounts of association rules easily can be organized by instating metrics that measure the interestingness of the discovered connections. An appropriate metric could be *Lift* [18]. To generate association rules many algorithms are available. The Apriory Algorithm is the classic procedure and works in two steps [19]. First, frequent itemsets are identified before the confident association rules are generated. In this research, association rule discovery will be applied to find relation rules between BI and CPM. Before applying the Apriory Algorithm to the compiled dataset the data will be transformed into binary variables. This transformation gives each of the items the two characteristics “distinct” or “not distinct”. After executing the Apriory Algorithm to the cleansed data it is assumed that many anticipated connections between BI and CPM are shown. But also interesting new results are expected. As the analysis will include all criteria from both CPM and BI, detailed outcomes are targeted. The results will be association rules, showing which BI items and which CPM items appear together in one of the above mentioned characteristics. For example, a discovered rule could state, that if BI item 1 and BI item 2 are distinct, then there will be a high chance that CPM item 3 as well is distinct. Furthermore, it can be reasoned that the investment of a company in the development of BI item 1 and 2 results most likely in a higher CPM stage of maturity.

The second research question R2 asks for similarities between the companies interviewed. The regarding results could provide information that states if the company size influences the successful use of BI and CPM. Also it could be shown whether companies with well-established CPM strategies also have a well-functioning BI system. Corresponding results may create room for conclusions of a positive connection between the successful use and implementation of the BI system in a company and an effective CPM. Patterns and groups in the research criteria can be found by using clustering. Clustering organizes data without previous knowledge of potential groups [18]. It organizes the examination objects by means of their similarity.

TABLE I. OVERVIEW OF THE APPLICAPLE DATA MINING TECHNIQUES

Research Issues and applicable Data Mining Techniques		
Research Issue	***Data Mining Method***	***Algorithm***
Examining the relationship between the CPM criteria and the BI criteria	Association Rule Discovery	Apriory Algorithm
Pattern discovery	Clustering	k-means algorithm
Predictability of the CPM development	Classification	decision tree modelling

The objects belonging to one group are as much as possible homogenous based on their characteristics [18]. The groups, however, are as heterogeneous as possible among themselves [18]. In clustering all attributes available can be used in parallel. This offers a detailed view of the cluster features and enables a thorough view on the relations between BI and CPM. Clustering is done by defining a similarity and distance measure, which is also known as proximity measure. For ordinal scaled variables the *City-Block-Metric* is an appropriate measure [20]. A first interesting result on the research data is believed to be accomplished by using the k-means algorithm as it is the best known partitioning algorithm [21]. The data are iteratively partitioned into *k* clusters by using a distance function. The quantity of clusters *k* has to be defined beforehand [21]. A hierarchical cluster analysis, like the Ward’s method, can help to define the number of clusters needed [21]. Results could be two or more clusters in which the companies surveyed can be divided. An example outcome could show that companies belonging to the same industrial sector group together in one cluster. The development of the BI items and CPM items in this cluster, furthermore, help to explore the usage and connection of BI and CPM in Germanys companies today. This facilitates the understanding of the connection BI and CPM have in Germany.

To evaluate the predictability of the CPM development in a company on the occurrence of the BI characteristics, as asked in question R3, decision tree modelling [18] can be applied. The decision tree learning is known as very effective and therefore a widely used technique for classification [21]. It is a hierarchical classification model which means that the research items are tested separately according to their importance. In that way, the possible classes are limited gradually and visually presented as a decision tree [19]. This is usually done by identifying successively homogeneous groups in a training dataset in concerning the classification variables [19]. The data collected will be used as training dataset. The decision tree results support the prediction of a company’s CPM maturity through evaluating the BI development. It also shows if a high development of the BI items leads to a high development of the CPM items. In detail, the BI items most important to a successful CPM can be identified. Conversely, it should be possible to derive the BI tools and

 ISBN: 978-1-61208-423-7

processes most important to the effective use of a certain CPM process. This supports, in turn, the definition of the business value of BI.

V. CONSLUSION AND FUTURE WORK

The Data Mining methods association rule discovery, clustering and decision tree modelling are seen as powerful research tools for determining the business value of BI on the effectiveness of CPM. Unlike the commonly used research methods like explanatory factor analysis and PLS, the Data Mining techniques include all research criteria, which is seen to give a more detailed insight and highly interesting new findings on the subject. After collecting the data sample, the above mentioned Data Mining methods will be applied. As indicated in Figure 1, the evaluation and discussion of the results will follow as a next step of this research. It is assumed that the findings will promote the clarification of the relationship between BI and CPM in more detail. In a future research, a time-lagged investigation is sought to evaluate the study results.

REFERENCES

[1] M. Hannula and V. Pirttimäki, “Business Intelligence Empirical Study on the top 50 Finnish Companies”, Journal of American Academy of Business, vol. 2, no. 2, 2003, pp. 593-599.

[2] M. Aho, "The Distinction between Business Intelligence and Corporate Performance Management-A Literature Study Combined with Empirical Findings", in Proceedings of the MCSP 2010 conference, 2010.

[3] O. Jacob and F. H. Lien Mbep, “Factors to determine the value of Business Intelligence to Corporate Performance Management”, Hochschule Neu-Ulm, unpublished.

[4] Pro-How Paper 2/00, “Measuring the benefits of Business Intelligence”, Available from http://legacy.wlu.ca/documents/22449/07_Measuring_the_Benefits_of_BI_Viva.pdf, retrieved 2015.02.12.

[5] S. Negash, "Business Intelligence", The Communications of the Association for Information Systems, vol. 13, no. 1, 2004, pp.177-195.

[6] S. Williams and N. Williams, "The Business Value of Business Intelligence", Business Intelligence Journal, vol. 8, 2003, pp. 30–39.

[7] I. B. Pugna, F. Albescu, and D. Babeanu, "The Role of Business Intelligence in Business Performance Management", Annals of Faculty of Economics, University of Oradea, vol. 4, 2009, pp. 1025–1029.

[8] ResearchandMarkets, “Business Intelligence: Corporate Performance Management”, Available from http://www.researchandmarkets.com/reports/1055897, retrieved 2015.02.19.

[9] J. Becker, D. Maßing, and C. Janiesch, „An Evolutionary Process Modell for the Introduction of Corporate Performance Management Systems“, Data Warehousing, 2006, pp. 247–62.

[10] F. H. Lien Mbep, O. Jacob and L. Fourie, “Critical Success Factors of Corporate Performance Management (CPM)”, BUSTECH, 2015, pp. 6-14..

[11] G. Richard, W. Yeoh, A. Y. Loong Chong, and A. Popovič, “An empirical study of Business Intelligence impact on Corporate Performance Management”, Proceedings of the PACIS 2014 conference, 2014, Paper 341.

[12] S. Miranda, “Beyond BI: Benefiting from Corporate Performance Management Solutions”, Financial Executive, vol. 2, 2004, pp. 58-61.

[13] N. Yogev, L. Fink, and A. Even, “How Business Intelligence Creates Value”, ECIS 2012 Proceedings, 2012, Paper 84.

[14] M. L. Gargano and B. G. Raggad, “Data Mining – a powerful information creating tool”, OCLC Systems and Services, vol. 15, no. 2, 1999, pp. 81-90.

[15] A. Stolzer and C. Halford, “Data Mining Methods Applied to Flight Operations Quality Assurance Data: A Comparison to Standard Statistical Methods”, Journal of Air Transportation, vol. 12, no. 1, 2007, pp. 6-24.

[16] J. Jackson, “Data Mining; A Conceptual Overview”, Communications of the Association for Information Systems, vol. 8, 2002, article 19.

[17] S. Natek and M. Zwilling, “Data Mining for small student dataset”, Management, Knowledge and Leanring Conference, 2013, pp.1379-1389.

[18] J. Cleve and U. Lämmel, Data Mining, Oldenbourg Wissenschaftsverlag GmbH, 2014.

[19] H. Peterson, Data Mining: Methods, Processes, Application Architectures, Odlenbourg Wissenschaftsverlag GmbH, 2005.

[20] K. Backhaus, B. Erichson, W. Plinke, and R. Weiber, Multivariate Analysis – an application-oriented introduction, Springer-Verlag Berlin Heidelberg, Aufl. 13, 2011.

[21] B. Liu, Web data mining exploring hyperlinks, contents, and usage data, Springer-Verlag Berling Heidelberg, 2011.

 ISBN: 978-1-61208-423-7

Data Processing Intervals through Dynamical Models Applied to the Analysis of Self-Degenerative Systems

Ricardo Tomás Ferreyra
Facultad de Ciencias Exactas, Físicas y Naturales
Universidad Nacional de Córdoba, UNC
Córdoba, Argentina
e-mail: ricardotf45@hotmail.com

***Abstract*— In this paper, an oscillatory model is proposed to provide the necessary period of time for the analysis of a system's data in order to reduce its attrition. The actual system was assumed to be a periodic complex dynamical system, since it deals with the human-machine daily routine activity. Typically, in the real life, is the maintenance developed by the employees in the industry. The model presented is suitable for the simulation of the periodicity of the reliability of the studied system, as well as prediction. Then, two types of models are considered: the first one is associated only with the degradation process of the system, while the second one is also associated with the periodic remedying process along the employees production time, from which the lifespan is extended.**

Keywords-dynamic; systems; attrition; measurements.

I. INTRODUCTION

The Poisson's distribution together with the reliability definition gives

$$X(\tau) = 1 - e^{-\alpha\tau} = 1 - C(\tau) \tag{1}$$

which is associated with the degeneration of a studied system. This concept is frequently used in both general and specific literature, as in [1]-[2]. The parameter α is adopted to be the rate of system's faults, t is the arbitrary time of reliability, $C(\tau) = e^{-\alpha\tau}$ is the reliability, while $X(\tau)$ is the attrition as a function of time. However, in addition to these results of the applications, a slaving procedure is implied by (1). This is due to the fact that it is always necessary to process the whole system's data (system or complex system) before emergent real time problems and attrition's values affect the system's operation. Consequently, it was necessary to fix the system's faults to continue its operation. The remedy system starts working again until the new emergent problems appear and affect the operation, see [3]-[5]. This cycle is repeated ad infinitum. Since the attrition is a cumulative process, a new time dependent function $\alpha = \alpha(t)$ replaces the original parameter α. Consequently, this is to be done in (1). In this case, the analyst must reset the observation at every period of time, which demands a lot of attention and effort, even when using a computer. This is due to the fact that the data must be updated at every cycle. The key for this work remains in the assumption that the system has oscillatory motion. This behavior seems a damped mass-spring response, as in [6]. In this context, it is assumed that the system includes humans and machines together. This is also to say that the system stores and loses energy, and also has inertia. Moreover, an appropriate damping implies oscillatory behavior and periodicity. The first objective of this work was to predict the system reliability's evolution through a feedback loop based on the model proposed in [3]-[5], but here the function $\alpha(t)$ is presented in such a fashion that generates a damped mass-spring periodic model. This dynamical model is likely to that developed in [6], but here it is being necessary only the lineal approximation. The second objective was to save administrative time and resources during the system's operation control by applying the generated dynamical damped mass-spring periodic model in obtaining the period of the system. This paper is organized as follows: In Section I, a brief state of the art and some introductory remarks are presented. In Section II, some developments associated with the process of data processing intervals of self-degenerative systems are described. Finally, in Section III, a set of conclusions is provided.

II. DATA PROCESSING INTERVALS FOR A SELF-DEGENERATIVE DYNAMICAL SYSTEM

Let us consider systems with inertial, spring and dissipative forces, as in [6]. Note that fewer forces than these three are qualitatively and quantitatively far from representing the real situation, whatever kind of dynamical system is considered. Here, a balance between the machines production and the human counterpart restoration to operate them is proposed. A typical dynamic equation is

$$\overset{**}{X} + \alpha \overset{*}{X} + W^2 X = 0 \tag{2}$$

, where $(\overset{*}{}) = \frac{d}{d\tau}$ is a derivative with respect to τ. The fonts α and W represent parameters or functions of the system respectively, and X is the unknown variable. Since

 ISBN: 978-1-61208-423-7

the human contribution to the system may be adopted in order to always follow (2), then W must be zero. Although an external and actual agent is needed in this oscillatory process to operate the machines, its contribution does not exist in the right member of (2), effectively. So, (2) is derived homogeneous with $W = 0$. In this context, the most relevant fact to point out is the assumption that X is a global, non-deterministic, and abstract property of the system. For instance, this property could be the degeneration, or the attrition, $X = X(\tau)$, of the system, and obeys (2). This equation is differential, linear, ordinary, second order and, in general, variable coefficients. Moreover, (2), when used, can model the linear asymmetric interaction due to self-degeneration with a sort of "null natural frequency" ($W = 0$). So, we have the equation

$$\overset{**}{X} + \alpha \overset{*}{X} = 0 \tag{3}$$

Then, by adding two initial conditions, such as $X(0) = 0$ together with $\overset{*}{X}(0) = \alpha$, in order to obtain $X(\tau) = 1 - e^{-\alpha\tau}$, agrees with the degeneration of the system obtained from the field of statistical analysis by applying the Poisson distribution as frequently reported in the literature, see [1]-[2]. The parameter α was adopted to be the rate of faults per unit of time, while $X(\tau)$ was the attrition as a function of time.

The following sequence of steps to define a new procedure was developed:

- The system's model suffers a change from stochastic to deterministic after it has been proof that the stochastic model also obeys (3), which is valid for a dynamical system like (2), together with $W = 0$.
- Then, a periodic behavior is introduced by updating the system's dynamical law from (3) to (2) by including energy storage capability $(W \neq 0)$. The negative feedback control-loop (human control or machine control) acts as a spring and generates this behavior.
- Now, by considering $W \neq 0$, and $\alpha = \alpha(\tau)$ in (1), and by using $C = C(\tau)$ in (1) and (2), and also considering explicitly the "forced" human activity as $\frac{\ddot{\gamma}(\tau)}{C(\tau)}$, which is proposed or measured, the period associated with the human-machine is

$$T = \frac{2\pi}{\sqrt{\frac{\ddot{\gamma}(\tau)}{C(\tau)} - \frac{\ddot{C}(\tau)}{C(\tau)} - \frac{2\dot{C}(\tau)\alpha(\tau)}{C(\tau)}}} \tag{4}$$

- Once the time period is generated, the automatic process for resetting the data intervals is implemented:
  - by measuring on site the system's parameters (or functions) α , W , γ and then by scaling the model and processing the macro-data.
  - by solving $\overset{**}{C} + 2\alpha\overset{*}{C} + W^2 C = \ddot{\gamma}(t)$ for the reliability $C(\tau)$, and then for the attrition $X(\tau)$. Then, the response of X over time is obtained and a superior limit for attrition is adopted.

Finally, it is also useful to apply both the Poisson distribution and the reliability definition repetitively at each time period, in order to allow the response to be obtained over time for a statistical tool. Therefore, a comparison should be made to calibrate both responses (dynamical and statistical).

III. CONCLUSIONS

A procedure for setting the data processing intervals for analysis of self-degenerative systems was given as an external support. A new tool is proposed:

- by avoiding the data lecture at each time period.
- by connecting theoretically and experimentally the Poisson distribution with the parameters and response of a dynamical system.

The tool developed was based on physical laws and will be applied to the fields of health care, economy and politics, as well as to other social sciences. This tool advises the operator which parameters should be changed in the actual dynamical system in order to obtain the desired response over time.

ACKNOWLEDGMENT

The author would like to thank the Ministerio de Educación de la República Argentina.

REFERENCES

1. J. M. Epstein, Measuring military power: "The soviet air thereat to Europe", Princeton University Press, 1984.
2. J. M. Epstein, Strategy & Force Planning: "The Case of the Persian Gulf", the Brookings Institution, Washington, D.C., 1987.

 ISBN: 978-1-61208-423-7

3. E. Fogliato, “A predictive model for dimensioning an air force”, Superior School of Air-War, Argentinian Air Force, 2001.
4. E. Fogliato. “A predictive dynamical model for the attrition of air systems in conflict”, ENIEF 2007, XVI**. C**ongresso of Numerical Methods and Applications, Córdoba, Argentina, 2007, pp. 2-5. [Online]. Available from: http://www.famaf.unc.edu.ar/~torres/enief2007/ 2007.10.07
5. E. Fogliato, R T. Ferreyra. Dynamic model of degradation of air systems in conflict. Journal of Mechanics Engineering and Automation vol. 3, pp. 453-457, 2013.
6. R. T. Ferreyra, E. Fogliato, M. A. Ferreyra, S. García, Oscilaciones de relajación en la dinámica no lineal de la depredación entre sistemas**,** Revista de **M**ecánica Computacional vol. XXVII, pp. 2411-2417, 2008. .

 ISBN: 978-1-61208-423-7

Predicting the Next Executions Using High-Frequency Data

Ko Sugiura
Graduate School of Economics
Keio University
Tokyo, Japan
Email: ko.sugiura.0720@gmail.com

Teruo Nakatsuma
Faculty of Economics
Keio University
Tokyo, Japan
Email: nakatuma@econ.keio.ac.jp

Kenichiro McAlinn
Department of Statistical Science
Duke University
Durham, USA
Email: kenmcalinn@gmail.com

***Abstract*—With the progression of computer technology, the term "big data" has become more and more popular in the financial markets. In the literature of finance, this term, in many cases, means high-frequency data, whose size almost reaches as much as 10 GB per day. High-frequency trading (HFT) is, now, widely practiced in the financial markets and has become one of the most important factors in price formulation of financial assets. At the same time, a huge amount of data on high-frequency transactions, so-called tick data, became accessible to both market participants as well as academic researchers, which paved the way for studies on the efficacy of the high-frequency trading and the microstructure of the financial markets. The tick data contain all the information of all trades and are recorded in a thousands of a second, or a millisecond. Nevertheless there have been a great deal of works on investigating the features of HFT, and there have been a few works on application of them in forecast. In this paper, we try to develop a new time series model to capture the characteristics in tick data and use it to predict executions in high-frequency trading.**

Keywords–High-Frequency Trading; Tick Data; Executions; Duration Models; Bid-Ask Clustering.

I. INTRODUCTION

With the progression of computer technology, the term "big data" has become more and more popular in the financial markets. In the literature of finance, that word, in many cases, means high-frequency data, whose size almost reach as much as 10 GB per day. High-frequency trading is, now, widely practiced in the financial markets and has become one of the most important factors in price formulation of financial assets. At the same time, a huge amount of data on high-frequency transactions, so-called tick data, became accessible to both market participants as well as academic researchers, which paved the way for studies on the efficacy of the high-frequency trading and the microstructure of the financial markets. The tick data contain all the information of all trades and are recorded in a thousands of a second, or a millisecond. HFT is used not only in the stock markets but also in the markets for stock options and futures. Increased number of attention has been paid to this data, because it may help the mechanism of price formulation for financial assets. In fact, since the end of twentieth century, many researchers have worked on the practical study using tick data, and a lot of characteristics about high-frequency data have been reported.

One of the most famous series of study in tick data is the study on durations. Naturally, when the next execution occurs or when the price moves is the prime interest for market participants, particularly for specialists. It has long been known that there are largely two difficulties in duration data: discreteness of duration data and the sparsity in duration data. In other words, transaction data arrives with irregularly spaced intervals. However, [2] tackled these problems by proposing a new time-series model. Their model succeeded capturing the feature of clustering of durations. Afterwards, many papers have been devoted to their model and the model has a lot of variations and extensions ([1], [6], [8], etc.).

Another fact which is most frequently documented and stylized on high-frequency transaction data is bid-ask bounce. Bid-ask bounce is a phenomenon that execution prices tend to move back and forth between the best-ask and the best-bid. But, it is also pointed that, particularly in much shorter periods, after an execution at best-ask (best-bid), the next execution occurs more likely at best-ask (best-bid). That is, we can observe the runs of executions, which we named *bid-ask clustering*. The histogram of the runs appears to be more fat-tailed than a fair coin toss suggests. This means that executions don't occur completely at random. Despite a vast amount of literature [5][9] on reproducing the bid-ask clustering, there is little literature on application of this feature into forecast of executions.

In this paper, we try to develop a new time series model with combining the duration models and the feature of bid-ask clustering for forecasting executions in stock markets in the context of tick data. Our contribution is that we take explicitely the bid-ask clustering into consideration and that we focus on the best ask/bid pries themselves, not on the spreads or the price movements. From a practical point of view, we need to specify simultaneously the time and the price for the execution. Since these two pieces of information can fortunately be assumed to be independent, we can identify the probability on these two pieces of information separately. Then, our model comprises two parts and is intuitively understandable.

II. MARKET MICROSTRUCTURE

A. Principles of Financial Market

In general, a market is the platform where people trade something they want. At that place, transactions are made based on the agreement between prospective buyers and prospective sellers. Particularly in the modern financial market, buyers and sellers are matched through electronic servers, and they haggle over the price at a place called the order book. Following certain rules, all actions that take place in the financial market are recorded in order books. As an example of order book, Table I shows a snapshot of the order book

 ISBN: 978-1-61208-423-7

for the stock of Toyota Motor Corporation on 31 April, 2012. In this table, the column labeled "Volume (Ask)" shows how many stocks are on sale and the corresponding price in the middle column is the price at which these stocks will be sold. Such a price is called an ask price. In the same table, the column labeled "Volume (Bid)" shows how many stocks they are willing to buy and the corresponding price in the middle column is the price at which these stocks will be bought. Such a price is called a bid price. The best ask price is the lowest among ask prices while the best bid price is the highest among bid prices. The difference between the best ask price and the best bid price is called the bid-ask spread. Since no one wants to buy stocks at a price above the best ask price or to sell at a price below the best bid price, cells above the best ask price in the left column and those below the best bid price in the right column are empty by construction. Therefore, if they want to sell some of their stocks, they need to look at bid prices. If they want to buy some stocks, on the other hand, they have to consider ask prices.

When it comes to order processing method, two types of method are used; one is a call market and the other is continuous trading. In the former, orders are collected without execution until the certain time, and when the market is called, they start to be simultaneously matched. This style is used in the beginning and the ending of the trading session. In the latter, on the other hand, orders can be executed intermittently while the market is open. This method is mostly used during the trading hours excepting for the opening and closing of the market.

TABLE I. Order Book (31 April, 2012)

Volume (Bid)	Price (Yen)	Volume (Ask)
	⋮	⋮
	2822	23400
	2821	4200
	2820	17200
	2819	10600
	2818	3000
	2817	2100
	2816	2000
	2815	15400
4700	2814	
4400	2813	
5300	2812	
7300	2811	
2100	2810	
8600	2809	
2200	2808	
8300	2807	
⋮	⋮	

Since 1970's, a great deal of attention have been paid to the question how difference in a trading mechanism affects on a price discovery process in financial markets. Studies on this topic caught on especially after 1980's and the field has gained its own name: market microstructure. [7] provides a comprehensive overview of this topic. Although there are a tremendous amount of researches on market microstructure, the characteristics of order executions in a market tend to be translated into three aspects of transactions; prices, volumes and durations. In this section, we review some of the prominent works relating to these variables.

B. Tick Data

Table II shows a typical format of tick data. They are excerpts from by the Nikkei NEEDS database which will be used for our empirical study. As shown in Table II, the data are composed of snapshots of order books. For example, the best ask price is in ③ and seven ask prices are in ④ $\sim$ ⑩ above the best one while the best bid price is in ⑫ and seven bid prices are in ⑬ $\sim$ ⑲ bellow the best one. Additionally, the data also have the information of executions (②). Each line contains a variety of information. A full description of the information is given in Table III.

TABLE II. Tick Data

① 150020120131111 11 7203 0953333002+00000000197 0+0001031200128
② 110020120131111 11 7203 0953 03003+00002814 16 0+0000000400 0
③ 120020120131111 11 7203 0953333004+00002815 0 0+0000015400128
④ 150020120131111 11 7203 0953333004+00002816 1 0+0000002000128
⑤ 150020120131111 11 7203 0953333004+00002817 2 0+0000002100128
⑥ 150020120131111 11 7203 0953333004+00002818 3 0+0000003000128
⑦ 150020120131111 11 7203 0953333004+00002819 4 0+0000010600128
⑧ 150020120131111 11 7203 0953333004+00002820 5 0+0000017200128
⑨ 150020120131111 11 7203 0953333004+00002821 6 0+0000004200128
⑩ 150020120131111 11 7203 0953333004+00002822 7 0+0000023400128
⑪ 150020120131111 11 7203 0953333004+00000000 97 0+0001237300128
⑫ 120020120131111 11 7203 0953333005+00002814128 0+0000004700128
⑬ 150020120131111 11 7203 0953333005+00002813129 0+0000004400128
⑭ 150020120131111 11 7203 0953333005+00002812130 0+0000005300128
⑮ 150020120131111 11 7203 0953333005+00002811131 0+0000007300128
⑯ 150020120131111 11 7203 0953333005+00002810132 0+0000002100128
⑰ 150020120131111 11 7203 0953333005+00002809133 0+0000008600128
⑱ 150020120131111 11 7203 0953333005+00002808134 0+0000002200128
⑲ 150020120131111 11 7203 0953333005+00002807135 0+0000008300128
⑳ 150020120131111 11 7203 0953333005+00000000197 0+0001031200128
㉑ 120020120131111 11 7203 0953333006+00002815 0 0+0000015400128

TABLE III. Definition of Items in Tick Data

$$\underbrace{1200}_{}\underbrace{20120131}_{(I)}\underbrace{11111}_{}\underbrace{7203}_{(II)}\underbrace{0953}_{(III)}\underbrace{33}_{(IV)}\underbrace{30}_{(V)}\underbrace{06}_{(VI)}\underbrace{+00002815}_{(VII)}\underbrace{0}_{(VIII)}\underbrace{0+0000015400}_{(IX)}\underbrace{128}_{}$$

Number	Item Name	Defenition
(I)	Date of Data	YYYYMMDD (Y: Year, M: Month, D: Day)
(II)	Companies' Codes	Four-digit numbers for companies
(III)	Time 1	HHMM (H: Hour, M: Minute)
(IV)	Classification of Records	"0": Executed "1": Not executed
(V)	Time 2	SS (S: Second)
(VI)	Consecutive Numbers	Consecutive Numbers in the same times
(VII)	Prices	Unit: Yen
(VIII)	Classification of Orders	"16": Executed at the best ask price "48": Executed at the best bid price "0": Other cases
(IX)	Volumes	Unit: Stocks

III. Non-Randomness of Executions

A. Bid-Ask Clustering

Despite the fact that the sample size of tick data is large enough to justify the use of the law of large number in the standard situation, it is recognized among researchers that the variance of a tick-data-based estimator such as realized volatility tends to be extremely high and difficult to obtain a stable estimate. Many researchers proposed possible explanations of this phenomena. One promising answer to this question is that high-frequency tick-by-tick price series we observe contain

some kinds of observation error. One of the most influential component of the error is called *bid-ask bounce*, which stems from back and forth movements of prices between bid and ask prices. There are many works treating this phenomena. Among them, [3] analyzes the mechanism of bid-ask bounce from the perspective of bid-ask spread, and gives an intuitively simple explanation about the cause. Here we shall briefly review his work.

Another well-known phenomenon found in tick data is *bid-ask clustering*. This term refers to the stylized fact that an execution at the best ask (bid) price tends to be followed by another execution at the best ask (bid). Figure 1 shows a histogram of the length of runs in executions[1]. As the length of a run increases, the number of runs are observed more than the geometric distribution (fair-coin toss) implies. This tendency of serial correlation has been analyzed in a number of works. Particularly, many have been devoted for elucidating nature of this feature, or reproducing the phenomena using the agent-based simulations. For example, [9] pointed that the investors' order submissions were exactly influenced by the state of the order book, and this fact indeed generated serial correlation in volume, volatility and order signs. Moreover, [5] considered an order splitting strategy of traders, which split their large orders into smaller ones. Although this strategy was originated from minimization of market impacts, they showed that the minimization strategy leads to the serial correlation.

As we have seen here, there have been a tremendous amount of works on market microstructure. However, there exist only a few number of papers which studied price movements in terms of best ask and bid prices, not bid-ask spreads or execution prices. In our proposed model, we explicitly treat whether execution occurs at the best ask or the best bid. We also incorporate bid-ask clustering into our model and try to take advantage of it in forecasting price movements and making investment strategies. In the next chapter, we will further elaborate these points and lay our framework for prediction of the future execution.

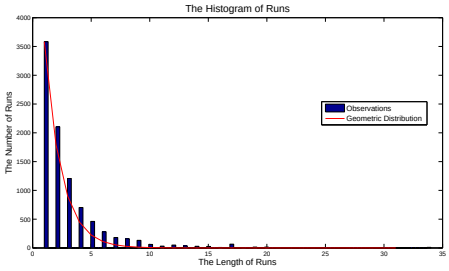


Figure 1. Histogram of Runs (Executions at Best Ask)

IV. Duration Models

A. Autoregressive Conditional Duration (ACD) Model

Although econometricians have traditionally worked on analyzing regularly spaced data, i.e. daily, monthly and yearly data, duration data have some difficulties in modeling. First of all, the data are recorded inherently in irregular time intervals. In order to address this matter, [2] assumed that the arrival times are random variables which follow a point process. The second problem in duration data is that they are necessarily non-negative. Traditionally in the context of finance, the random variables we are interested in may take both negative and positive values. When it comes to duration, however, it is essential to pose a restriction of no-negative on the model. Lastly, it is a well known fact that clusterings can be seen in duration data. This phenomena is thought to stem from a simple causality: the more active a market become, the more transactions we observe. Since the same feature was recognized in volatility and it was modeled by GARCH models, [2] introduced the similar method in duration models.

For the sake of tackling the problems just mentioned above, [2] introduced Autoregressive Conditional Duration (ACD) models. As its name suggests, the ACD models are specified in terms of the conditional density of the durations. Although we recall here its simplest version for simplicity, the discussion can be generalized into higher orders. Letting $\delta_n = t_i - t_{i-1}$ and ψ_i be the interval between two arrival times and the conditional expectation of the i-th duration, respectively, we have:

$$\psi_i = E_{i-1}(\delta_i | x_{i-1}, \theta), \tag{1}$$

where θ is the other parameters. The ACD models consists of this parameterizations and the following assumption:

$$\delta_i = \psi_i \epsilon_i, \tag{2}$$

where $\{\epsilon_i\}$ is a sequence of $i.i.d.$ random variables with positive support. Although the general form of ACD models can be written by the combination of (1) and (2), there are proposed a number of variations on the assumption of $\{\epsilon_i\}$. Engle and Russell, in their paper, introduced the EACD model in which the "E" represented the exponential assumption on the innovation terms. They mentioned the first order one of the EACD models is often the very successful and this is represented as:

$$\begin{aligned} \psi_i &= \omega + \phi\delta_{i-1} + \kappa\psi_{i-1} \\ \delta_i &= \psi_i \epsilon_i, \end{aligned}$$

where $\{\epsilon_i\}$ follows $Exponential(\lambda)$, $\omega > 0$, and $\phi, \kappa \geq 0$.

B. Stochastic Conditional Duration (SCD) Model

About fifteen years after the appearance of ACD models, [1] introduced a state-space class of parametric models for durations, which they called *stochastic conditional duration (SCD)* models. In their models, a latent variable cause the evolution of the duration, and equally it capture the information which cannot be observed directly. Then, SCD models are composed of two stochastic equations, namely state equation and observation equation, whereas ACD models have a stochastic equation and a deterministic equation. In SCD models, the conditional expected duration of ACD model become a random variable. In terms of shapes of models, ACD models and SCD models are similar to GARCH models and SV models, respectively. The simplest version of SCD models is expressed as

$$\begin{aligned} \psi_i &= \omega + \theta\psi_{i-1} + u_i \\ \delta_i &= exp(\psi_i)\epsilon_i, \end{aligned}$$

where $\{u_i\}$ follows a Gaussian distribution and $\{\epsilon_i\}$ a distribution with positive support. The innovation term of the observation equation can take some form, and [1] mentioned the case

[1] In fact, this series of data reject the null hypothesis in run test.

 ISBN: 978-1-61208-423-7

of Weibull distribution and gamma distribution. Although [1] used the combination of quasi-maximum likelihood estimation and Kalman filter in parameter estimation, we employed a more general method called *particle filter*.

C. Parameter Estimation: Particle Filter

When it comes to the parameter estimation of state-space models, there arise two problems: filtering hidden state variables and estimating model parameters. After the development of Kalman filter, these problems have been discussed in Bayesian framework, which is called *particle filter*. Take a general form of non-Gaussian nonlinear state-space model for time series y_t, for example;

$$\begin{aligned} x_t &= f(x_{t-1}, v_t) \\ y_t &= h(x_t, w_t), \end{aligned}$$

where x_t is a hidden state variable, and v_t and w_t are both noise terms. This model implies the information about two types of distribution: the distribution of x_t conditioned to x_{t-1}, $p(x_t|x_{t-1}, \theta)$, and the distribution of y_t conditioned on x_t, $p(y_t|x_t, \theta)$, where θ represents model parameters. Besides, let the distribution of x_0 and the distribution of θ be $p(x_0|\theta)$ and $p(\theta)$, respectively. Thus, the state-space model can be denoted by

$$\begin{aligned} x_t|x_{t-1} &\sim p(x_t|x_{t-1}, \theta) \\ y_t|x_t &\sim p(y_t|x_t, \theta) \\ x_0 &\sim p(x_0|\theta) \\ \theta &\sim p(\theta), \qquad \text{for } t = 1, \cdots, T. \end{aligned}$$

Ordinary particle filter is interested in only hidden state variables given the model parameters, and its procedure consists prediction step and filtering step. Prediction distribution at $t-1$ is given by filtering distribution at $t-1$ and prediction distribution at $t-1$.

$$\begin{aligned} p(x_t|y_{1:t-1}) &= \int p(x_t, x_{t-1}|y_{1:t-1}) dx_{t-1} \\ &= \int p(x_t|x_{t-1}, y_{1:t-1}) p(x_{t-1}|y_{1:t-1}) dx_{t-1} \\ &= \int p(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1}) dx_{t-1} \end{aligned}$$

Filtering distribution at time t is obtained by observation distribution at t and prediction distribution at $t-1$.

$$\begin{aligned} p(x_t|y_{1:t}) &= p(x_t|y_{1:t-1}, y_t) \\ &= \frac{p(x_t, y_t|y_{1:t-1})}{p(y_t|y_{1:t-1})} \\ &= \frac{p(y_t|x_t, y_{1:t-1}) p(x_t|y_{1:t-1})}{p(y_t|y_{1:t-1})} \\ &= \frac{p(y_t|x_t) p(x_t|y_{1:t-1})}{p(y_t|y_{1:t-1})} \\ &= \frac{p(y_t|x_t) p(x_t|y_{1:t-1})}{\int p(y_t, x_t|y_{1:t-1}) dx_t} \\ &= \frac{p(y_t|x_t) p(x_t|y_{1:t-1})}{\int p(y_t|x_t) p(x_t|y_{1:t-1}) dx_t}. \end{aligned}$$

Naturally, this equation is nothing but Bayes' theorem[2]. As is often the case with non-linear and non-Gaussian state-space models, these computations, especially integrations are too complicated for analytical implementation. Then, Markov Chain Monte Carlo (MCMC) method started to be used rapidly to accomplish the integration calculus in wide range of the research area at the end of twentieth century, thanks to a remarkable development in computer technology that helps to simulate a good amount of calculation. Particle filter is also a feat of MCMC method[3], and is always implemented by a numerical way. The merit of particle filter is that it enables us to make a on-line estimation of parameters and predictions. In order to estimate hidden states variables and model parameters jointly, [4] proposes the application of extended state vector for parameter estimation, which he calls it *self-organiged state-space model*. We employed his method in our research.

Algorithm 1 Algorithm for Particle Filter

(1) Give an initial set of particles $\{x_{0|0}^{(i)}\}_{i=1}^m$, where m is the number of particles.
(2) Repeat the following steps for $t = 1, \cdots, T$, where T is the length of data.
a. Generate a random numbers which represent state noise $v_t^{(i)} \sim q(v_t)$, for $i = 1, \cdots, m$.
b. Compute $x_{t|t-1}^{(i)} = f(x_{t-1|t-1}^{(i)}, v_t^{(i)})$, for $i = 1, \cdots, m$.
c. Compute $\lambda_t^{(i)} = p(y_t|x_{t|t-1}^{(i)})$, for $i = 1, \cdots, m$.
d. Compute $\beta_t^{(i)} = \lambda_t^{(i)} / \sum_{i=1}^m \lambda_t^{(i)}$, for $i = 1, \cdots, m$.
e. Resample particles $\{x_{t|t}^{(i)}\}_{i=1}^m$ from $\{x_{t|t-1}^{(i)}\}_{i=1}^m$ with the weight $\beta_t^{(i)}$.

V. Empirical Analysis

A. Model Description

We introduced some notations: n, τ, δ, r, and X. Let t be the time measured in millisecond, and n be the number of execution observed by time t. And τ_n is a random variable for representing the time when the n-th execution is observed, and the duration in our interest is represented by δ_{n+1}, satisfying

$$\delta_{n+1} = \tau_{n+1} - \tau_n. \tag{3}$$

We defined X_n as a random variable representing at which price the next execution occurs. That is, X_n equals to -1 when we observe the n-th execution at best ask price, and to 1 when we observe the n-th execution at best bid price:

$$X_n = \begin{cases} -1 & if \ \ best\ ask \\ 1 & if \ \ best\ bid. \end{cases}$$

Since we highlight the continuity of executions in our research, we define r_n as the length of the last run including X_n, and we count it up as follow;

$$r_n = \begin{cases} 1 & if \ \ X_n \neq X_{n-1} \\ r_{n-1} + 1 & if \ \ X_n = X_{n-1}. \end{cases}$$

From the practical view point, we need to know two pieces of information: when the next execution will occur and at

[2] $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

[3] In fact, some articles call particle filter as Monte Carlo filter.

 ISBN: 978-1-61208-423-7

which price the execution will occur. For this purpose, we set the target probability as bellow:

$$P(X_{n+1} = k, \tau_{n+1} \in (t, t + \Delta t] | X_n, \tau_n, r_n) \qquad k = -1, 1$$

where Δt denotes a time window which will be fixed before simulation[4]. Under the condition of independence on X_{n+1} and τ_{n+1}, the target probability can be decomposed into two parts by the law of conditional probability:

$$\begin{aligned} &P(X_{n+1} = k, \tau_{n+1} \in (t, t + \Delta t] | X_n, \tau_n, r_n) \\ &= P(\tau_{n+1} \in (t, t + \Delta t] | \tau_n, r_n) P(X_{n+1} = k | X_n, r_n), \end{aligned}$$

and they were estimated separately: $P(\tau_{n+1} \in (t, t + \Delta t] | \tau_n, r_n)$ was estimated by duration models and $P(X_{n+1} = k | X_n, r_n)$ was by historical frequency.

For the sake of applying the duration models, we rewrite $P(\tau_{n+1} \in (t, t + \Delta t] | \tau_n, r_n)$ in the context of durations using the equation (3). Substituting it, we obtain a following duration representation:

$$\begin{aligned} &P(\tau_{n+1} \in (t, t + \Delta t] | \tau_n, r_n) \\ &= P(t < \tau_{n+1} \leq t + \Delta t | \tau_n, r_n) \\ &= P(t < \tau_n + \delta_{n+1} \leq t + \Delta t | \tau_n, r_n) \\ &= P(t - \tau_n < \delta_{n+1} \leq t - \tau_n + \Delta t | \tau_n, r_n). \end{aligned}$$

We estimated this by the ACD model and the SCD model. The parameter estimation of both models were conducted through particle filter, because it enables us to update on line the parameter estimates. Although particle filter is usable in continuous time, we, in the process of particle filter, update this probability as we observe a new order or execution. When we observe an execution, we update the probability with recalculating a predictive distribution. On the other hand, when we observe an order, we update the probability without recalculating a predictive distribution. Calibrations of the probability were conducted by moving a time window, Δt, on the predictive distribution. Thus, when we observe an execution, the probability is calculated as

$$\begin{aligned} &P(\tau_{n+1} \in (t, t + \Delta t] | \tau_n, r_n) \\ &= P(0 < \delta_{n+1} \leq \Delta t | \tau_n, r_n) \\ &= \frac{\int_0^{\Delta t} f(\delta_{n+1} | \tau_n) d\delta_{n+1}}{\int_0^{\infty} f(\delta_{n+1|} | \tau_n) d\delta_{n+1}} \\ &= \frac{\int_0^{\Delta t} f(\delta_{n+1} | \tau_n) d\delta_{n+1}}{\int_0^{\Delta t} f(\delta_{n+1|} | \tau_n) d\delta_{n+1} + \int_{\Delta t}^{\infty} f(\delta_{n+1|} | \tau_n) d\delta_{n+1}} \\ &= \int_0^{\Delta t} f(\delta_{n+1} | \tau_n) d\delta_{n+1}, \end{aligned}$$

where $f(\cdot)$ denotes a predictive distribution. Similarly, when we observe an order, the probability is given by

$$\begin{aligned} &P(\tau_{n+1} \in (t, t + \Delta t] | \tau_n, r_n) \\ &= P(t - \tau_n < \delta_{n+1} \leq t - \tau_n + \Delta t | \tau_n, r_n) \\ &= \frac{\int_{t-\tau_n}^{t-\tau_n+\Delta t} f(\delta_{n+1} | \tau_n) d\delta_{n+1}}{\int_{t-\tau_n}^{\infty} f(\delta_{n+1|} | \tau_n) d\delta_{n+1}} \\ &= \frac{\int_{t-\tau_n}^{t-\tau_n+\Delta t} f(\delta_{n+1} | \tau_n) d\delta_{n+1}}{\int_{t-\tau_n}^{t-\tau_n+\Delta t} f(\delta_{n+1|} | \tau_n) d\delta_{n+1} + \int_{t-\tau_n+\Delta t}^{\infty} f(\delta_{n+1|} | \tau_n) d\delta_{n+1}}. \end{aligned}$$

After estimating the duration, we calculate the probability $P(X_{n+1} = k | X_n, r_n)$ using the histogram of length of runs. When we observed $X_n = k$ and a run of executions whose length was $\bar{r}$, the probability we wanted to know was given by

$$\begin{aligned} &P(X_{n+1} = k | X_n = k, r_n = \bar{r}) \\ &= \frac{\sum_{i=n+1}^{\infty} P(X_{i+1} = k | X_i = k, r_i = \bar{r} + i - n)}{P(X_{n+1} \neq k | X_n = k, r_n = \bar{r}) + \sum_{i=n+1}^{\infty} P(X_{i+1} = k | X_i = k, r_i = \bar{r} + i - n)}. \end{aligned}$$

B. Algorithms

In order to compare the performance of our model, we introduced 5 types of algorithms. The difference comes from the estimation method of two probabilities we divided. In *Model 1* and *Model 2*, $P(\tau_{n+1} \in (t, t + \Delta t] | \tau_n, r_n)$ of both models were estimated by the SCD models. But $P(X_{n+1} = k | X_n, r_n)$ of the former model was given by the "bid-ask clustering" or the histogram of length of runs, while that of the latter was by a completely random method, namely a fair coin toss. Similarly in *Model 3* and *Model 4*, the $P(\tau_{n+1} \in (t, t + \Delta t] | \tau_n, r_n)$ in both models were calculated through the ACD models, whereas $P(X_{n+1} = k | X_n, r_n)$ of the former was by the "bid-ask clustering" and that of the latter was by a fair coin toss. Lastly, *Model 5* was comprise of completely and totally random method, that is, both probability $P(\tau_{n+1} \in (t, t + \Delta t] | \tau_n, r_n)$ and $P(X_{n+1} = k | X_n, r_n)$ were given by fair coin tosses. Since it is reported that the SCD model fit better than the ACD model, we expected the Model 1 to show the best performance. Using these algorithms, we made predictions about executions: whether execution occurs in Δt or not, and if does, at which best prices the execution occurs. Then, our prediction was categorized into three types: *no execution*, *execution at best ask price* and *execution at best bid price*. The algorithms of the Model 1 is stated bellow as an example:

Algorithm 2 Model 1

(Step 1) *Execution* **or** *No Execution*
We estimate $P(\tau_{n+1} \in (t, t+\Delta t] | \tau_n, r_n)$ by the SCD model, and we predict

$$\begin{cases} No\ Execution & if\ P(\tau_{n+1} \in (t, t + \Delta t] | \tau_n, r_n) < 0.5 \\ Execution & if\ P(\tau_{n+1} \in (t, t + \Delta t] | \tau_n, r_n) > 0.5 \end{cases}$$

(Step 2) *Best Ask* **or** *Best Bid*
If we predict *Execution*, we predict

$$X_{n+1} = \begin{cases} 1 & if\ P(X_{n+1} | X_n, \tau_n) > 0.5 \\ -1 & if\ P(X_{n+1} | X_n, \tau_n) < 0.5 \end{cases}$$

[4]Note that a trivial fact:

$$\begin{aligned} &P(X_{n+1} = -k, \tau_{n+1} \in (t, t + \Delta t] | X_n, \tau_n, r_n) \\ &= 1 - P(X_{n+1} = k, \tau_{n+1} \in (t, t + \Delta t] | X_n, \tau_n, r_n). \end{aligned}$$

C. Data Description

We applied the proposed model into the real stock data of Toyota Motor Corporation which contained the signs of every order and execution. We used the data of 4th-18th January, 2012 (10 trading days) as a learning period and the data of 19th-31st January, 2012 (9 trading days) as a prediction period. And we omitted the first 30 minutes, because we intended to eliminate the influence of call market method adopted just before opening of the market. Then, the time of the data ranges from 9:30 to 11:30 and from 13:00 to 15:00.

TABLE IV. Statistical Information of the Data Used

	4th-18th Jan	19th-31st Jan
Number of Observations	301517	367320
Best Ask (%)	5.50%	6.54%
No Execution (%)	88.86%	87.74%
Best Bid (%)	5.64%	5.72%

D. Empirical Results

For the sake of summarizing the results, we broke the observations down into the following table, which was used in [10]:

		Actual			
		Best Ask	No Execution	Best Bid	
Predicted	Best Ask	N_{11}	N_{12}	N_{13}	$N_{1.}$
	No Execution	N_{21}	N_{22}	N_{23}	$N_{2.}$
	Best Bid	N_{31}	N_{32}	N_{33}	$N_{3.}$
		$N_{.1}$	$N_{.2}$	$N_{.3}$	N

In the empirical analysis, we made a forecast about executions as we observed an order and/or execution. And Δt after, we examined whether the forecasts were right or wrong. For example, if we forecast there will be a execution at best ask price in Δt at time s, and actually there is a execution at best ask between time s and $s+\Delta t$, we count this prediction adding one to N_{11}. In order to summarize this table, we defined some measures to compare the performance:

- $\alpha = \dfrac{N_{11} + N_{22} + N_{33}}{N}$
- $\beta = \dfrac{N_{11} + N_{33}}{N_{.1} + N_{.3}}$
- $\gamma = \dfrac{N_{11} + N_{33}}{(N_{11} + N_{13}) + (N_{31} + N_{33})}$
- $\delta_1 = \dfrac{N_{11}}{N_{.1}}$, $\delta_2 = \dfrac{N_{22}}{N_{.2}}$, $\delta_3 = \dfrac{N_{33}}{N_{.3}}$

α is the ratio of correct predictions among all the predictions. β is the ratio of correct predictions when we observe executions. γ is the ratio of correct predictions when we predicted executions. δ_1, δ_2 and δ_3 are the ratio of correct predictions when we predicted executions at best ask, when we predicted no executions and when we predicted executions at best bid, respectively.

The simulation results are summarized in the TABLE V, using the measures mentioned above. As for the case with $\Delta t = 1$, Model 5 performed best in β, δ_1 and δ_3. Model 2 was the best model for δ_2. The remaining measures α and γ are takes the highest in Model 1, which shows the second best performance in terms of the other measures. Regarding the case with $\Delta t = 2$, Model 1 outperformed all the other models in all measures.

TABLE V. Performance Measures for the Five Models

	Model 1	Model 2	Model 3	Model 4	Model 5
α	0.5612	0.5174	0.4815	0.4702	0.3675
β	0.2322	0.1458	0.1091	0.0854	0.2500
γ	0.7696	0.4985	0.6314	0.4943	0.5011
δ_1	0.2288	0.1409	0.1087	0.0844	0.2511
δ_2	0.9331	0.9373	0.9026	0.9052	0.5002
δ_3	0.2360	0.1513	0.1096	0.0866	0.2488

	Model 1	Model 2	Model 3	Model 4	Model 5
α	0.5292	0.4557	0.4542	0.4004	0.3403
β	0.4435	0.3301	0.4022	0.3217	0.2505
γ	0.6726	0.4998	0.6308	0.5006	0.5002
δ_1	0.4402	0.3260	0.4006	0.3140	0.2515
δ_2	0.6808	0.6775	0.5460	0.5396	0.4988
δ_3	0.4472	0.3348	0.4041	0.3305	0.2495

$*$ The above table is for the case with $\Delta t = 1$ and the bellow one is for the case with $\Delta t = 2$

VI. Conclusion & Discussion

In our model, we take the feature of bid-ask clustering explicitly into consideration. This arrangement makes it possible to forecast next executions more precisely. Despite the good performance of our model, this doesn't immediately suggest that people can make money from the financial markets, because there is a general rule of price-priority and time-priority in the markets. However, it may bring us an insight about formation of market trends. Moreover, with further studies on the bid-ask clustering, the accuracy of the model can be improved. For example, it might be useful if we take not only the length of runs but also volumes and prices into consideration.

Acknowledgment

This work is supported in part by a Grant-in-Aid for the Leading Graduate School program for "Science for Development of Super Mature Society" from the Ministry of Education, Culture, Sport, Science, and Technology in Japan.

References

[1] Bauwens, L., and Veredas, D. (2004), The Stochastic Conditional Duration Model: A Latent Variable Model for the Analysis of Financial Durations, Journal of Econometrics, **119(2)**, 381-482.

[2] Engle, R.F., and Russell, J.E. (1998a), Autoregressive Conditional Duration: a new model for irregularly spaced transaction data, Econometrica, **66**, 1127- 1162.

[3] Harris, L. (2002) Trading and Exchanges: Market Microstrucuture for Practitioners, *Oxford University Press*

[4] Kitagawa, G. (1998) A Self-Organizing State-Space Model, *Journal of the American Statistical Association*, **93**, 443.

[5] Lillo, F. and Farmer, D. (2004) The Long Memory of the Efficient Market, *Studies in Nonlinear Dynamics & Econometrics*, **8(3)**, 1.

[6] Ng, K. H, Allen, D. E, and Peiris, S. (2009) Fitting Weibull ACD Models to High Frequency Transactions Data: A Semi-parametric Approach based on Estimating Functions, *Working paper of the School of Accounting, Finance and Economics*, Edith Cowan University.

[7] O'Hara, M. (1995) Market Microstructure Theory, Blackwall.

[8] Vuorenmaa, T. A. (2011) A q-Weibull Autoregressive Conditional Duration Model with an Application to NYSE and HSE Data, *SSRN Working Paper Series.*

[9] Yamamoto, R. (2010) Order Aggressiveness, Pre-Trade Transparency, and Long Memory in an Order-Driven Market, *Journal of Economic Dynamics and Control*, **35**, 1938-1963.

[10] Zuccolotto, P. (2004), Forecasting tick-by-tick price movements, *Statistica & Applicazioni*, **II**, 1.

 ISBN: 978-1-61208-423-7

Profit-based Logistic Regression: A Case Study in Credit Card Fraud Detection

Azamat Kibekbaev, Ekrem Duman
Industrial Engineering Department
Özyeğin University
Istanbul, Turkey
E-mail: kibekbaev.azamat@ozu.edu.tr , ekrem.duman@ozyegin.edu.tr

Abstract— **Credit card fraud is a serious and growing problem which became increasingly rampant in recent years. In practice, many predictive models are used to identify fraudulent transactions. In this study, we developed a new profit-based logistic regression model. In order to do this, we modified the cost function in Maximum Likelihood Estimator (MLE) by changing its values according to the profit of each instance. We did this in four different scenarios and tested the results on real-life data of credit card transactions from an international Turkish bank. According to our findings, original Logistic Regression (LR) has the best performance in terms of TP rate. In terms of saving or net profit, profit-based LR scenarios outperformed others.**

Keywords-Fraud detection; Profit-based Logistic regression; MLE; cost function.

I. INTRODUCTION

Logistic Regression (LR) [17] is now widely used in credit scoring and credit card fraud more often than discriminant analysis because of the improvement of the statistical software for logistic regression. Moreover, LR is based on an estimation algorithm that requires less assumptions (assumption of normality, assumption of linearity, assumption of homogeneity of variance) than discriminant analysis. Prior work in related areas has estimated logit models (logit regression or logistic regression) of fraudulent claims in insurance, food stamp programs, and so forth [3][7][10]. It has been argued that identifying fraudulent claims is similar in nature to several other problems in real life including medical and epidemiological problems [13].

In credit card fraud detection, the dependent variable would take on a value of 0 (legitimate transaction) or 1 (fraudulent transaction). In this study, our dependent variable is binary and we estimate a LR model to predict fraud using primary and derived attributes as independent variables. In literature, a commonly used technique to detect credit fraud is LR. Such an econometric tool, together with the above mentioned techniques, is mostly employed within the credit scoring process to help institutions and organizations decide whether to issue credit to consumers who apply for it [1][4][5][6][16].

According to literature, Persons [12] developed a stepwise logistic regression model and provided evidence that accounting data is useful in detecting fraudulent financial reporting. Summer and Sweeney [15] report that a logistic model including insider trading variables differentiates between fraud and non-fraud firms. Lee, Ingram and Howard [9] document that a self-developed LR model has greater predictive ability when including the excess of cash flow over earnings as an explanatory variable, compared to only utilizing traditional financial statement variables. Bell and Carcello [2] construct a LR model based on multiple fraud-risk factors. They find that their relatively simple model consisting of several corporate governance and performance variables successfully differentiates between fraudulent and non-fraudulent observations. On the other hand, Kaminski et al. [8] present evidence that two regression models solely relying on basic financial ratios have limited use in detecting fraudulent financial statements. Sanjeev et al. [14] evaluated support vector machines and random forests, together with the LR, as part of an attempt to better detect credit card fraud. Random forests demonstrated overall better performance across performance measures.

In recent years, among all pattern recognition models, LR has become one of the outstanding linear algorithms with various applications from thrift failures and stock price predictions to bankruptcy prediction. Most of the previous studies have focused on cost of misclassification because in most of the problems, correct classification has no profit and there are just equal or different costs for different types of misclassifications. In above example regarding diagnosis problems, there are different costs for various misclassifications of healthy and unhealthy people. However, in most of the business problems, there is a cost-benefit wise perspective because correct classifications have some kinds of profit. For example, in “credit card fraud” if the base scenario is to take all of the instances as legitimate, if a model correctly detects a fraudulent transaction, it will save the accessible limit of the card and consequently will save it. In the direct marketing context, if a model correctly detects a potential customer for a campaign, there will be a profit of gaining that customer. Due to aforementioned reasons, in most of business problems, we have to develop a profit-cost wise prediction model. In the original version of LR, all of the misclassifications have same costs, which is not a realistic assumption in most of the real-world problems. For instance, in patient diagnosis problems,

 ISBN: 978-1-61208-423-7

misclassification of an unhealthy as healthy is more risky and costly than misclassification of a healthy person as unhealthy. This issue motivated most of researchers to investigate the effect of different misclassification costs on classification models. For this reason, most of the works are related to cost-sensitive LR.

The remainder of the paper is organized as follows: the next section presents a brief literature survey on LR. Section 3 outlines modified error function or profit-based LR which takes the individual net profit into account and four applicable scenarios are presented to generate individual weights. Section 4 introduces the experimental results and discussions. Finally, Section 5 draws the conclusions of the study and indicates some possible future work areas.

II. ORIGINAL AND PROFIT-BASED LOGISTIC REGRESSION

LR is a statistical classification technique that has been developed in 1940's and since then has been widely used in real life. It is similar to a linear regression model but is suited to models where the dependent variable is dichotomous. LR is often used when the dependent variable takes only two values and the independent variables are continuous, categorical, or both. The goal in LR is to find the best fitting, and most parsimonious model, to describe the relationship between a response or outcome variable, and a set of explanatory or predictor variables. LR model predicts the probability of occurrences, so if the odds of occurrences are higher than fifty percent, then the prediction will be assigned to class denoted by binary variable "1", if less it is class "0". The LR model is [18]:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \quad (1)$$

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \quad (2)$$

$$P(y = 1 \mid x; \theta) = h_\theta(x)$$
$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x) \quad (3)$$

where the θ_i's are the parameters and x_i are independent variables. Then, we can reformulate it as:

$$g(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

is called the logistic function or the sigmoid function as shown in Figure 1:

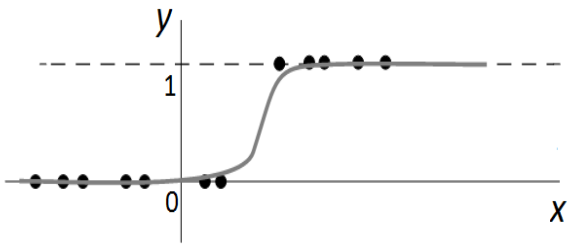


Figure 1. Sigmoid function

Then, we can write it more compactly as:

$$P(y \mid x; \theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y} \quad (5)$$

Assuming that, the *m* training examples were generated independently, likelihood of the parameters will be:

$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X; \theta) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) \\ &= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}} \end{aligned} \quad (6)$$

It will be easier to maximize the log likelihood:

$$\begin{aligned} \ell(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) \end{aligned} \quad (7)$$

After this, we now have to solve the maximization of likelihood. We used Newton's method [19] (also called the Newton-Raphson method) given by:

$$\theta = \theta - H^{-1} \nabla_\theta l(\theta) \quad (8)$$

where, $\nabla\theta\ell(\theta)$ is, as usual, the vector of partial derivatives of $\ell(\theta)$ with respect to the θ_i 's; and H is an n-by-n matrix of second partial derivatives (actually, n +1-by-n + 1, assuming that we include the intercept term) called the Hessian:

$$H_{ij} = \frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \quad (9)$$

Newton's method typically enjoys faster convergence than (batch) gradient descent, and requires much less iteration to get very close to the minimum. The aim of Maximum Likelihood Estimator is to find the parameter values that make the observed data most likely to be predicted.

 ISBN: 978-1-61208-423-7

This paper proposes a new error function which modifies the original cost function to increase the total net profit. In this study, we defined four different scenarios to modify the error function and focused on profitability in the model building step. The key contribution entails that the proposed framework incorporates individual costs and benefits relevant for a business setting, as opposed to the current practice, which focuses on the statistical properties of classification algorithm. It seems obvious that these benefits and losses originating from correct and incorrect classifications should be taken into account. Note that allowing models to optimize the profitability criterion during the model construction step, leads to models with a higher performance in terms of profit although, it may decrease statistical performance of the model in comparison to previous models. Next section will explain our new modified error functions.

III. PROFIT-BASED LR SCENARIOS

Our main goal is to correctly classify the profitable instances as much as possible so that there is less decrease in the accuracy of detecting other instances (i.e. not profitable ones). For this reason, an indicator has been used in the error function to make the algorithm more sensitive to high profitable instances without affecting others. Accordingly, we used a multiplier to intensify the individual penalty of profitable false negatives (in CC Fraud, fraudulent misclassifications which their usable limit is more than average).

We can consider this modification from another point of view. A learning rate is user-defined value to determine how much the weights of examples can be modified at each iteration. We can assume that the learning rate has been modified to assign an appropriate individual penalty for each example and penalize the misclassified important examples considering their individual importance.

The indicator should indicate the profitable (important) instances using their attribute which shows the importance of instance which is Usable Limit (UL) in the context of credit card fraud and the customer revenue (balance) in direct marketing. Thus, indicator has been defined as:

$$P_i = \begin{cases} 1 & if\ UL_i > AvgUL \\ 0 & otherwise \end{cases} \quad (10)$$

$$AvgUL = \frac{1}{n}\sum_{i=1}^{n} UL_i \quad (11)$$

A. Scenario1

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log\left(h_\theta(x_i)\right) + (1-y_i) * \log\left(1-h_\theta(x_i)\right)\right) * \left(\frac{UL_i}{Avg\ UL}\right)^{P_i} \quad (12)$$

where UL_i is the individual profit of instance *i* and $AvgUL$ is average usable limit of an instance. Our main goal is to correctly classify the profitable instances as much as possible with minimum decrease in the accuracy of detecting other instances.

B. Scenario2

As the ratio $\frac{UL_i}{Avg\ UL}$ in the previous scenario can give out large values it may cause instability in the model, so for the sake of making the multiplier not a very large value, we can use logarithm function in an alternative scenario. Hence, the penalty for each instance can be defined as:

$$R_k = \ln\left(1 + \frac{UL_i}{Avg\ UL}\right) \quad (13)$$

The value of one inside the logarithm guarantees that the output will always be positive as the ratio $\frac{UL_i}{Avg\ UL}$ is a positive real number. The penalty function and weight updating equations can be expressed as:

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log\left(h_\theta(x_i)\right) + (1-y_i) * \log\left(1-h_\theta(x_i)\right)\right) * \left(\ln\left(1+\frac{UL_i}{Avg\ UL}\right)\right)^{P_i} \quad (14)$$

C. Scenario3

This scenario is based on modified Fisher [11]. In this scenario, there is no indicator for profitable instances where all of the instances are given a weight related to their potential profit. The error function for this scenario is as follows:

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log\left(h_\theta(x_i)\right) + (1-y_i) * \log\left(1-h_\theta(x_i)\right)\right) * \left(1+\frac{UL_i}{Avg\ UL}\right)^{1/2} \quad (15)$$

D. Scenario 4

This scenario gives different weights for different instances considering their profit of correct classification. Instead of average usable limit we divided it by the maximum of limits. For this reason, this Max_LR error function is:

 ISBN: 978-1-61208-423-7

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log\left(h_\theta(x_i)\right) + (1-y_i) * \log\left(1-h_\theta(x_i)\right)\right) * \left(1 + \frac{UL_i}{\max\{UL_i\}}\right) \quad (16)$$

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The credit card (CC) fraud data set has been gathered from a well-known Turkish bank and it contains 9243 transaction where 8304 of them are legitimate and 939 are fraudulent ones. In the empirical study of each data, the data set has been divided in a way that 2/3 proportion is used to train the model and 1/3 is used to test the trained model. Therefore, there are 313 fraudulent instances and 2817 legitimate ones in the test set. In all the scenarios, the train sets and test sets are the same. However, as the initial weights are generated randomly from standard normal distribution to cope for the effects of randomness related with the solution of train/test sets and the algorithm parameters. Also, each of the models has been run ten times and the average of runs is considered as classifiers' final performance.

In the context of credit card fraud, the most important profit-based attribute is the usable limit of each card. If we correctly detect fraudulent cases, we save their usable limit subject to a cost of contact. Let us consider the base scenario as the case where all transactions are supposed to be legitimate. It is a common approach for evaluating the profit of applying data mining algorithms. Then, the following expression demonstrates how to calculate the amount of net profit (saving) for each model:

$$NP = \sum_{i=1}^{N_{TP}} (UL_i - c) + \sum_{k=1}^{N_{FP}} (-c) \quad (17)$$

where c is the fixed cost for each alarm (cost of contacting the customer) and N_{TP} and N_{FP} indicate the number of true positives and false positives, respectively. As mentioned above, UL_i is the amount of profit gained when the instance i is classified correctly. The threshold has been changed from 0.5 to the number of cases (positives) in test set to show that in the top most probable instances, which of the classifiers is successful.

"*Saving*" measures the amount of profit in each model with threshold 0.5. The **"*Net profit in top n*"** (***n*** is the number of actual positives in test set) evaluates net profit when the cutoff point is output of top n^{th} instance. This measure has an advantage that doesn't care about the number of total positives in each classifier, but it gives more importance to the actual number of positives detected in the first top positives in each model and sums their net profits.

Tables 1-3 illustrate the performances of the four scenarios and original LR on the given data set. According to statistical measure, original LR has the greatest TPR as it tries to correctly classify instances as much as possible where instance's profitability is not important. Also, profit-driven LR in 3rd scenario has also compatible TPR. However, in savings profit-based LR showed better performance (especially 3rd and 4th scenarios). In the average results, Modified Fisher scenario (3rd) has highest amount when threshold is on top 313th instance and Max_LR (4th) outperformed in total savings.

TABLE I. TRUE POSITIVE RATE

Scenario	TP rate		
	Min	Avg	Max
Original	**0,765**	**0,778**	**0,782**
1st	0,764	0,768	0,775
2nd	0,758	0,767	0,778
3rd	0,756	0,772	0,780
4th	0,763	0,769	0,774

TABLE II. TOTAL SAVINGS ON TEST SET

Scenario	Total Saving (%)		
	Min	Avg	Max
Original	0,730	0,762	0,798
1st	0,761	0,775	0,808
2nd	0,766	0,782	0,814
3rd	**0,780**	0,795	0,810
4th	0,770	**0,797**	**0,834**

TABLE III. TOP 10% SAVING ON TEST SET

Scenario	Saving (%) on top 313		
	Min	Avg	Max
Original	0,775	0,793	0,810
1st	0,775	0,800	0,827
2nd	0,787	0,804	0,820
3rd	**0,790**	**0,820**	0,840
4th	0,773	0,815	**0,846**

V. CONCLUSION AND FUTURE WORK

In this study, a novel profit-based logistic regression has been proposed which makes the classification considering all individual costs and profits of instances and

consequently maximizes the total net profit captured from applying the classification model. For this purpose, we modified the logistic regression error function which is sensitive to instances' profitability's. Different scenarios have been proposed to generate weights (penalties) for modification of error function. All scenarios have been tested on a real-life fraud data set. In order to evaluate the classifiers, both TP rate and Savings performance metrics have been used. According to results, original LR has the best performance in terms of TP rate. While, in terms of saving profit-based LR (Modified Fisher and Max_LR) scenarios outperformed others.

As for the future research, we are working on models which assign an individual profit for the non-cases which have been classified correctly. As there is a variable cost of making a contact with each customer, they may get annoyed by this action of being contacted and there might be a cost of missing a customer and consequently missing his/her life time value or future profits.

ACKNOWLEDGMENT

With a deep sense of gratitude the authors would like to thank The Scientific and Technological Research Council of Turkey (TÜBİTAK) under Project No. 113M063.

REFERENCES

[1] H.A. Adbu, "An evaluation of alternative scoring models in private banking," Journal of Risk Finance, vol. 10 (1), 2009, pp. 38-53.

[2] T.B. Bell and J.V. Carcello, "Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting," Auditing: A Journal of Practice & Theory, vol. 19, 2000, pp. 169-184.

[3] C.R. Bollinger and M.H. David, "Modeling discrete choice with response error: food stamp participation," Journal of the American Statistical Association, vol. 92, 1997, pp. 827–835.

[4] J. Crook, and J. Banasik, "Does reject inference really improve the performance of application scoring models?" Journal of Banking & Finance, vol. 28 (4), 2004, pp. 857-874.

[5] V.C. Desai, J.N. Crook and J.G.A. Overstreet, "A comparison of neural networks and linear scoring models in the credit union environment," European Journal of Operational Research, vol. 95 (1), 1996, pp. 24-37.

[6] W.H. Greene, "Sample selection in credit-scoring models," Japan and the World Economy, vol. 10, 1998, pp. 299-316.

[7] J.A. Hausman, J. Abrevaya and F.M. Scott-Morton, "Misclassification of a dependent variable in a discrete-response setting," Journal of Econometrics, vol. 87, 1998, pp.239–269.

[8] K.A. Kaminski, T.S. Wetzel and L. Guan, "Can financial ratios detect fraudulent financial reporting?" Managerial Auditing Journal,vol. 19, 2004, pp. 15-28.

[9] T.A. Lee, R.W. Ingram and T.P. Howard, "The Difference between Earnings and Operating Cash Flow as an Indicator of Financial Reporting Fraud." Contemporary Accounting Research, vol. 16, 1999, pp. 749-786.

[10] M. Artis, M. Ayuso and M. Guillen, "Detection of automobile insurance fraud with discrete choice models and misclassified claims," The Journal of Risk and Insurance, vol. 69 (3), 2002, pp. 325–340.

[11] N. Mahmoudi and E. Duman, "Detecting credit card fraud by Modified Fisher Discriminant Analysis," Expert Syst. Appl., Nov. 2014.

[12] O.S. Persons, "Using financial statement data to identify factors associated with fraudulent financial reporting," Journal of Applied Business Research, vol. 11, 1995, pp. 38-46.

[13] S.B. Caudill, M. Ayuso and M. Guillen, "Fraud detection using a multinomial logit model with missing information," The Journal of Risk and Insurance, vol. 72 (4), 2005, pp. 539–550.

[14] J. Sanjeev, M. Guillen and J.C. Westland, "Employing transaction aggregation strategy to detect credit card fraud," Expert Systems with Applications, vol. 39, 2012, pp. 12650–12657.

[15] S.L. Summers and J.T. Sweeney, "Fraudulently misstated financial statements and insider trading: An empirical analysis," The Accounting Review,vol. 73, 1998, pp. 131-146.

[16] L.C.A. Thomas, "Survey of credit and behavioural scoring: forecasting financial risk of lending to consumers," International Journal of Forecasting, vol. 16 (2), 2000, pp. 149-172.

[17] C. Spathis, "Detecting False Financial Statement Using Published Data: Some Evidence from Greece," Managerial Auditing Journal, vol 17, April 2002, pp.179-191.

[18] D.W. Hosmer and S. Lemeshow, " Applied Logistic Regression (2nd ed.)," Wiley, 2000.

[19] P. Komarek and A. W. Moore, "Making logistic regression a core data mining tool: A practical investigation of accuracy, speed, and simplicity," Robotics Institute, Carnegie Mellon University, 2005.

 ISBN: 978-1-61208-423-7

Big & Deep Data Analytics using Statistical Significance: An Introductory Survey

Sourav Dutta

Databases and Information Systems
Max-Planck Institute for Informatics
Saarbrücken, Germany
Email: sdutta@mpi-inf.mpg.de

***Abstract*—The explosion of diverse and rich information sources across the world wide web has fostered the need of extremely efficient approaches for storage, management, and retrieval of such enormous data in the order of hundreds of petabytes. Scalable data mining or extraction of interesting summaries, patterns, and association rules from such huge text and sequence data-stores caters to a multitude of applications, such as search engines, financial modeling, climate monitoring, computational biology, text analysis, and social graph mining to name a few. This necessity has led to the growth of recent research directions in *big data analytics* and *deep learning*.**

Statistical significance attributes the occurrence of an event to chance alone or to the presence of an interesting phenomenon. Such techniques enable the detection of anomalies or deviations from the expected distribution, enabling faster and highly accurate approximate data mining or retrieval by quantization into "normal" or "significant" observational sub-classes. This paper provides a brief survey of interesting recent works and possible future exploratory directions incorporating statistical significance for sub-text mining (in blog analysis, spell checks, etc.), outlier detection, and graph mining in the context of big data analytics.

Keywords–statistical big data analytics; χ^2 significance; text and graph mining; clustering; survey.

I. INTRODUCTION

The surge of information sources and the explosion of data generated world-wide, catering to diverse applications, such as online transactions, financial data, climate systems, computational biology, natural language processing, and social network graphs among others, has necessitated efficient information management and retrieval. This wealth of data provides a rich opportunity to explore, study, and extract interesting user behavioral rules and interactions, natural patterns, or latent semantic models that might provide crucial operational or framework insights not only for the industry (e.g., user-interface design for new online applications, user recommendations, click/shopping behavior, and rule mining), but also for the academia (e.g., pattern analysis, behavior modeling, and algorithm design), and government (e.g., prevention of natural calamities, security, and telecommunications). The study of efficient and scalable analysis and mining of latent structures from such enormous amounts of data has led to the advent of modern research domains, like *Big Data Analytics* [1] and *Deep Learning* [2][3].

Data analytics involve a range of operations such as prediction (user rating of items [4]), extraction of latent patterns (association rules and market basket for online shops [5]), clustering (recommender systems), outlier or anomaly detection (intrusion detection [6]), etc., based on identification of relationships among objects and data observations. Mathematically, statistical significance forms the framework for establishing whether the outcome of an experiment can be ascribed to some latent phenomena affecting the system or to pure chance alone. As such, this enables the quantification of an observation as "interesting" wherein large deviations from the expected cannot be attributed to randomness alone. Detection of such statistically relevant patterns (potentially hidden) using measures such as the *p-value*, *z-score* [7], etc., within a sequence of events indicating the possible existence of hidden parameters and attributes, caters to large modern data mining applications across diverse fields of study.

In this survey paper, we introduce and discuss several state-of-the-art algorithmic approaches, in applications such as *sub-string mining* (text analytics), *motif extraction* (gene mutations in bio-informatics), approximate string matching (spell checks), subgraph mining (social network graph analytics), etc., that involve novel and efficient use of statistical significance in observations for large data analytic purposes.

Roadmap: Section II introduces a background on the measures and computation of statistical significance. Section III presents different approaches to extract statistically significant sub-sequences from an input sequence. Application of such algorithms for text and graph mining in the context of Big Data is next described in Sections IV and V. We also propose several interesting directions of future research in Section VI, while Section VII concludes the paper, followed by an extensive reference of existing literature in this domain.

II. STATISTICAL MEASURES

Statistical methods capture the degree of uniqueness of a pattern and help classify it as "significant" (or not), i.e., depicting a large deviation from the expected analysis, and also inherently take into account the *Bonferroni's Principle* [8], which informally states that the real instances of an event should be considered bogus if the number of such instances are smaller than the expected number of occurrences under a uniform distribution model. We next discuss a few popular statistical analysis measures:

 ISBN: 978-1-61208-423-7

- **p-value:** Given a sample observation O with score $S(O)$, the classical *p-value* of the observation O characterizes the probability that a random sample drawn from the same probability distribution obtains either the same or a greater score [9], i.e. in effect similar to the tail bound analysis. Formally, the underlying *null hypothesis* (H_0) states that the random sample is indeed drawn from identical probability model, while the p-value measure the chance of rejecting H_0 (based on a pre-defined significance level α). Hence, lower the p-value, less likely is it for H_0 being true and hence the observation tends to be significant. The p-value is mathematically represented by the cumulative probability distribution function (cdf) of O as:

$$p - value(O) = 1 - cdf(O) \tag{1}$$

  However, in most scenarios the probability distribution function is hard to estimate or is non-parametric, leading to the enumeration of exponential number of all possible outcomes (along with the associated scores) for accurate p-value computation, making the computation of p-value practically infeasible. To alleviate such problems, *branch-and-bound* techniques have been proposed [10] or other statistical methods are used for asymptotically approximating the p-value in large samples [11].

- **z-score:** The *z-score* or *standard score* [7][9] measures the number of standard deviations by which an observation differs from the mean or expected value under a normal distribution. It is suitable for outlier detection in applications where the data about the entire population (of all possible observations) is known apriori. Otherwise, it is referred to as the *Student's t-measure* when sample based parameters are considered. Mathematically, for an observation O,

$$Z(O) = \frac{O - \mu_O}{\sigma_O} \tag{2}$$

  where μ_O and σ_O are the mean and standard deviation of the population, respectively. The z-score operates only on the mean and variance of the data, ignoring the probability distribution curve at other points [11], thus rendering it less precise than the p-value.

- **Hotelling's T^2 measure:** The T^2 measure provides a generalization of the Student's t-measure by considering a multivariate distribution of the possible outcomes [12]. It considers the difference in the mean of different outcome populations as,

$$T^2 = n(\mathbf{x} - \mu)^T C^{-1}(\mathbf{x} - \mu) \tag{3}$$

  where n is the number of observations, $\mathbf{x}$ is a column vector of observations with corresponding mean μ, and C is the covariance matrix.

- **Log-Likelihood ratio:** The *likelihood ratio* between two models expresses how likely the data fits under one model than the other. The logarithm of this ratio, or the *log-likelihood ratio* (G^2) [9][13] essentially quantifies the deviation of the observed outcome from the expected behavior by using the theoretical distribution with k possible outcomes as,

$$G^2(O) = 2\sum_{i=1}^{k}\left(O_i \ln \frac{O_i}{E_i}\right) \tag{4}$$

  where O_i and E_i represent the observed and expected number of outcomes for the i^{th} possibility, respectively. G^2 is characterized by its *degrees of freedom*; however suffers from logarithmic instability for low (approaching 0) expected or observed values. The log-likelihood ratio can also be approximated using the *Wilk's theorem* [14], which states that as the sample size tends to infinite, the G^2 statistic becomes asymptotically χ^2 distributed (described next). Further, under no parameter assumption, the likelihood ratio demonstrates the best performance as justified by the *Neyman-Pearson lemma* [15].

- **Chi-square (χ^2) measure:** The χ^2 distribution is generally used to model the goodness-of-fit of a set of observations to the null hypothesis model. Although, for small sample sizes, the distribution tends to degenerate to a normal distribution, it provides a good approximation of the p-value in most scenarios [13]. The *Pearson's χ^2 measure* [16] uses the frequency of occurrences of categorical data to fit the observation model to that proposed by theory. The events are assumed to be independent and mutually exclusive. Similar to G^2, the *Pearson's χ^2* is defined as,

$$\chi^2 = \sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i} \tag{5}$$

  The chi-square distribution is also characterized by degrees of freedom and is anti-monotonic with the p-value, i.e., larger the deviation from the expected, lower is the p-value; hence greater is the χ^2 value and more significant is the observation. Even for multinomial models, the χ^2 measure approximates the statistical importance more closely than the G^2 measure. Hence, we observe that the *chi-square statistic* provides a good approximation to the p-value by diminishing the probability of type-I errors (false positives) and is widely used to estimate the significance of an event (categorical set).

In the remainder of the paper, we present recently proposed algorithms, for data mining on enormous data stores, utilizing χ^2 and other statistical significance measures to categorize and extract interesting patterns or observations.

III. Mining Statistically Significant Strings

Consider, an automated temperature monitoring system (e.g., in industrial combustion chambers) composed of interconnected sensors or a computer server sniffing the network for possible intrusion attacks involving real-time decision based on a sequence of observed events. Such scenarios require the detection of certain "important" events pre-defined as trigger points (e.g., temperature increasing beyond threshold, etc.). However, for many data analytic settings, such as financial modeling, stock prediction, gene mutation characteristics, etc., apriori categorization of events as normal or otherwise is not

 ISBN: 978-1-61208-423-7

possible. Hence, significant pattern detection (using statistical significance) over a sequence of observables, like telecommunication traffic [6][17], time series transactions [18], and others [19][20][21] have been studied.

In this context, we now study the problem of extracting the statistically most significant sub-sequence from an event stream; and provide insights into the working of novel algorithms and theoretical bounds present in the literature. We later show that several other data analytic settings are systematically mapped to solving this central problem, stated formally as:

Problem Statement: Given a sequence of length l composed of event symbols s_i taken from a finite alphabet set Σ with cardinality m, let p_{σ_i} denote the associated probability of occurrence of σ_i such that $\sum p_{\sigma_i} = 1$. For θ_{σ_i} observed number of occurrence of event σ_i, we need to efficiently compute the sub-sequence demonstrating the maximum *chi-squared* value or maximum deviation from the normal.

For the remaining discussion in this section, we use the following example: Consider the input event sequence $I = \{1, 0, 0, 0, 1, 1\}$ of length $l = 6$ and alphabet set $\Sigma = \{0, 1\}$ of size $m = 2$. Assume the probability distribution of the events to be $p_0 = 0.9$ and $p_1 = 0.1$.

1) **Naïve Algorithm:** The simplest approach to identify interesting patterns involve the brute-force extraction of all the sub-sequences present in the input, trivially compute the individual χ^2 score, and finally return the *top-k* events (using a heap data structure) demonstrating the maximum chi-squared value. However, it suffers from $O(l^2 m)$ (quadratic) computation cost for sequence length l and event cardinality m, making it infeasible for real-time large data analytics.
   In our example, the sub-sequences and their corresponding χ^2 values (using Eq. (5)) are 1 (9), 0 (0.11), 10 (3.5), 11 (18), and so on and so forth.
2) **Blocking Algorithm:** To reduce the practical running of the naïve algorithm, [22] proposed partitioning the input symbols sequence into *blocks* consisting of adjacent identical symbols, and each block being replaced by only one of its symbols. Hence, our example input is modified to $I' = \{1, 0, 1\}$.
   The naïve algorithm was then executed on this "block"-ed input to obtain the top-k significant subsequence. Although the theoretical complexity remained the same, significant gains in run-time were shown. Interestingly, for binary symbol settings it was proven that the most significant sub-sequence starts and ends with same event symbol.
3) **Local Maxima Approach:** A combination of blocking and the use of local maxima (based on chi-square scores) was presented in [23][24]. The input events were read serially and the positions of local maxima (based on the chi-square measure) were stored as *begin* candidates. The local maxima found in our running example are $\{1\}$ and $\{0, 0, 0, 1, 1\}$. Similarly, the input sequence is reverse and the local maxima is re-computed for *end* candidates extraction ($\{1, 0, 0, 0\}$ and $\{1, 1\}$).
   The Cartesian product of the positions of the $begin$ and end candidates is then considered for finding the maximum significant locality, and corresponding χ^2 value computed. The candidate positions were conjectured to be necessary and sufficient to find the global maxima, and the subsequent pruning of the search space reduced the run-time of the procedure. Alternatively, a linear time probabilistic approach was also proposed based on the positions by treating the 2 categories of candidates separately. This $O(lm)$ algorithm was shown to attain 90% accuracy under different empirical settings; making it applicable for big data scenarios with slight error-tolerance demanding high run-time efficiency, such as transaction management, spurious web clicks, etc.
4) **Bernoulli Modeling:** Although the above approaches provided significant run-time efficiency, they suffered from a worst case complexity similar to that of the naïve algorithm. The theoretical complexity of a deterministic approach for significant sub-sequence mining was recently reduced to $O(l^{3/2} m)$ [25] with high probability. The proposed algorithm considered the events to be generated from a *memoryless Bernoulli* model and explored possible candidates with all possible lengths of sub-sequences for maximizing the χ^2 value for a chain of events.
5) **Motif Discovery:** The extraction of significant substructures and their interactions have also been exhaustively studied in the domain of bio-informatics for DNA sequencing, protein structure interactions, gene mutations, etc., [26][27] and is referred to as the *motif* discovery problem. Several online tools, such as $WebMOTIFS$ [28], $CompleteMotifs$ [29], etc., have also been designed to offer complete frameworks for motif discovery, scoring, analysis, and visualization. However, the underlying methodology in such methods remains the same, i.e., extracting and exploring the cause of statistically significant observations and structures. However, for modeling statistical significance of events in higher dimensional (matrices, tensors, etc.) settings prevalent in biological domains, generally the *log-likelihood* measure under a Poisson distribution is used [30].

Interestingly, the probabilities of occurrences of the different events can be considered as a combination of different distribution functions, and hence the above algorithms provide a generic framework for diverse model working scenarios. An exhaustive performance comparison of the proposed methods with real as well as synthetic datasets can be found in [25]. It was shown that the Bernoulli modeling and the Local Maxima approaches deterministically obtained the sub-string with the maximum χ^2 value with at least $3\times$ improvement in run-time. Although, the probabilistic algorithm (using Local Maxima) ran faster than the other, the accuracy was observed to vary from $80 - 90\%$ under various data inputs.

IV. Approximate Text Matching

Natural language processing (NLP) and text mining applications extract patterns of words and sentiment usage from blogs, twitter posts, articles, etc., to obtain behavioral rules of users for varied mining tasks, such as recommending product advertisements, studying the veracity of information, prevailing public sentiments, or security measures. Further, several applications involving auto-suggest, text correction, spell checks, and web search require robust approximate text matching [31]

to report documents or resources similar to the user query. Traditional methods employ *Levenshtein distance* [32] and other similarity metrics (e.g., Jaro-Winkler, cosine, etc.) to obtain the closest match, but suffer from high computation complexity – quadratic in the query length for pair-wise similarity computation – for a large dictionary of vocabulary. Several approaches for reduce the complexity involving indexing schemes [33], variable length n-grams [34], along with dynamic programming based filtering techniques [35] was proposed to partially solve the scalability challenge.

To alleviate the above problems, we now discuss a recent *approximate text matching* algorithm using statistically significant sub-sequence mining (of Section III) based on *n-grams* with 1-sliding window protocol [31]. The algorithm proposed a unique mapping of tri-grams present in the document texts onto symbols based on the degree of its matching with triplets present in the query. The similarity between two 3-grams was pre-defined into 4 hierarchical classes, and the probabilities of occurrences of the symbols correspondingly computed (assuming an independent and uniform distribution on the alphabet set). The intuition was to transform the document into a symbol sequence (based on triplet similarity) and thus closely matching words or phrases would lead to multiple adjacent trigram matches represented by high similarity symbols (having low probability) in the documents leading to a high χ^2 value. The probabilistic linear-time local maxima based sub-sequence mining approach (described in Section III) was used on the modified documents to extract the approximately matched texts with efficient run-time complexity.

For example, consider a document $D = abcdef$ and a query $Q = bcde$ with alphabet set $\Sigma = \{a, b, c, d, e, f\}$; where the triplets in D (namely, abc, bcd, cde, and def) are matched with those in Q (namely, bcd and cde). For simplicity, assume an exact match of a 3-gram in D with a triplet in Q to be represented by symbol 1, or by 0 otherwise. Hence, depicting D by similarity symbols, we obtain $D' = 0110$. Observe that the probability of exact triplet match (symbol 1) is very small, and hence the sub-sequence 11 of D' (representing $bcde$ in D) providing the highest χ^2 value is extracted as the most statistically significant string (i.e., best approximate matching to the query Q).

The proposed algorithm [31] is linear in run-time (efficiently bypassing the expensive edit distance computations) and hence provides real-time characteristics applicable to the scenario of big data. Further, it was shown to be $7\times$ faster compared to the naïve algorithm while attaining similar accuracy in results.

V. Sub-Graph Mining

The popularity of social networking communities provide large graphical network structures containing hidden or latent patterns for user-user interaction, influence, and behavior. Efficient mining of association rules from such huge network structures caters to enormous research interest in the multimedia and advertisement domains for collaborative based applications, product recommendations, etc. Similarly, analytics based on *belief propagation* [36], effect of influence, recommendations, and community detection on hugely connected graphs involve efficient and scalable sub-graph mining procedures. Analysis of computer network structures to identify security weak-points and other connectivity problems, along with road networks, etc., also involve mining of network graphs, albeit at varying operational scales. The use of graph mining is also pertinent in computational biology for detecting hidden structural patterns in protein-protein interactions and their associated effects.

Unfortunately, no polynomial time solution exists for the *graph isomorphism problem* and thus the similarity between two graph (with vertex and edge labels) for huge structures is computationally infeasible. Hence, traditional sub-graph mining involve a threshold based frequent pattern search with intelligent indexing schemes, and correspondingly approximate similarity computation to an input query [37]. Extraction and indexing of individual sub-structures of graphs such as k-length cliques for aggregated query reporting (via merging) was proposed in [38], providing a divide-and-conquer strategy using smaller sub-graphs as the working model. However, such methods involve complicated pre-processing stages and expensive merge step at query time.

The use of statistical significance for mining connected subgraphs from vertex labeled graphs was recently studied in [39]. Based on the vertex labeling (for example, discrete set of biochemical entities ranging from molecules to genes [40]), the input graph was *compressed* using rule-based edge and vertex fusion (*contracting edges*) to form a smaller super-graph. The super-graph enabled a faster run-time complexity and was shown to preserve certain properties of the original graph (such as connected sub-graphs, etc.) along with preservation of 96% of the optimal χ^2 value. For each vertex in the super-graph, its *z-score* is computed using the weighted average of the neighborhood attributes (vertex label symbols, edge weights, etc.), thus modeling the structure of the current sub-graph under consideration.

Detection of *spatial outliers* is then performed by combining the individual z-scores, and a *chi-squared* based statistical score is computed from the multi-dimension z-scores to obtain a contiguous region with high significance (i.e., connected sub-graph outlier). This approach provided a framework for generic outlier detection for vertex labeled graphs with discrete as well as continuous labels. The approach was shown to provide an analysis of statistically significant connected sub-graphs (specifically, outliers) within large social networks, such as Orkut, DBLP, etc. within 3 hours considering continuous vertex labels.

VI. Open Directions of Research

The intelligent mapping of various data mining problems to statistical significance computations in the above applications have led to a reduction in run-time with high accuracy of results, forming the basic strategy for tackling queries on huge data stores. Hence, we observe that pattern mining using statistical significance holds potential for efficiently handling Web-scale data for diverse applications. In this section, we discuss a few further directions of research involving data significance as applied to real-time mining tasks.

- **Clustering:** Clustering involves the task of grouping together items depicting similar attributes. The analysis of clusters and its use thereof for recommendation, collaborative filtering, etc., forms a basic approach in data mining and information retrieval. However, certain scenarios such as relief-help distribution, traffic

 ISBN: 978-1-61208-423-7

congestion, social community popularity, etc., require detection of only the top-k clusters based on cardinality. They depict the most "crucial" areas and help resource concentration for better management. Hence, end-to-end clustering in such scenarios provides an inefficient approach.
However, the modeling of search space into k-dimensional matrix structures, and corresponding mapping of data points (represented by symbols with associated probability of occurrence) onto the cells for statistical significance computation (where more symbols generate more significance) might provide an alternative shortcut to intelligently tackle such huge data volumes. Further, the early and efficient identification of the most populous clusters and their analysis with *centrality measures*, such as *Katz centrality* [41] might help in faster epidemic control. The real-time nature of such approaches would help combat decision delays in situations of calamity.

- **Sub-graph Matching:** The problem of sub-graph isomorphism search has myriad applications for graph classification, electronics circuits, and protein interactions to name a few. However, finding sub-graph isomorphism is NP-hard; leading to the proposal of pruning-rule based approaches [42] and *combine and permute* indexing strategies [43] for approximate sub-graph matching to a query graph.
Similar to the approximate text matching, neighboring edge and vertex locality based sub-graph matching using χ^2 significance score might be performed by mapping vertices to the symbols based on their degree of similarity to structures in the query graph. This would enable efficient approximate sub-graph isomorphism for analyzing social community and other huge network graphs. Generalizing such approaches to graph mining and connected component association provides further research interest.

- **Skyline Queries:** Skyline involves the ranking of search results based on user-define preferences using the *Pareto dominance* criteria. However, the computation of relationship for every item-item pair provides a bottleneck for scalability of such methods. Hence, the expensive computation of skyline queries have been reduced by a number of caching approaches and efficient indexing structures, such as *closed skycubes* [44]. However, similar to the clustering approach, the encoding of data points in each dimension (user preference) to matching symbols, and the corresponding significance computation promises to capture the data points respecting the user constraints (at least in most of the specified preferences). Use of pruning mechanisms based on the significance score and extraction of top-k results might provide significant run-time and storage improvements in such scenarios.

Additionally, theoretical analysis of algorithms, under the ambit of statistical significance testing approach, to derive performance bounds for varying probability distributions of symbols also provides a pertinent area of future research across different communities.

VII. CONCLUSION

This paper presented an introductory survey of recent algorithmic trends in the applicability of classical *statistical significance* testing for the domain of big data analytics and deep mining from varied and huge data sources. We initially provide a brief background of the statistical measures commonly used and then discuss state-of-the-art approaches based on the Pearson's χ^2 measure (and others) to efficiently solve graph mining, text analysis, and approximate matching problems, among others. The use of statistical significance for mining tasks to extract interesting patterns and rules across diverse domains such as computational biology, social networks, etc., has been shown to provide enhanced accuracy, run-time, and scalability performance compared to state-of-the-art methods. We also enumerated a few possible exciting further research directions involving graph isomorphism and clustering, based on the statistical significance of observations.

ACKNOWLEDGMENT

The author would like to thank the *Google European Doctoral Fellowship* for financially supporting this work.

REFERENCES

[1] J. Manyika et al., "Big Data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, Tech. Rep., June 2011.

[2] J. Dean, "Large Scale Deep Learning," Keynote at CIKM (research.google.com/people/jeff/CIKM-keynote-Nov2014.pdf), 2014, retrieved: June 7, 2015.

[3] L. Deng and D. Yu, "Deep Learning: Methods and Trends," Foundations and Trends in Signal Processing, vol. 7, no. 3-4, 2013/14, pp. 197–387.

[4] A. Moreno et al., "Hybrid Model Rating Prediction with Linked Open Data for Recommender Systems," Communications in Computer and Information Science, vol. 475, 2014, pp. 193–198.

[5] H. Aguinis, L. Forcum, and H. Joo, "Using Market Basket Analysis in Management Research," Journal of Management, vol. 39, no. 7, 2013, pp. 1799–1824.

[6] N. Ye and Q. Chen, "An anomaly detection technique based on chi-square statistics for detecting intrusions into information systems," Quality and Reliability Engineering International, vol. 17, no. 2, 2001, pp. 105–112.

[7] M. Regnier and M. Vandenbogaert, "Comparison of statistical significance criteria," Journal of Bioinformatics and Computational Biology, vol. 4, no. 2, 2006, pp. 537–551.

[8] J. Leskovec, A. Rajaraman, and J. Ullman, Mining of Massive Datasets. Infolab: Stanford University (http://www.mmds.org/), 2015, retrieved on: June 7, 2015.

[9] T. Read and N. Cressie, Goodness-of-fit statistics for discrete multivariate data. Springer, 1988.

[10] G. Bejerano, N. Friedman, and N. Tishby, "Efficient exact p-value computation for small sample, sparse and surprisingly categorical data," Journal of Computational Biology, vol. 11, no. 5, 2004, pp. 867–886.

[11] S. Rahmann, "Dynamic programming algorithms for two statistical problems in computational biology," in Workshop on Algorithms in Bioinformatics (WABI), 2003, pp. 151–164.

[12] H. Hotelling, "Multivariate quality control," Techniques of Statistical Analysis, vol. 54, 1947, pp. 111–184.

[13] T. Read and N. Cressie, "Pearson's χ^2 and the likelihood ratio statistic G^2: a comparative review," International Statistical Review, vol. 57, no. 1, 1989, pp. 19–43.

[14] S. S. Wilks, "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses," The Annals of Mathematical Statistics, vol. 9, 1938, pp. 60–62.

[15] J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 231, no. 694706, 1933, pp. 289–337.

 ISBN: 978-1-61208-423-7

[16] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," Philosophical Magazine Series 5, vol. 50, no. 302, 1900, pp. 157–175.

[17] R. Goonatilake, A. Herath, S. Herath, S. Herath, and J. Herath, "Intrusion detection using the chi-square goodness-of-fit test for information assurance, network, forensics and software security," Journal of Computing Sciences, vol. 23, no. 1, 2007, pp. 255–263.

[18] R. Povinelli, "Identifying temporal patterns for characterization and prediction of financial time series events," IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 2, 2003, pp. 339–352.

[19] M. Kaboudan, "Genetic programming prediction of stock prices," Computational Economics, vol. 16, no. 3, 2000, pp. 207–236.

[20] A. Denise, M. Regnier, and M. Vandenbogaert, "Assessing the statistical significance of overrepresented oligonucleotides," in WABI, 2001, pp. 537–552.

[21] I. Kuznetsov and S. Rackovsky, "Identification of non-random patterns in structural and mutational data: the case of Prion protein," in CSB, 2003, pp. 604–608.

[22] S. Agarwal, "On Finding the most statistically significant substring using the chi-square measure," Master's thesis, Indian Institute of Technology, Kanpur, 2009.

[23] S. Dutta and A. Bhattacharya, "Most Significant Substring Mining Based on Chi-Square Measure," in PAKDD, 2010, pp. 319–327.

[24] A. Bhattacharya and S. Dutta, "Mining Statistically Significant Substrings Based on the Chi-Square Measure," in Pattern Discovery and Sequence Mining: Applications and Studies. IGI Global, 2011.

[25] M. Sachan and A. Bhattacharya, "Mining Statistically Significant Substrings using the Chi-Square Statistic," VLDB, vol. 5, no. 10, 2012, pp. 1052–1063.

[26] P. Ng, "Statistical Significance for DNA Motif Discovery," Ph.D. dissertation, Cornell University, 2011.

[27] D. Lovell, "Biological Importance and Statistical Significance," Journal of Agricultural and Food Chemistry, vol. 61, no. 35, 2013, pp. 8340–8348.

[28] K. A. Romer, G. R. Kayombya, and E. Fraenkel, "WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches," Nucleic Acids Research, vol. 35, 2007, pp. 17–20.

[29] L. Kuttippurathu et al., "CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments," Bioinformatics, vol. 27, no. 5, 2011, pp. 715–717.

[30] M. Frith, J. Spouge, U. Hansen, and Z. Weng, "Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences," Nucleic Acids Research, vol. 30, no. 14, 2002, pp. 3214–3224.

[31] S. Dutta, "MIST: Top-k Approximate Sub-String Mining using Triplet Statistical Significance," in ECIR, 2015.

[32] V. Levenshtein, "Binary Codes capable of Correcting Deletions, Insertions, and Reversals," Soviet Physics Doklady, vol. 10, no. 8, 1966, pp. 707–710.

[33] D. Fenz, D. Lange, A. Rheinlnder, F. Naumann, and U. Leser, "Efficient Similarity Search in Very Large String Sets," in Springer, 2012, pp. 262–279.

[34] C. Li, B. Wang, and X. Yang, "VGRAM: Improving Performance of Approximate Queries on String Collections using Variable-length Grams," in VLDB, 2007, pp. 303–314.

[35] D. Deng, G. Li, J. Feng, and W. Li, "Top-k string similarity search with edit-distance constraints," in ICDE, 2013, pp. 925–936.

[36] J. Yedidia, W. Freeman, and Y. Weiss, "Understanding Belief Propagation and Its Generalizations," in Exploring Artificial Intelligence in the New Millennium. Morgan Kaufmann Publishers Inc., 2003, pp. 239–269.

[37] C. Jiang, F. Coenen, and M. Zito, "A Survey of Frequent Subgraph Mining Algorithms," The Knowledge Engineering Review, vol. 28, no. 1, 2013, pp. 75–105.

[38] L. Zhu, W. Ng, and C. J., "Structure and Attribute index for approximate graph matching in large graphs," Information Systems, vol. 36, 2011, pp. 958–972.

[39] A. Arora, M. Sachan, and A. Bhattacharya, "Mining Statistically Significant Connected Subgraphs in Vertex Labeled Graphs," in SIGMOD, 2014, pp. 1003–1014.

[40] C. You, L. Holder, and D. Cook, "Temporal and structural analysis of biological networks in combination with microarray data," in CIBCB, 2008, pp. 62–69.

[41] L. Katz, "A New Status Index Derived from Sociometric Index," Psychometrika, vol. 18, 1953, pp. 39–43.

[42] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "A (sub)graph isomorphism algorithm for matching large graphs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 10, 2004, pp. 1367–1372.

[43] W. Han, J. Lee, and J. Lee, "$Turbo_{ISO}$: Towards UltraFast and Robust Subgraph Isomorphism Search in Large Graph Databases," in SIGMOD, 2013, pp. 337–348.

[44] C. Raïssi, J. Pei, and T. Kister, "Computing Closed Skycubes," VLDB, vol. 3, 2010, pp. 838–847.

 ISBN: 978-1-61208-423-7

Combining Machine Learning with Shortest Path Methods

Discovering, Visualizing, and Analyzing Hollywood's Power Clusters to Go From Six Degrees of Kevin Bacon to Knowing Colin Firth

Armand Prieditis
Neustar R&D
San Francisco, CA USA
email: armand.prieditis@neustar.biz

Chris Lee
Neustar Labs
Mountain View, CA
email: chris.lee@neustar.biz

***Abstract*—This paper describes a method to model, discover, and visualize communities in social networks. It makes use of a novel method based on the "Six Degrees of Kevin Bacon" principle: find the shortest path between entities in a social graph and then discover communities based on clustering with those shortest-path distances. We have applied this idea to find Hollywood's power clusters based on IMDB (Internet Movie Database), which links actors to movies. Using this method, we found roughly three clusters of Hollywood elite actors, the largest of which contained many of Hollywood's best-known actors. For living actors, we found Colin Firth (who played *Pride and Prejudice*'s Mr. Darcy), Javier Bardem (who played a psychopathic killer in *No Country for Old Men)*, and Joaquin Phoenix (who played Johnny Cash and a Roman Emperor in *Gladiator*) to be some of the most well-connected actors in Hollywood. This suggests that analyzing a social network using our method can lead to some surprising results**.

Keywords-Social networks; modeling; disovery; visualization; clustering; influence analysis; machine learning.

I. INTRODUCTION AND MOTIVATION

What is a **social network?** Typically, names such as Facebook, Twitter, or Google+ spring to mind when one thinks of a social network because that is the moniker these websites adopt. While these websites are not the only type of social networks, they are good examples of networks that are "social." This is because they comprise: a set of **entities** that participate in the network. In social networks such as Facebook, Twitter, and Google+, the entities are people. In general, entities do not have to be people. They also comprise a set of **relations** between the entities. For example, in Facebook, the relations are called "friends." In Twitter, they are follower and followee relationships. Finally, social networks comprise **weights** on those relations. For example, the higher the weight, the stronger the relationship. While most social networks such as Facebook, Twitter, and Google+ are all or none weights (i.e., you are either a friend or not; either a weight of a 1 or a 0), other social networks could have the degree of the relationship expressed as a weight. This degree might not be explicit. For example, how often someone reads the postings of someone they follow could be could be used to determine the weight.

Most real-world networks are not random and exhibit locality. That is, a randomly constructed network rarely looks like a real-world network and the intuition behind locality is that the relationships among entities tend to cluster somehow. For example, if X knows both Y and Z, then Y and Z probably know each other. One reason that Y and Z might know each other is that that they both comment on X's postings and hence they eventually discover and befriend each other based on those postings. Or, X could have introduced Y to Z other either online or in the real-world, based on X as a mutual friend.

This paper considers methods by which Y and Z could (or should) get to know each other by the modeling, discovery, and visualization of local communities that they share. More generally, this article touches on social influence in the sense of how a community influences the individuals in the community. Social influence is an active area of research because it aims to understand how information, memes, ideas, knowledge, experience, and innovation spread in a social network. Thus, analyzing and mining social networks can provide insights into how people interact and why certain ideas, memes, and opinions spread in the network and others do not. Although this paper describes a specific clustering method, it is not about clustering. That is, many different clustering methods could be used and we would expect comparable results. This paper is about how shortest path methods can improve upon clustering in social networks.

Discovery of communities can also be viewed as **link prediction**. Clearly, social networks are dynamic and constantly evolving and methods that can *anticipate* future links, such as link prediction, are important. As the network evolves, two unconnected nodes in the same community may eventually form a link between them. The intuition is that if future links can be predicted, the growth of a social network can be facilitated. Moreover, the relationships of the entities might be more satisfying from discovering other like-minded people faster. Thus, link prediction can be used to model how a social network evolves over time.

A. Social Networks are Ubiquitous

While social networks such as Facebook, Twitter, and Google+ capture the mindshare of the term "social network," social networks go beyond mere friend networks. In fact, the entities do not even have to be people to be

 ISBN: 978-1-61208-423-7

considered a social network. That is, a social network does not necessarily have to be in a social context. For example,

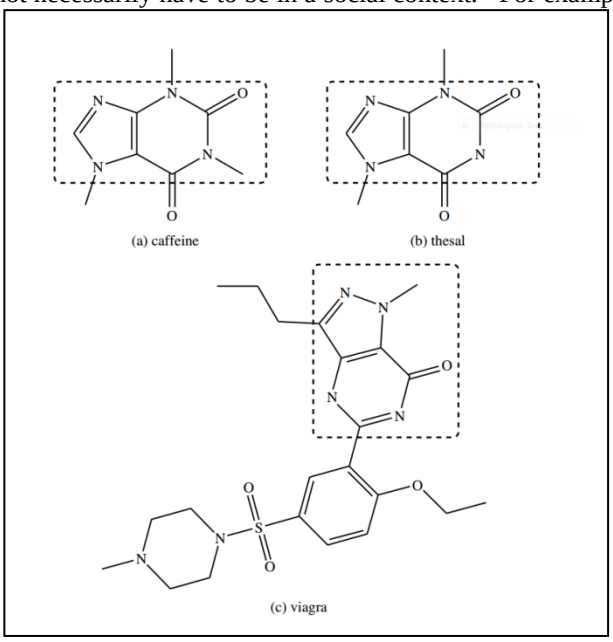


Figure 1. A "Social' Network in Chemistry

social networks *that* are non-social in context include electrical power grids, telephone call graphs, the spread of computer viruses, the World Wide Web, and co-authorship and citation networks of researchers. In a citation network, the entities might represent individuals who have published research papers and the relations between the entities might be researchers who jointly co-authored one or more papers. Weights might include the number of joint publications—the higher the weight, the more joint publications. The communities one might be able to discover in this network might include researchers working in the same area. Other such social networks might be possible to construct. For example, two Wikipedia editors can be related if they've edited the same article. Alternatively, the articles themselves can be the entities, which are linked if they have been edited by the same person.

More generally, social networks and their characteristics can often be generalized to networks found in a diversity of fields such as biology, chemistry economics, mathematics, and physics. For example, Figure 1 shows a social network in chemistry, where the entities are atoms and the relations are bonds. In the figure, (a) shows the caffeine molecule, (b) shows the thesal molecule, and (c) shows the Viagra molecule. All of these molecules are biologically and pharmaceutically important and hence their network analysis of activity is important.

Social networks can also include collaborative filters, where recommendations are based on customer preferences. Such networks can be viewed from the point of view of the customers as entities and the relations expressing customers who bought the same products. Such networks can also be viewed from the point of view of the products as the entities and the relations expressing products that were bought by the same customer.

Determining the entity vs. the relation can get complicated. For example, **users** can place **tags** on **websites** on social tagging sites, such as deli.cio.us. Users can be connected to other users based on tags they place on the same website. Alternatively, users can be connected to other users based on the *type* of tags they use. Both of these, can, of course, be flipped: websites can be connected based on the same users; tags can be connected based on the same users or websites.

Biological networks include epidemiological models, cellular and metabolic networks, food webs, and neuronal connections. The exchange of email or communication messages can also form social networks within corporations, newsgroups, chat rooms, friendships, dating sites, and corporate control (i.e., who serves on what boards). The entities in an email network represent individuals and a relation between entities can include an email exchange in any direction between two individuals. A weight might mean the number of emails between two individuals in a given period. This view distinguishes normal emailers from spammers: a normal emailer has higher frequency communication with a small set of individuals whereas a spammer has low frequency communication with a large set of individuals.

In a telephone network, the nodes might represent the phone numbers and relations might include two phones that have been connected over some period of time. Weights might include the number of calls.

Thus, many different networks bear similarities in terms of how social networks can be explicitly or implicitly derived from them: for paper citation networks entities might be papers or people and a relation exists if one paper cites another or the same paper was co-authored by two people. For collaboration **networks** entities might be people

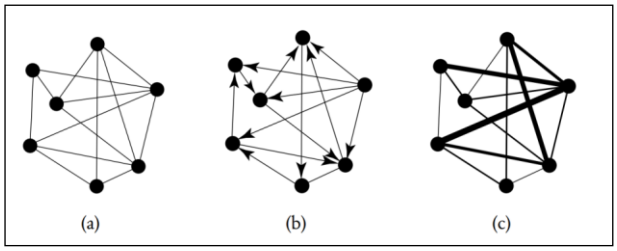


Figure 2. Social Networks as Graphs: Undirected (a), Directed (b), and Weighted (c)

and a relation expresses one person working with another. For semantic word graphs, such as in a dictionary or a thesaurus, entities might be words and a relation exists between two words if they are associated with each other. For biological networks, entities might be processes and a relation exists if two processes are related (e.g., protein or drug interactions). For news networks entities might be events, people, or words and relations might be causal links or people in common.

 ISBN: 978-1-61208-423-7

B. *Social Networks as Graphs*

A reasonable way to model a social network is as an undirected or directed or undirected graph. In an undirected graph, the entities are modeled as nodes and the relations are modeled as edges. The weight is represented by a labeled edge. Typically, the relations require a directed graph because the distinction between a follower and a followee is important. That is, if X follows Y then there is a directed edge between X and Y, but not necessarily vice versa. Informally, one can say that X "points to" Y. Note that this relationship could have been modelled the other way, with Y pointing to X, but the in-pointers to a node are typically more important than the out-pointers. That is, the people who follow you are a stronger sign of a relationship than the people who you follow because you have control over who you follow but not vice versa. For example, one could follow Lady Gaga, but that means little to most people. But if Lady Gaga follows you, that means a lot. In short, a directed graph can capture relationships that are one way, but not the other. Social relationships modeled as directed graphs are common in the real-world, so common that phrases such as "unrequited love" have been invented in order to capture them.

Figure 2 illustrates undirected, directed, and weighted graphs. For example, (a) shows an undirected graph. The nodes are the dark circles and the undirected edges are the lines. Graph (b) illustrates a directed graph. The nodes are the same, except the lines are now directed. Graph (c) illustrates an undirected weighted graph, where the thickness of the edge is proportional to the weight of the edge.

C. *Discovering Communities in a Social Network*

Figure 3 shows a social network represented as a graph with nodes A, B, C, D, E, F, and G and undirected edges as the relations between nodes. Visually, nodes A, B, and C seem more closely related to each other than to the other

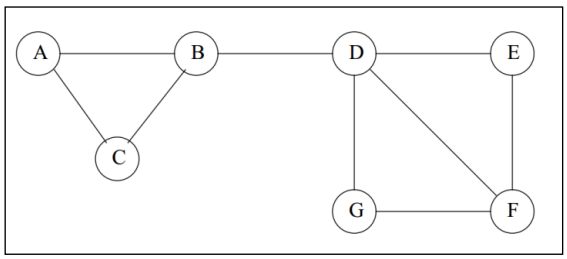


Figure 3. Visual Discovery of Communities by Distance

nodes. Similarly, nodes D, E, F, and G seem more closely related to each other than to the other nodes. Thus, one way in which the nodes can be clustered or groups is in terms of distance to each other. More specifically, the act of clustering can be viewed as discovering a **community** in a social network. The intuition is that nodes A, B, and C might have something in common, at least more in common than with nodes D, E, F, and G. In short, one way to discover communities is to group by distance in terms of relations. An important aspect of a social network is that it can be implicit, by virtue of liking the same things, visiting the same sites, or having similar attributes. One important task is to discover *homophily*, which can be viewed as discovering communities in a social network.

Another way to discover communities is to form groups based on common attributes. For example, Figure 4 shows a graph coloring based on interest in music, sports, and cooking. In this case, the nodes A, B, C, and D form one cluster, nodes C, H, I, and J form another cluster, and nodes D, E, F, G form a third cluster. Note that in this case, the nodes might have been grouped similarly by their relations instead of their attributes. Thus, it might be likely that nodes sharing relations are interested in some of the same things (i.e., have the same attributes). Otherwise, such nodes would have little basis for interacting with each other.

Although the concerns are different in different fields, the idea of community discovery can be treated similarly, as described here. Indeed, one way that complex networks and complex systems can be understood is by discovering structures in the form of communities in them. Human cognition often prevents analyzing the network as a whole; finding communities is a way to simplify a network into a small set of communities, so that human cognition can then take over. In short, this paper recognizes that modeling,

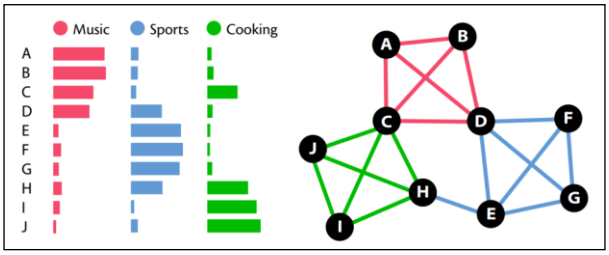


Figure 4. Communities Discovered by Common Interests

discovery, and visualization of communities in networks is a general methodology applicable to most real-world networks. It also recognizes that finding an appropriate division of labor between humans and machines is important to combine the unique cognitive strengths of humans with the tireless computational abilities of machines.

D. *Modeling, discovering, and visualizing communities in Hollywood actor networks*

We wanted to test our ideas for modeling, discovering, and visualizing communities on a large enough data set to produce interesting results. For this reason, we choose the IMDB (Internet Movie Database) [1], which is publicly available. The database contains hundreds of thousands of movies, many of which are obscure, and thousands of actors, most of whom are obscure bit players. This database can be viewed as a bipartite graph where each node either corresponds to an actor or to a movie. In this graph, an edge between an actor and a movie means that the actor appeared

 ISBN: 978-1-61208-423-7

in the movie. The task is to model this data somehow to make it easy to discover and visualize communities.

E. Organization of the Rest of this Paper

The rest of this paper is organized as follows. Section II describes related work. Section III describes our approach to modeling, discovering, and visualizing communities in a social network. Section IV summarizes our results. Finally, Section V presents our conclusions and several promising areas for research.

II. RELATED WORK

Discovering communities in a social network can be viewed as clustering. As such, researchers have used two general approaches to clustering: hierarchical or agglomerative [2] and divisive [3]. Both approaches require a distance metric. When the edges have weights, those weights can be used as a distance measure. The difficulty arises when the edges are unlabeled, as in most online social networks: the "friends" network. It's possible to use a weight of 1 or 0 for a direct edge and a large weight for those without a direct "friends" relationship, but these measures violate the triangle inequality principle of a distance metric, which generally causes anomalies in clustering.

Assuming a suitable distance measure can be found, researchers have defined the distance between clusters as the minimum distance between two nodes of each cluster. Hierarchical clustering first combines two nodes connected by an edge. It then chooses at random edges that are not between the two nodes in the cluster to combine the clusters to which each of the two nodes below. This agglomeration continues until an appropriate criterion is met. Divisive clustering proceeds in the opposite direction: starting with one giant cluster, it successively seeks edges that break the cluster into smaller and smaller parts.

These standard clustering methods have produced somewhat unsatisfactory results in social networks. As a result, researchers have developed specialized clustering methods aimed specifically at finding communities in social networks. One method, a divisive one, is based on finding an edge that is least likely to be in cluster and then removing it. This method uses the Girvan-Newman (GN) algorithm [4] to calculate the number of shortest paths running between every pair of nodes. An edge with a high GN score is a candidate for removal. The GN algorithm essentially conducts a breadth-first search of the graph and counts the number of times the same edge is encountered for all pairs of nodes.

Thus, by using the GN-based score, this specialized clustering method removes edges, which has the effect of decomposing the graph into subcomponents. The process beings with the initial graph and then each time it removes that edge with the highest GN score until the graph is decomposed into an appropriate number of connected components.

Another approach uses matrix theory (i.e., spectral methods as in [5]) to partition a graph such that the number of edges that connect different components is minimized. But such "cut-based" methods are unstable because cuts are not desirable that break the two components into unequal size.

In general, the approaches to finding communities in social networks have been somewhat unsatisfactory, often relying on arbitrary distance measures.

Social network analysis is an active area of research and this paper can be considered part of that work. For example, a Google search reveals nearly three hundred conferences on or related to social network analysis. A recent book [6] describes some of the network relational structures described here. Moreover, the Web Science conferences have been publishing leading work in social network analysis since 2009. Related work in these conferences includes research on six degrees of separation in social networks [7], clustering users on social discussion forums based on roles [8], topic-author networks [9], influence detection in networks [10], status evaluation [11], four (not six) degrees of separation [12], the spread of misinformation in a social network [13], and social graph annotation based on activities [14]. All of these results are consistent with the results presented here. For example, we found, just as in [12], that much less than six degrees separate most actors.

III. OUR APPROACH TO MODELING, DISCOVERING, AND VISUALIZI'G COMMUNITIES IN SOCIAL NETWORKS

We began our process with the IMDB [1], which links actors to movies. Next, we converted this bipartite graph into a social network where actors are the entities and a relation between one actor and another means that the two actors have appeared in the same movie.

Even though we built our graph with the entire IMDB, we focused on the top 100 actors of all time (based on IMDB, 62 of which were all in the same connected component, which we focused on for computational efficiency and presentation brevity): Jack Nicholson, Marlon Brando, Al Pacino, Daniel Day-Lewis, Dustin Hoffman, Tom Hanks, Anthony Hopkins, Denzel Washington, Spencer Tracy, Laurence Olivier, Jack Lemmon, Gene Hackman, Sean Penn, Johnny Depp, Jeff Bridges, Gregory Peck, Ben Kingsley, Leonardo DiCaprio, Tommy Lee Jones, Alec Guinness, Kevin Spacey, Javier Bardem, Humphrey Bogart, Clark Gable, George C. Scott, Jason Robards, Peter Finch, Charles Chaplin, James Cagney, Burt Lancaster, Cary Grant, Sidney Poitier, Alan Arkin, Samuel L. Jackson, Sean Connery, Christopher Walken, Heath Ledger, Jamie Foxx, Colin Firth, Joaquin Phoenix, Jeremy Irons, George Clooney, Tom Cruise, Matt Damon, John Hurt, Brad Pitt, Nicolas Cage, John Travolta, Clint Eastwood, Orson Welles, Charlton Heston, Henry

 ISBN: 978-1-61208-423-7

Fonda, Ian McKellen, Liam Neeson, Woody Allen, John Malkovich, Mickey Rourke, Danny DeVito, Robert Mitchum, Buster Keaton, Harvey Keitel, and Martin Sheen.

We also explored the top 250 and top 1000 actors (as ranked by IMDB) and obtained similar results. However, we found that the top 100 list adequately captured the core ideas well while making the results convenient for presentation here. Another reason we focused on this top 100 list of well-known actors is that when we ran our system on the largest set of actors (i.e., the entire set of actors in the IMDB movie database) we found that the cluster centers were comprised of these relatively unknown actors: Stéphanie Sokolinski, Olivier Rittano, Magid Bouali, David Luraschi, Simon Muterthies, David Vincent, Stéphanie Blanc, Anne Comte, Juliette Goudot, and Anne Nissile. Since no one among our associates could recognize even a single actor in these clusters, we felt that the interested reader would get a better feel for our system if the actors were "well-known" even though the clustering is on *all* the actors in the IMDB movie database. We simply ignore the less-known actors even though they are behind the scenes in the clustering. Note that the appearance of these less-known actors near the cluster centers does not mean that our approach does not work. It merely means these less-known actors happened to locate near the center because they greatly outnumber well-known actors. That is, because of their large numbers, a less-known actor is more likely to appear near a center than a well-known actor. Indeed, watching the credits roll by at the end of any typical modern movie confirms that only a few actors in that roll are well-known.

Next, we added an edge between each actor in the same movie. For example, Danny DeVito and Jack Nicholson were in *One Flew Over the Cuckoo's Nest* and hence they are connected with a single link. Thus, the initial graph we built contains only *direct* social relationships between actors. In our social network the entities are the actors and the relations that link them are joint appearance in a movie, but we could have just as easily built a social network where the entities are the movies and the relations that link them are joint appearance of actors in both movies. We choose the former because we were more interested in finding out the "Hollywood power clusters." That is, we were interested in discovering which well-known actors would turn out to be at the center of the largest clusters. We were also interested in finding out which actors were central to multiple clusters—which actors act as articulators in multiple clusters. Note that this type of analysis is unrelated to clustering, but is a post-clustering analysis.

As a result, we wanted to link *all* the actors together somehow. The problem, as with the distance measures that we mentioned, is that actors either have a link (i.e., a weight of 1) or they do not (i.e., a weight of 0). A high weight, as assigned in the previously mentioned research, is clearly unacceptable because there could just be a few actors separating any two actors. For example, the game of "Six Degrees of Kevin Bacon," assumes that any actor can be linked through his or her film roles to Kevin Bacon within six steps. (Sadly, Kevin Bacon does not make an appearance in this paper even though the title mentions his name. This is because he is not a member of the Hollywood power clusters we found.)

To combat what we call the "binary problem" of edges (i.e., either a 1 or nothing), we ran a shortest-path algorithm between all pairs of actors in all connected components, one such algorithm per connected component. We then focused on the largest connected component, which contained roughly 5000 actors. Here is an example of the edge weights between a few selected pairs of actors, emanating from Buster Keaton, silent movie star of the 1920's: Humphrey Bogart:1, Daniel Day-Lewis: 2, Matt Damon: 2, Javier Bardem: 2, Jamie Foxx: 2, Joaquin Phoenix:2, Henry Fonda:2, and Johnny Depp:2.

This example illustrates Buster Keaton's connection to both modern and old-time movie stars. For example, he's directly connected to Humphrey Bogart (having starred in the same film), but is only two connections away from Matt Damon. That is, he starred in a film with someone who starred in a film with Matt Damon, a modern movie star. We were not surprised that the "Six Degrees of Kevin Bacon" holds true, but we were surprised at how few steps away an actor of the 1920's was from actors of the new millennium, over 80 years later. Going the other way, from recent actors to old-time actors, we see that Jamie Foxx is similarly connected to both old and new actors: Humphrey Bogart: 2, Daniel Day-Lewis: 2, Matt Damon: 2, Javier Bardem: 1, Johnny Depp: 1, and Charles Chaplin: 2.

We would not have guessed that Jamie Foxx, who recently appeared in Tarantino's *Django Unchained*, is a mere two steps away from Charles Chaplin, silent movie star of the 1920's. Conducting a shortest-path analysis reveals such connections between actors. Thus, the motivation behind the shortest-path analysis is to compute *indirect* relationships, which we believe are as important as direct relationships in clustering and in discovering communities.

Next, we applied a clustering algorithm to find how the actors clustered based on these shortest-path distances in the largest connected component of actor relations. We choose **K-Means** clustering as the method to cluster the actors. K-Means clustering partitions the data points into K clusters such that each data point belongs to the cluster with the nearest mean [15]. Thus, each cluster's mean serves as a summary of the data points in the cluster. The resulting partition can be viewed as a set of Voronoi cells. We used Lloyd's algorithm to find the K means [16]. This algorithm begins with an initial random set of K means. Next, it assigns each data point to the nearest mean of the K means. It then recalculates the K means for each cluster and repeats the assignment. This continues until the assignments no longer change. Although there is no guarantee that a globally optimum set of assignments can be obtained (i.e.,

 ISBN: 978-1-61208-423-7

those that minimize the sum of a least squares fit between the data points and their closest clusters), multiple random restarts can increase the confidence that a globally optimum set of assignments can be founds. To start with good initial parameters, we used the K means ++ assignment algorithm [17], which is an effective way to ensure faster convergence by choosing better initial values. We choose the K-Means clustering method both because of its simplicity and because of its ability to deal with numerical values through a straightforward distance measure, which is consistent with the distance measure in our application.

IV. SUMMARY OF RESULTS

Using the standard estimate of the mean-squared error over all the data points, we obtained the following results for K-means clustering: K = 5: 32790, K = 25: 24957, K = 50: 21781. After K = 50, the train-and-test error rate began climb, so we stopped with K=50 and used that as the baseline K for all the results described here.

The largest cluster contained the following actors: Jack Nicholson, Marlon Brando, Al Pacino, Daniel Day-Lewis, Dustin Hoffman, Tom Hanks, Anthony Hopkins, Denzel Washington, Laurence Olivier, Jack Lemmon, Gene Hackman, Johnny Depp, Jeff Bridges, Gregory Peck, Ben Kingsley, Leonardo DiCaprio, Tommy Lee Jones, Alec Guinness, Kevin Spacey, George C. Scott, Jason Robards, James Cagney, Burt Lancaster, Cary Grant, Sidney Poitier, Samuel L. Jackson, Sean Connery, Christopher Walken, Heath Ledger, Colin Firth, Jeremy Irons, Tom Cruise, John Hurt, Brad Pitt, Nicolas Cage, John Travolta, Clint Eastwood, Orson Welles, Charlton Heston, Henry Fonda, Ian McKellen, Liam Neeson, Woody Allen, John Malkovich, Mickey Rourke, Danny DeVito, Robert Mitchum, Buster Keaton, Harvey Keitel, and Martin Sheen. Based on our cluster analysis, this largest cluster can be viewed as Hollywood's true power brokers in terms of their connections. In other words, cluster analysis shows this to be the true "A-list" of actors. Actors on this list tend to be tightly connected to each other.

The next largest cluster contained Spencer Tracy, Humphrey Bogart, Clark Gable, Peter Finch, Charles Chaplin, Jamie Foxx, and Joaquin Phoenix. These are also well-known power-brokers, but nothing like the first list. Finally, the next largest cluster contained the remaining actors: Sean Penn, Javier Bardem, Alan Arkin, George Clooney, and Matt Damon. The rest of the clusters (i.e., the other 47) contain nearly all unknown actors and hence we will not discuss them here. This suggests that it is difficult to break into the top three Hollywood power clusters.

Next, we "softened" the notion of cluster membership in K-means and found the list of the 10 closest actors to each cluster's center. Membership is "soft" because these actors might not necessarily be in the cluster. Names such as Jamie Foxx, Javier Bardem, and Spencer Tracy appear on many of clusters. Subsequently, we counted the number of times each actor appeared in the top 10 closest actors in each cluster and obtained the following results: Peter Finch (50), Spencer Tracy (49), Colin Firth (49), Charles Chaplain (48), Javier Bardem (48), Heath Ledger (48), and Joaquin Phoenix (48). After a big gap, Matt Damon comes in at 30. The rest of the actors do not appear in as many clusters as these. Among dead actors, Peter Finch, Spencer Tracy, and Heath Ledger would have been the ones to get to know to make Hollywood connections. Among living actors, Colin Firth, Javier Bardem, and Joaquin Phoenix appear to be the go-to guys to make connections. These actors can be viewed as major "articulators" who are well-connected to nearly everyone. Intuitively, this means that if you get to know these actors they might help you unlock the doors to the most power clusters in Hollywood. Colin Firth was a surprise to us. But then, upon closer examination, we found out that Colin Firth's films have earned more than $936 million and that he's had over 42 movie releases worldwide. Based on our analysis, our advice to a young actor interested in Hollywood social climbing is to get to know Colin Firth.

For the largest cluster (the "Jack Nicholson" one), Figure 5 shows a visualization of the ten closest actors in that cluster and their distance apart. We used the NetworkX Python facility [18] to produce a planar graph, given the inter-node distances. What is interesting about this visualization is that James Cagney appears to be the prototypical actor in this largest cluster. That is, he is most like the average member of this cluster than anyone else. For the next largest cluster (the "Spencer Tracy" one), Figure 6 shows a visualization of the ten closest actors in that cluster and their distance apart. Spencer Tracy sits comfortably in the middle of this cluster, even though he died over half-century ago. This visualization vividly demonstrates the temporal reach of good actors: they can die, but they never really leave Hollywood. For the third-largest cluster, Figure 7 shows that Colin Firth, who we have already said is worth getting to know for social climbing, is at the center of this web of actors.

V. CONCLUSIONS AND FUTURE WORK

Evaluating the quality of these clusters is difficult as there is no standard grouping of actors against which we can compare our results. It may be possible to borrow evaluation ideas from the research focused on "power" users in social networks [19], but this work lacks a clustering component.

Short of such an evaluation, these results can be viewed as the discovery of power communities among Hollywood actors. We believe that the process of **Modeling**, **Discovery**, and **Visualization of Communities,** as we have presented it, is a powerful way to analyze social networks. **Modeling** comprises **Choosing Entities and Relations → Building a Social Graph Based on Relations → Calculating an All-Pairs Shortest-Path Metric**. **Discovery** comprises **Finding the Parameters of a Piece-wise Linear Function** (i.e., this is what K-Means clustering discovers). **Visualization** comprises **Laying out the Nodes and Their Relations In the Discovered Communities on a Planar**

Graph such that the layout preserves the distance metric between nodes.

We believe this process is an appropriate division of labor between machines, which are good at mind-numbing calculations, and humans, who are good at detecting visual patterns. It is difficult to perceive visual patterns in a large multi-dimensional space such as that produced after the all-pairs shortest-path metric is calculated. However, once the discovery process is completed and the resulting communities are displayed on a planar graph, the human visual system, with all its virtues, can take over and unlock patterns difficult for machines to see. Without this discovery, these patterns are nearly impossible to visually unlock. We believe this type of discovery and analysis might be important in determining how to spend advertising dollars on the Internet: find those nodes that are most influential and spend the most money there. The scientific contribution of this paper is a way to combine shortest-path methods with clustering to yield better results.

Based on our results, our conclusion is that when it comes to well-known actors, there are only three Hollywood power clusters, with one cluster dominating the other two in terms of size. Some actors are more well-connected than others, namely Colin Firth, Javier Bardem, and Joaquin Phoenix.

Could similar results be expected for other types of networks? We have applied the same idea to geo-locating the world's routers [20]. This work builds a map of directly connected internet routers based on time delays between routers (as returned by the trace command), calculates the shortest path time-delay between all pairs of routers based on this map, and then clusters the results. In this application, the time delay is analogous to the degree of separation between actors.

We are currently investigating several promising directions for future work including a more sophisticated clustering algorithm (e.g., EM), adding attributes for additional clustering knowledge (e.g., when and where each actor was born and the types of roles for which they are known), and applying our idea to predicting the geographic location of the world's routers (the entities) based on the round-trip transit time between the routers (the relations).

Working with large social networks can be computationally difficult. We believe our method can be extended to networks will millions of nodes by making use of frameworks, such as Hadoop and Spark, which we are currently investigating. An important advantage of our method is that every step in the process we have described can easily be parallelized to make it scalable.

REFERENCES

[1] J. J. Jung. (2012). "Attribute selection-based recommendation framework for short-head user group: An empirical study by MovieLens and IMDB." Expert Systems with Applications, 39(4), 4049-4054.

[2] K. C. Gowda and G. Krishna. (1978). "Agglomerative clustering using the concept of mutual nearest neighbourhood." Pattern Recognition, 10(2), 105-112.

[3] S. M. Savaresi, D. Bolev, S. Bittanti, and G. Gazzaniga. (2002, April). "Cluster Selection in Divisive Clustering Algorithms." In SDM.

[4] M. E. Newman and M. Girvan. (2004). "Finding and evaluating community structure in networks." Physical review E, 69(2), 026113.

[5] U. Von Luxburg. (2007). "A tutorial on spectral clustering." Statistics and computing, 17(4), 395-416.

[6] J. Scott. (2012). Social network analysis. Sage.

[7] L. Zhang and W. Tu. (2009) "Six Degrees of Separation in Online Society." In: Proceedings of the WebSci'09: Society On-Line, 18-20 March 2009, Athens, Greece.

[8] J. Chan, C. Hayes, and E. Daly. (2010) "Decomposing Discussion Forums using Common User Roles." In: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, April 26-27th, 2010, Raleigh, NC: US.

[9] N. Naveed, S. Sizov, S and S. Staab. (2011) "ATT: Analyzing Temporal Dynamics of Topics and Authors in Social Media." In: Proceedings of the ACM WebSci'11, June 14-17, Koblenz, Germany, pp. 1-7.

[10] Chandra, P., & Kalyanasundaram, A. (2012, June). "A network pruning based approach for subset-specific influential detection." In Proceedings of the 3rd Annual ACM Web Science Conference (pp. 57-66). ACM.

[11] B. State, B. Abrahao, and K. Cook. (2012, June). "From Power to Status in Online Exchange." In Proceedings of the 3rd Annual ACM Web Science Conference (pp. 57-66). ACM.

[12] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. (2012, June). "Four degrees of separation." In Proceedings of the 4th Annual ACM Web Science Conference (pp. 33-42). ACM.

[13] N. P. Nguyen, G. Yan, M. T. Thai, and S. Eidenbenz. (2012, June). "Containment of misinformation spread in online social networks." In Proceedings of the 4th Annual ACM Web Science Conference (pp. 213-222). ACM.

[14] A. V. Sathanur and V. Jandhyala. (2014, June). "An activity-based information-theoretic annotation of social graphs." In Proceedings of the 2014 ACM conference on Web science (pp. 187-191). ACM.

[15] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. (2002). "An efficient k-means clustering algorithm: Analysis and implementation." Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(7), 881-892.

[16] C. N. Vasconcelos, A. Sá, P. C. Carvalho, and M. Gattass. (2008). "Lloyd's algorithm on GPU." In Advances in Visual Computing (pp. 953-964). Springer Berlin Heidelberg.

[17] S. Agarwal, S. Yadav, and K. Singh. (2012, March). "K-means versus k-means++ clustering technique." In Engineering and Systems (SCES), 2012 Students Conference on (pp. 1-6). IEEE.

[18] A. Hagberg, P. Swart, and D. Chult. (2008). "Exploring network structure, dynamics, and function using NetworkX." (No. LA-UR-08-5495). Los Alamos National Laboratory (LANL).

[19] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao. (2009, April). "User interactions in social networks and their implications." In Proceedings of the 4th ACM European conference on Computer systems (pp. 205-218). ACM.

 ISBN: 978-1-61208-423-7

[20] A. Prieditis and G. Chen. (2013). "Mapping the Internet: Geolocating Routers by Using Machine Learning." In Computing for Geospatial Research and Application (COM.Geo), 2013 Fourth International Conference on Computing for Geospatial Research and Application (pp. 101-105). IEEE.

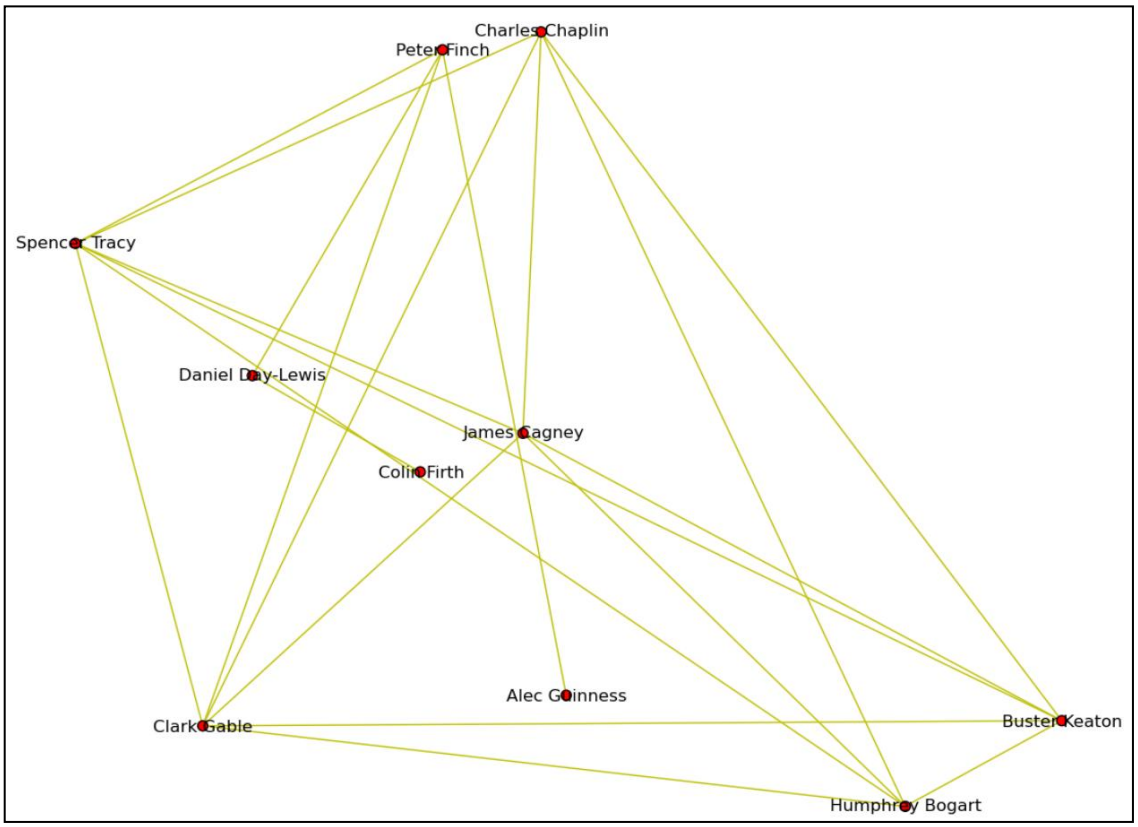


Figure 5. The Ten Actors Closest to the Center of the Largest Cluster

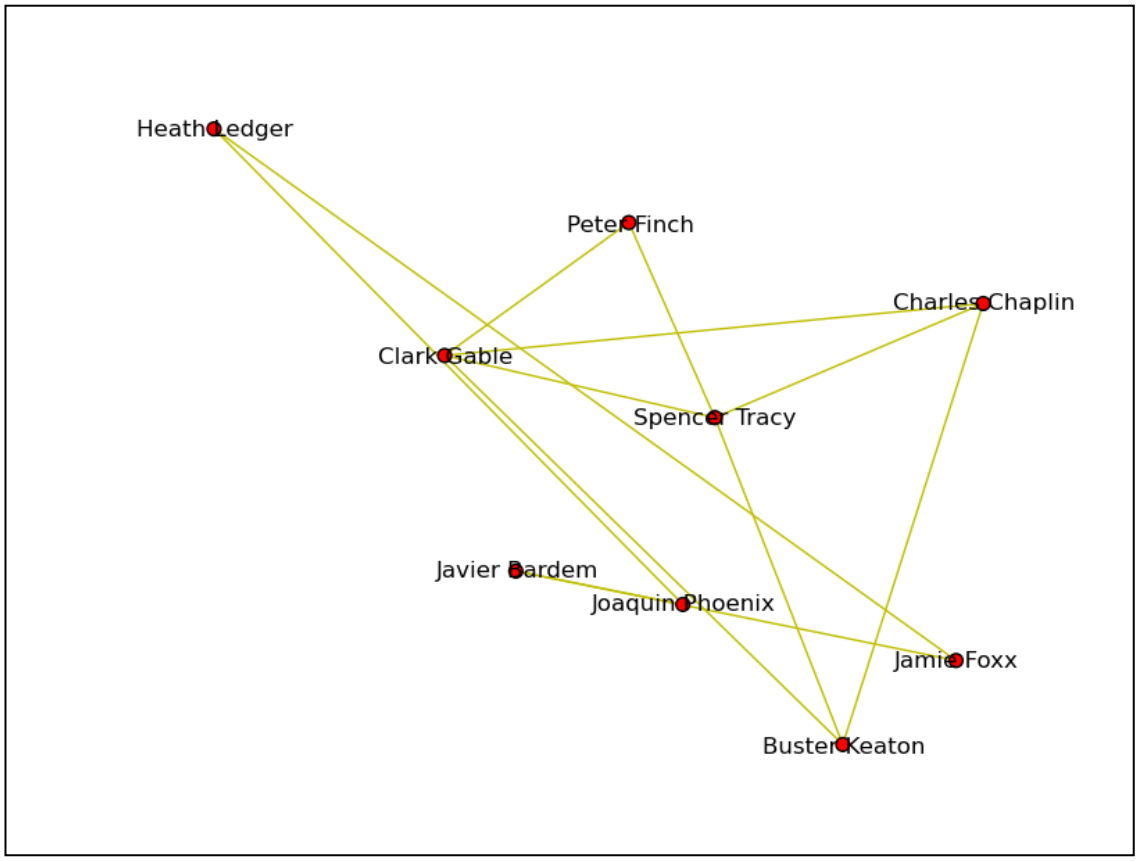


Figure 6. The Ten Actors Closest to the Center of the Next-Largest Cluster

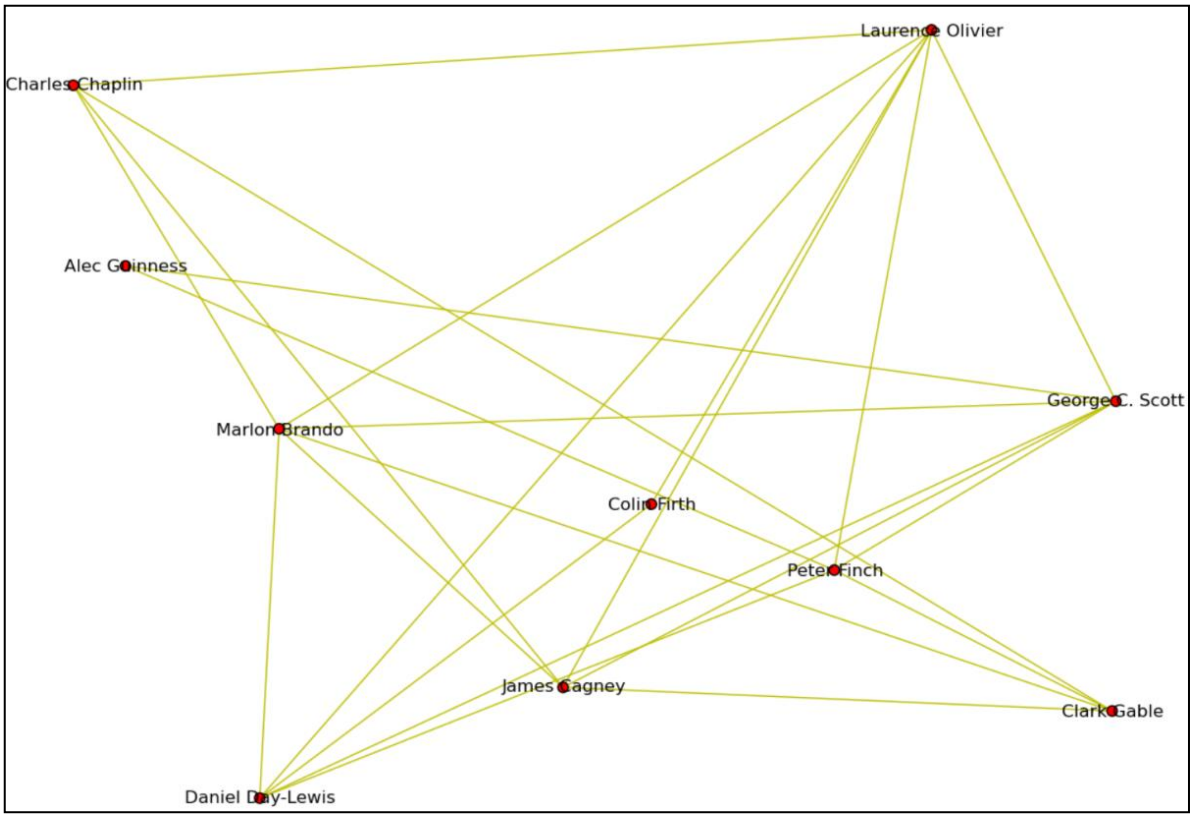


Figure 7. The Ten Actors Closest to the Center of the Third-Largest Cluster

 ISBN: 978-1-61208-423-7

Towards the Automated Identification of Orphan Diseases From Case Descriptions

Christian Rohrdantz*, Andreas Stoffel*, Franz Wanner*, Martin Drees†
*Department of Computer & Information Science, University of Konstanz, Germany
Email: firstname.lastname@uni-konstanz.de
†coliquio GmbH, Konstanz, Germany

***Abstract*—Orphan diseases are very rare diseases that are not well-known to many medical doctors. Patients suffering from them often remain without the correct diagnosis. Yet, there is a potential that advice-seeking doctors, posting medical case descriptions in web forums, may be automatically given a hint to matching orphan diseases. In this work-in-progress paper, we investigate opportunities and issues for an automated identification of orphan diseases in medical case descriptions through text mining and data analytics.**

Keywords–Data Analytics; Text Mining; Medical Data Analytics.

I. INTRODUCTION

It is the daily work of most medical doctors to examine patients and then combine observations on clinical signs and symptoms with their knowledge and experience in order to arrive at diagnoses. Accurate and timely diagnoses are crucial for initiating successful treatments. Yet, there are cases where medical doctors are confronted with patients showing symptoms that do not fit well into the known patterns. In these cases, physicians consult literature and seek the advice of experienced colleagues.

But what if the disease of the patient is just so extremely rare that only a handful of experts worldwide would be able to identify it based on the given clinical signs? There is a quite high chance that these cases end up with unspecific or wrong diagnoses and do not get the optimal treatment. Yet, it is known that there are quite a number of such rare, so-called orphan diseases.

In recent years, doctors have increasingly made use of medical web forums in order to seek advice from colleagues and discuss cases. In Germany, the largest and most active online community of medical doctors is coliquio [1]. It is experiencing a fast growth and has currently already more than 125,000 members. Without a doubt, it would be of great value if the case descriptions of advice-seeking physicians could be automatically matched with the known orphan diseases. These physicians could then be hinted by the system to the corresponding orphan disease in cases where a match seems likely.

Consequently, in this paper we describe work-in-progress towards such an automated identification of orphan diseases. The contribution of this initial study is twofold. First, we investigate how freely available, stuctured resources of medical knowledge, like orphanet [2], [3], and large repositories of medical texts, like the Wikipedia Portal Medicine [4] , can be exploited for our purpose. Second, we present a statistical method for the extraction of contextual knowledge and terminology, and show that it yields a lot of relevant information that could not be gathered from common dictionaries and medical ontologies.

The remainder of this paper is organized as follows. First, in Section II, we provide the necessary background information on rare or orphan diseases and the related work in this area, before we describe the data sources used in Section III. Next, in Section IV, we introduce our novel approach of leveraging the described data for the potential detection of orphan diseases from medical case descriptions. In Section V, we present first results, evaluate the performance, and identify issues and opportunities. Finally, we draw conclusions and identify key challenges setting the agenda for future work in Section VI.

II. RELATED WORK

Orphan or rare diseases fit into the broader context of research regarding rare events [5] and events in text data [6], but to date have not been treated in these areas.

The definitions of orphan disease vary slightly in the literature. The European Organisation for Rare Diseases (EURORDIS) states on their Websites:

"A rare disease, also referred to as an orphan disease, is any disease that affects a small percentage of the population. Most rare diseases are genetic, and are present throughout a person's entire life, even if symptoms do not immediately appear. In Europe, a disease or disorder is defined as rare when it affects less than 1 in 2000 citizens." [7].

Despite of the rareness of individual diseases, the overall quantity of affected people is still quite high:

"There are more than 6000 rare diseases. On the whole, rare diseases may affect 30 million European Union citizens." [8]

The coverage of orphan diseases in standard terminologies is very limited [9]. Rath et al. [10] state that only 446 orphan diseases have a specific code in the ICD10 disease classification, which most European countries use in their health information systems.

In general, text mining for clinical medical records is an important field of research [11], but there is few work on symptom or disease identification. Koeling et al. [12] manually annotate symptoms in patient records and provide statistical information on the frequency distribution of symptoms. They come to the conclusion that "there is great variation in the expressions used to describe the same symptom". Data from orphanet has been used exploiting the given mapping between diseases and disease-causing genes [13], but not for text mining purposes. Our approach fills a clear gap in the current research.

III. DATA

A. Orphadata

Orphanet provides structured textual data on orphan diseases and indicative clinical signs as part of their orphadata

 ISBN: 978-1-61208-423-7

service [14]. We made use of the XML version of the data in German. The data contains information about 2689 different orphan diseases and their clinical signs. Overall, the data contains 1362 different clinical signs and information on their frequency for different diseases. Moreover, the clinical signs are organized hierarchically in a thesaurus structure from rather general to more specific signs. Each clinical sign is typically described by one or more synonyms or alternative expressions. For example, one of the clinical signs is named "Nausea/vomiting/regurgitation/merycism/hyperemesis". We will refer to each of these alternative expressions as *symptoms*. For each orphan disease, different clinical signs may have three different frequency values: *very frequent*, *frequent*, and *occasional*. While the data is available for different languages, in this initial study we use the German version only.

B. Wikipedia Portal Medicine

The Wikipedia Portal Medicine constitutes a rich body of diverse textual medical information. We leverage the German version of this resource in order to automatically extract context knowledge and feed it into our text mining models.

IV. Mining and Modeling Medical Knowledge from Text

One of the services orphanet provides is that a user can select different clinical signs from the controlled thesaurus through a web interface and retrieve potentially matching orphan diseases. The big challenge we face, however, is to automatically identify mentions of clinical signs in medical case descriptions. In only very few cases, physicians use explicitly and exactly the terminology given in controlled vocabularies like the orphanet thesaurus. Mostly, they will use either inflected word forms, alternative wordings, varying multi-word expressions, paraphrasing or abbreviations. Our approach consists in learning the alternative terminology applying advanced statistical methods to large text repositories, such as the Wikipedia Portal Medicine. The advantage is that such a source contains expressions as they are actually used by phycisians rather than controlled idealized language use. For the mining of medical knowledge from texts, we proceed different consecutive steps.

A. Step 1: Identifying Descriptive Contexts

As mentioned before, each clinical sign in the orphadata is described by a set of symptoms. In order to get hold of textual contexts describing symptoms, we query Wikipedia. For each symptom, our first attempt is finding an article where the title exactly matches the given symptom. Such an article has basically been written to describe the symptom and consequently we consider all of the text in the article to be related to the symptom. For us, it constitutes what we define as a *descriptive context*. If, however, there is no matching article, we make use of the common search capability of Wikipedia. We use the symptom as a query and then sift through the retrieved articles. For each of these articles, we first check whether it belongs to the category "medicine" or one of its more than 400 subcategories. From each article meeting this criterion, we extract those paragraphs, where the symptom appears and save them as descriptive contexts.

B. Step 2: Extracting Knowledge from Descriptive Contexts

Next, from the available descriptive contexts we build two kinds of data co-occurrence tables. First, for each noun we count how often it occurs within the descriptive contexts of each symptom. This gives us a noun-symptom co-occurrence table. Next, for each pair of nouns, we count how frequently they co-occur within descriptive contexts. This gives us a noun-noun co-occurrence table for symptom contexts. From now on, we will refer to nouns within the tables as *descriptors*.

C. Statistics-based Identification of Relevant Descriptors

Next, we perform a statistical analysis of the co-occurrence tables. First, we aggregate the descriptor counts for all symptoms belonging to the same clinical sign. Next, we extract those descriptors that are highly correlated with individual clinical signs. The assumption is that if these descriptors can be identified in a medical case description, they will be likely to point to the correlated clinical sign.

V. Preliminary Results & Evaluation

In order to gain a better feeling for the feasibility of an automatic detection, we systematically analyze both the information contained in the orphadata and that extracted from Wikipedia. We perform different statistical analyses in order to learn more about potential issues and opportunities.

A. Knowledge Extraction from Orphadata

The listing of orphan diseases and their clinical signs builds the backbone of our approach. The nature of this data may therefore impose limitations on the overall proceeding. In a first step, we have to examine and evaluate this resource.

An orphan disease contained in orphadata has between as few as one and as many as 180 different clinical signs. On average an orphan disease contains 19.5 clinical signs, with a standard deviation of 15.7. Yet, the distribution is somewhat skewed and the peak is with 9 clinical signs per disease, see Figure 1. At the lower end the data sparsity may be an issue for the identification of orphan diseases: 32 orphan diseases have only one clinical sign each, and 58 have only two signs each. It is questionable whether the automated detection is feasible for these cases as it will be based on a lightweight evidence. Orphan diseases at the other side of the range, the ones having a plethora of clinical signs, are challenging for the analysis, too. Yet, there is a chance to narrow the list of clinical signs down to the most indicative ones. The disease with most clinical signs has 180 of them, out of which 48 are classified as *very frequent*, 50 as *frequent*, and 82 as *occasional* only. It can be observed as a general tendency that diseases with more signs tend to have a disproportionately high amount of *occasional* signs.

A clinical sign contained in orphadata points to as few as one and as many as 987 different orphan diseases. On average a clinical signs points to 41.2 different orphan diseases, with a standard deviation of 71.5. Again, we are confronted with a quite skewed distribution with a peak at two different diseases per sign, see Figure 2. At the lower end of the scale, we find clinical signs that are quite specific and indicative for certain orphan diseases. For example, 66 signs point to exactly one disease each, and 73 signs point two different diseases each. The discriminatory power of these clinical signs within the set of orphan diseases can be considered to be quite high.

 ISBN: 978-1-61208-423-7

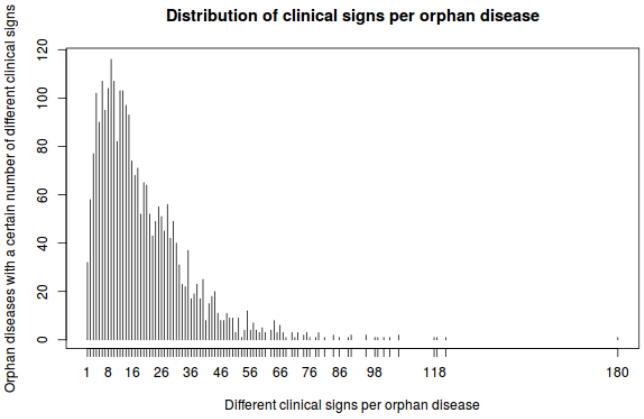


Figure 1. Skewed distribution: some orphan diseases have far more different clinical signs than others.

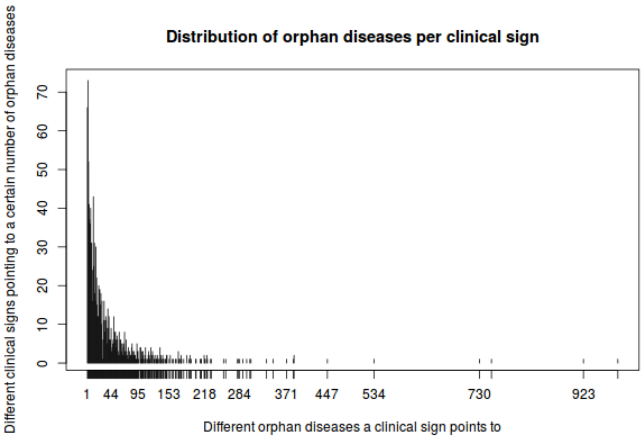


Figure 2. Skewed distribution: some clinical signs point to far more different orphan diseases than others.

At the other end of the range, we find widely spread signs like the one named "Intellectual deficit/mental/psychomotor retardation/learning disability", which points to 987 different diseases. When omitting orphan diseases for which clinical signs are *occasional* only, the tendency remains the same. The most widely spread sign occurs for 876 different diseases *frequently* or *very frequently*. Still, those clinical signs pointing to a wide range of different diseases may be quite useful for the higher-level classification whether a patient might suffer from an orphan disease or not. For a distinction within the set of orphan diseases, more specific, hardly spread signs are useful.

With $n = 2689$ different orphan diseases we can make $(n * (n-1))/2 = 3,614,016$ pairwise comparisons. In particular, we can determine for each pair of diseases in how many clinical signs they coincide and in how many they are distinct. Figure 3 provides a visual summary of performing all pairwise comparisons. The strong left shift of the resulting distribution clearly shows that for almost all of these pairwise comparisons, the corresponding orphan diseases are quite distinct: they share only very few clinical signs (x-axis) while there are many clinical signs in which they can be distinguished (y-axis). This shows that most orphan diseases have quite unique combinations of signs, which makes an automated distinction become very realistic.

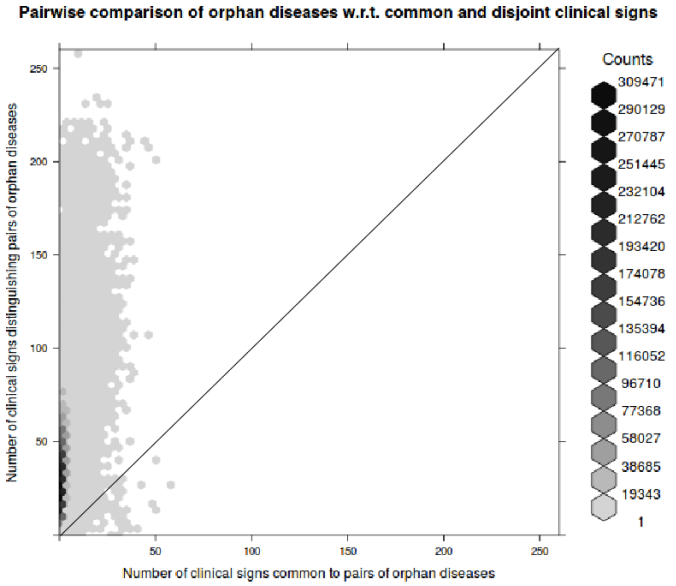


Figure 3. The lion's share of the data is located to the upper left of the diagonal, i.e., almost all pairs of diseases differ in more clinical signs than they have in common.

B. Knowledge Extraction from Wikipedia

The 1362 clinical signs from orphadata contain more than 2500 symptoms. Currently, for 37.2% of the symptoms Wikipedia articles exist, for 15,8% of the symptoms related articles can be retrieved, and for the remaining 47% no articles are found. Overall, information from 2479 Wikipedia pages was gathered. Yet, as described, for a considerable number of symptoms no information was available. This implies that for 423 clinical signs, roughly one third of all signs, we do not dispose of descriptive contexts.

From the available contexts, descriptors were extracted as described in Section IV. When investigating the extracted descriptors, it becomes evident that they are typically very useful and can be divided into 8 categories. For a better illustration, Table I provides the top 50 correlations to clinical signs together with the categories they fall in. The categories and their relative frequencies within the top 50 descriptors:

a) SYNM (28%): Synonyms, e.g., *Lichtscheu* for *Photophobie*. As in this case, often one synonym has a German origin, while the other is of Latin or Greek provenance.

b) ORTH (6%): Orthographical variations, e.g., *Hydrocephalus* for *Hydrozephalus*.

c) ABBR (4%): Abbreviations, e.g., *AVSD* for *atrioventrikulärer Kanal*.

d) DISE (8%): Diseases that relate to the given symptom, e.g., *KBG-Syndrom* for *EEG-Anomalien*.

e) THRP (2%): Terms indicating a common therapy for the given clinical sign.

 ISBN: 978-1-61208-423-7

f) GNRL (10%): Hypernyms or otherwise more general terms with and without morphological relation, e.g., *Volvulus* for *Magenvolvulus*, or *Gliedergürteldystrophien* for *Klauenzehe/Beugekontrakturen der Zehen*.

g) RELA (34%): Clearly related terms that do not fall into any of the other previous categories, e.g., *Fusionsgene* for *interstitielle Deletion/subtelomere Mikrodeletion*.

h) ERRD (8%): Errors typically due to wrong data. In all of the investigated cases, we could trace them back to the erroneous retrieval of an unrelated article in Wikipedia.

It can be concluded that a number of descriptors could potentially have been discovered by other means as well. For example, spell-checkers could have uncovered orthographical variations. Synonyms and abbreviations could potentially have been retrieved from digital dictionaries. Specialized ontologies could have helped in extracting related diseases and therapies. That makes a total of about 48% of descriptors, for which there is hope to obtain them by alternative means.

For the 10% of more general terms (GNRL), however, it is doubtful whether these could have been gathered from controlled vocabularies. For the 34% of remaining related terms (RELA), finally, there is basically no other way than learning them from data. Our approach does a very good job with respect to this and on top it extracts descriptors from all of the other categories with the same proceeding. The automated method yields whole semantic fields with a precision of more than 90% and provides more than a third more descriptors than available by the most optimistic usage scenario of traditional resources. It is hard to estimate a reasonable recall value, though. Finally, apart from a word lemmatization processing step, our method is language-agnostic and can readily be transferred to other natural languages.

VI. Conclusion & Future Work

This work-in-progress paper lays the foundation for the automated extraction of orphan diseases from medical case descriptions and uncovers a number of challenges and limitations mainly regarding the available data.

For some orphan diseases, orphanet lists only one or two symptoms. In these cases, we have a limited and uncertain foundation for identification. In addition, for one third of all clinical signs we could not retrieve descriptive contexts from Wikipedia. This limits the identification of these signs in medical case descriptions to mere exact matches.

While in some cases, due to nature of the data or the lack of available data, the automated identification of orphan diseases is currently quite challenging, for the majority of orphan diseases we indeed do see a very good chance. Moreover, the used data sources are permanently growing and being improved, which will ease the identification and put it in on more solid ground in the future. Regarding the extraction of descriptors, the first results can be considered as very promising. The precision is above 90%. In the future, we plan to extract more than just nouns as descriptors, namely words with other parts-of-speech, word ngrams and phrases. This will increase the recall further. Finally, we will start experimenting with different ways of incorporating the descriptors for an automated identification of orphan diseases within case descriptions posted to the medical community coliquio.

TABLE I. THE 50 STRONGEST CORRELATIONS BETWEEN CLINICAL SIGNS AND DESCRIPTOR WORDS TOGETHER WITH THE CATEGORY THEY FALL IN.

Clinical Sign (with orphanet id)	Descriptor	Cat.
47800 Kryoglobulinämie	Kryoglobuline	RELA
20360 Trommelschlegelfinger	Trommelschlägelfinger	ORTH
5720 Photophobie	Lichtscheu	SYNM
46560 Kniescheibenverrenkung	Patellaluxation	SYNM
23020 Hypohidrose/...	Anhidrose	SYNM
7150 Blepharophimose/...	Blepharophimose-Syndrom	RELA
21320 Anom. d. unt. Extremitäten/...	Epiphyseodese	THRP
15640 Pectus carinatum	Kielbrust	SYNM
33350 Ausweitung der Bronchien/...	Bronchiektasen	RELA
52480 interstitielle Deletion/...	Fusionsgene	RELA
52540 Chromosomenbrüchigkeit	Nijmegen-Breakage-Syndrom	DISE
41870 Galaktorrhö	Milchfluss	SYNM
43140 EEG-Anomalien	KBG-Syndrom	DISE
3700 Kinngrübchen/Kinnspalte	Grübchen	GNRL
22320 Klauenzehe/...	Gliedergürteldystrophien	GNRL
41750 vorzeitige Pubertät	Pubertas	SYNM
15400 überzählige Mamillen/...	Milchleiste	SYNM
4260 Melanose der Iris/...	Melanosis	ORTH
49680 Vitamin B3/PP-Mangel	Nicotinsäure	SYNM
35270 .../Raynaud-Phänomen/...	Raynaud-Syndrom	SYNM
23330 negatives Nikolski-Zeichen	Pemphigus	DISE
17880 persistierender Urachus/...	Allantois	RELA
23060 Hautdehnungsstreifen/Striae	Dehnungsstreifen	GNRL
27630 Darmverschluss/...	Ileus	SYNM
43200 Gangstörung/auffälliger Gang	Gangbild	RELA
10490 Ankyloglossie/...	Zungenbändchen	RELA
42450 Hydrozephalus	Hydrocephalus	ORTH
35480 Ödem der Beine/...	Frakturen	ERRD
49260 Hyperkalziurie	Nephrokalzinose	RELA
5060 Glaskörpertrübungen/...	Vitrektomie	RELA
23190 chron. Infektion der Haut/...	Ulcus	RELA
54210 Durst	Durstgefühl	RELA
34500 atrioventrikulärer Kanal	AVSD	ABBR
18880 kutanes/amniotisches Band/...	Amniotisches-Band-Syndrom	DISE
12250 überzählige Zähne/Polydontie	Hyperdontie	SYNM
49140 Hypokaliämie	Kalium	RELA
26420 Magenvolvulus	Volvulus	GNRL
2600 Kopfhaut/Schädeldefekt	Skalp	SYNM
2600 Kopfhaut/Schädeldefekt	Kopfschwarte	SYNM
2600 Kopfhaut/Schädeldefekt	Skalpieren	ERRD
41150 Kropf	Kropfmilch	ERRD
44450 Anomale Muskel-Biopsie/Muskelenzyme/CPK/LDH/...	Enzym	GNRL
same as above	Blutentnahme	RELA
same as above	Röhrchen	ERRD
21680 Knickfuß	Knöchel	RELA
21280 tarsale Anomalie/Fusion/...	Verschmelzung	SYNM
23500 Pigmentanomalien der Haut	FA-Patienten	RELA
50900 vaskulärer Tumor	EHE	ABBR
44500 Fasziitis	Faszien	RELA
24200 Lanugo/Wollhaar	Lanugobehaarung	RELA

Acknowledgment

This work has partly been funded by the research project "Visual Analytics of Text Data in Business Applications" at the University of Konstanz. We would like to thank the Vidatics GmbH for providing the software implementations.

References

[1] "coliquio," Available on: https://www.coliquio.de/ [Date accessed: 2015-05-26].

[2] "Orphanet: The Portal for Rare Diseases and Orphan Drugs." Available on: http://www.orpha.net/consor/cgi-bin/index.php [Date accessed: 2015-02-27].

[3] A. Rath, A. Olry, F. Dhombres, M. M. Brandt, B. Urbero, and S. Ayme, "Representation of rare diseases in health information systems: the orphanet approach to serve a wide range of end users," *Human mutation*, vol. 33, no. 5, pp. 803–808, 2012.

 ISBN: 978-1-61208-423-7

[4] "Wikipedia Portal Medicine (German Version)." Available on: https://de.wikipedia.org/wiki/Portal%3AMedizin [Date accessed: 2015-02-04].

[5] S. Uryasev and P. M. Pardalos, *Stochastic optimization*. Springer Science & Business Media, 2001, vol. 54.

[6] F. Wanner, A. Stoffel, D. Jäckle, B. C. Kwon, A. Weiler, and D. A. Keim, "State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams," in *EuroVis - STARs*, R. Borgo, R. Maciejewski, and I. Viola, Eds. Swansea, UK: Eurographics Association, 2014, pp. 125–139.

[7] "EURORDIS, Rare Diseases Europe: What is a rare disease?" Available on: http://www.eurordis.org/content/what-rare-diseas [Date accessed: 2015-03-03].

[8] "EURORDIS, Rare Diseases Europe: About Rare Diseases," Available on: http://www.eurordis.org/about-rare-diseases [Date accessed: 2015-03-03].

[9] K. W. Fung, R. Richesson, and O. Bodenreider, "Coverage of rare disease names in standard terminologies and implications for patients, providers, and research," in *AMIA Annu Symp Proc*, vol. 564, 2014, p. 570.

[10] A. Rath, B. Bellet, A. Olry, C. Gonthier, and S. Aymé, "How to code rare diseases with international terminologies?" *Orphanet Journal of Rare Diseases*, vol. 9, no. Suppl 1, p. O11, 2014.

[11] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks, "Approaches to text mining for clinical medical records," in *Proceedings of the 2006 ACM symposium on Applied computing*. ACM, 2006, pp. 235–239.

[12] R. Koeling, J. Carroll, A. R. Tate, and A. Nicholson, "Annotating a corpus of clinical text records for learning to recognize symptoms automatically," in *Proceedings of the 3rd Louhi Workshop on Text and Data Mining of Health Documents*, 2011, pp. 43–50.

[13] S. Köhler *et al.*, "Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research." *F1000Research*, vol. 2, Feb. 2013. [Online]. Available: http://dx.doi.org/10.12688/f1000research.2-30.v1

[14] "Orphadata: Free access data from Orphanet. © INSERM 1997." Available on: http://www.orphadata.org/ [Date accessed: 2015-02-04].

 ISBN: 978-1-61208-423-7

How can Plan Ceibal Land into the Age of Big Data?

Martina Bailón*, Mauro Carballo*, Cristóbal Cobo†, Soledad Magnone*,
Cecilia Marconi*, Matías Mateu* and Hernán Susunday*
* Plan Ceibal
†Fundación Ceibal
Montevideo, Uruguay
Emails:{mbailon,mcarballo,ccobo,smagnone,cmarconi,mmateu,hsusunday}@ceibal.edu.uy

***Abstract*—In 2007, Plan Ceibal became the first nationwide ubiquitous educational computer program in the world based on the 1:1 model. It is one of the most important programs implemented by Uruguay's Government to minimize digital divide and is based upon three pillars: equity, learning and technology. As of 2007, Plan Ceibal has covered all public schools, providing every student and teacher in kindergarten, primary and middle school with a laptop or tablet and internet access in the school. To date, Plan Ceibal has close to 700,000 beneficiaries, each with their own device. Since 2011, the Plan has focused on providing the learning community with a wide range of digital content to enhance the teaching and learning process, most notably Learning Management Systems, Mathematics Adaptive Platform, remote English teaching and an online library. Today, Plan Ceibal operates and integrates a large scale of databases fed by a number of management and educational activities. This abundance of data presents a great challenge and a large opportunity to exploit and transform mass data into rich information. The main goal of this article is to describe the most relevant data sources and present an ongoing data analysis research grounded by a case study. In addition, this paper suggests next steps required to implement a learning analytics strategy within Plan Ceibal. If well exploited, this evidence based data can be used to support and improve the current technology and learning educational policies.**

Keywords–Plan Ceibal; Big Data; Learning Analytics.

I. INTRODUCTION

The amount of data in the world has exploded; the analysis of large data sets is expected to become a new platform for new business, underpinning new waves of productivity growth, innovation, and consumer surplus [1]. This rapidly increasing amount of information, due to the expansion of social computing and the Internet has been coined as Big Data, and this time period described as the age of Big Data [2]. It is expected that Big Data will have a significant impact in the field of education, the design of curricula and the study of general patterns of teaching and learning activities [3][4].

Learning analytics is established as a proper field of knowledge [5]. The use of predictive modeling and other advanced analytic techniques help target instructional, curricular and support resources, to enhance the achievement of specific learning goals [6]. In 2006, the Government of Uruguay deployed Plan Ceibal as its main strategy to introduce Information and Communication Technologies on a large scale across the entire education system [7][8]. Plan Ceibal has successfully provided a device to all teachers and students in public schools from kindergarten to middle school, as well as internet access in all educational centers and outdoor areas. As stated by Fullan et al. [9], the delivery of computers (in some cases tablets) and internet connection represent the first stage of Plan Ceibal. From 2011, a second stage arrived with the introduction of a strong initiative to deliver digital content to teachers and students. The Plan has focused on providing the learning community with digital content to enhance and personalize the teaching and learning processes. More specifically Learning Management Systems, Mathematics Adaptive Platform, remote English teaching, online library with digital and media content, amongst many others. As of 2013, the third stage of Plan Ceibal is being overwhelmed by sustainability and quality aspects of its large-scale development and set of platforms, resources and services for educational community.

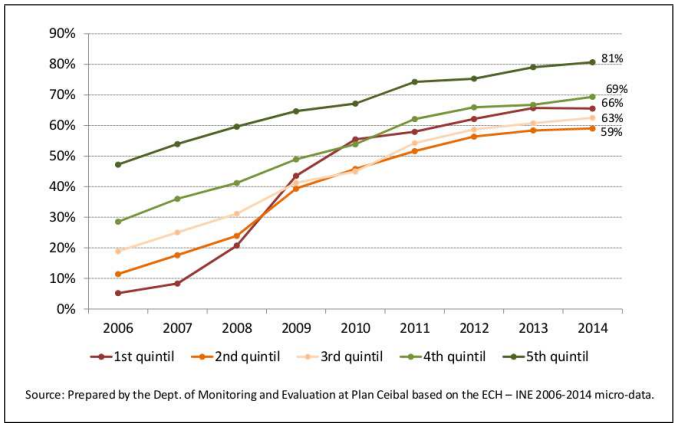


Figure 1. Personal computer access by quintile of per capita income. Whole country, percentage of households. Source: Monitoring and Evaluation Department, Plan Ceibal

The distribution of over 700,000 laptops and tablets, as well as the deployment of internet connections for schools and communities, facilitated higher levels of social inclusion. Moreover, it also provided equity via reduction of the 'digital gap' between the Information and Communication Technologies "haves" and "have nots" in all socio-economic contexts [10]; see Figure 1.

The decreased digital gap has enabled Plan Ceibal to focus on improving the quality of education by integrating technology in classrooms, schools and student's households.

Plan Ceibal is currently pursuing a consistent strategy for the improved implementation of the Data Exploitation stage, outlined in the value chain model of Data Analytics in Figure 2. This stage is one of many Plan Ceibal is simultaneously working on.

This paper provides an initial exploratory review to understand the data sources available in the existing databases

 ISBN: 978-1-61208-423-7

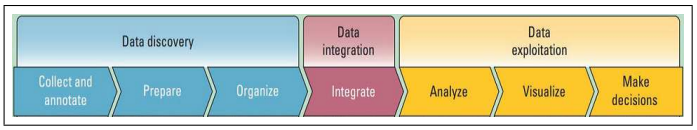


Figure 2. Value chain model of Data Analytics

in Section 2. In Section 3 the key questions to integrate and exploit these sources are presented. In Section 4 a case study that correlates the use of Adaptive Maths Platform (PAM) with infrastructure performance as well as social and demographical variables is described. Section 5 presents some preliminary results and Section 6 discusses the future associated to Learning Analytics and the Plan.

II. DATA SOURCES

Plan Ceibal is currently addressing many obstacles related to Big Data. Today, Plan Ceibal manages national scale databases that are necessary to pursue a number of management (deployment and management of devices and wifi network) and educational activities (teachers and students use of the tablet, laptops, platforms, websites, contents, etc.). The matrix in Table I shows some of the most important sources of information available as well as the volume, variety and frequency of data generation:

TABLE I. MATRIX OF THE MAIN DATA SOURCES

Source	Dimension	Frequency	Unit	Size
CRM	Socio-demographic Features	Annual	User ID	260.000 Primary students, 130.000 Secondary students
	Human and Physical Infrastructure	Annual	School ID	700.000 beneficiaries, 3000 endpoints
	Support Service	Hourly	User ID	3680 tickets a day
PAM, Adaptive Maths Platform	Performance	Daily	User ID	95.000 users in 2015
Tracker System Statistical Monitoring	Computer Usage	Daily	User ID	50 schools with 5000 users
Zabbix Monitoring System	IT Infrastructure Performance	Hourly	School ID	building 3000 facilities
CREA 2 LMS	Performance	Hourly	User ID	85.000 users in 2015

For instance, the *Tracker System for Statistically Monitoring Computer Usage* enables us to know what the most popular applications and activities are among students, how much time they spent on the computers, at what time of the day, and drive behavior patterns throughout the weeks. From this specific data source, we know that the main student activity is web browsing, followed by video camera applications, followed by drawing application is associated to 83% of students involved in Plan Ceibal's initiative between 6-11 years of age, in the primary school system.

It is worth mentioning that in the customer relationship management (CRM) database, Plan Ceibal records the devices owned per user. The PAM database provides information regarding the intensity of its usage by students and teachers. Zabbix database is the School's Network monitoring system and CREA 2 is the learning management system supported by Ceibal in schools based on a social network philosophy.

The different data sources are vital to understand the impact and use of the laptops on the national scale[11].

Regarding the size of generated data, Table II shows the amount of daily information generated in some of the components of the matrix.

TABLE II. SIZE OF DAILY GENERATED DATA

Source	Size (Mega Bytes)
Zabbix	200
CRM	4
Tracker	6
PAM activity	10
Internet activity	150
Total	370

Regarding the structure of the different data sources, some of the most relevant issues today are:

- Lack of Integration: there is very limited interoperability among databases.
- Lack of a common processing and visualization framework: a unique interface for processing and visualization is necesary.
- Lack of traceability: the bulk is generated from different databases, providers, interfaces and the unit of analysis depends on the database (school, device, end user or classroom based).

Most of the variables and data generated through these databases are collected and processed for managerial and operational needs. Simultaneously, a business intelligence (BI) system is being built on top of all databases for management purposes.

III. KEY QUESTIONS

What are the key parameters, significant variables and required data sources to include in the "data integration" and "data exploitation" stages? These questions can be described as:

- How can we improve integration of the different data sources in a more comprehensive and meaningful way?
- How to enable interoperability and consistency between information and variables retrieved from different data sources (i.e: the unit of analysis in some cases are schools, classrooms or individual based information)?
- What are the more reliable analytical techniques to identify strong correlations amongst key variables?
- How can the integration of the different data sources be applied to better understand ways of improving institutional and pedagogical strategies?

 ISBN: 978-1-61208-423-7

IV. CASE STUDY

A. Motivation

The primary objective of Plan Ceibal is the use of the technological development to boost pedagogic tools. In 2012, Ceibal took a step forward by acquiring an adaptive platform for mathematics (in Spanish PAM) that allowed the beneficiaries to exercise and learn algebra, geometry, etc. That platform allowed teachers to monitor and control the classroom's performance in PAM.

To improve the conditions for the use of PAM, Plan Ceibal made a strategic decision towards deploying an appropriate connectivity network and the acquisition of more modern devices. In order to provide students with high performance experiences via the new equipments a High Performance Network (HPN) is being deployed since 2014 in every urban school.

B. Methodology

The first hypothesis driving the investigation is:

- *The more powerful network infrastructure will facilitate a higher amount of exercises completed by students in PAM.*

The second hypothesis is:

- *The social-demographic features (metropolitan vs. interior urban, and socio-cultural context) affect the use of PAM.*

C. Research Questions

1) To what extent does network performance correlate with PAM use? And what are the most reliable techniques and data sources to explore this correlation?
2) To what extent does the social-demographic features mediate and moderate the relationships between High Performance Networks and PAM intensity of use?

D. Universe and sample

The universe are schools and students belonging to those schools that had its High Performance Newtork deployed during 2014. That represents 100 schools with 13800 students from 4th to 6th level.

From the schools a **random** and **stratified** by socio-demographic context sample is taken: **18 schools** with **3,823 students**.

The time frame will be data collected in PAM for the 2014 school year.

V. PRELIMINARY RESULTS

Regarding first hypothesis, an increase of 35.6% active PAM users has been detected between t_0 and t_1, being t_0 the period of time in 2014 the school had its initial wifi/internet access and t_1 the period of 2014 in which the school had its HPN installed. This is illustrated in Table III

Regarding the second hypothesis, two analyses have been made:

1) Comparison of average of usage between t_0 and t_1 by **location**, i.e, interior urban (IU) vs. Montevideo metropolitan Area (MVD) shown in Figure 3.

TABLE III. NUMBER OF PAM ACTIVE USERS BEFORE AND AFTER HPN DEPLOYMENT.

	Before HPN	After HPN
# PAM active users	806	1093
# activities	53179	67523

2) Comparison of average of usage between t_0 and t_1 by **socio-cultural context**, i.e, favorable vs. very unfavorable contexts(data not shown).

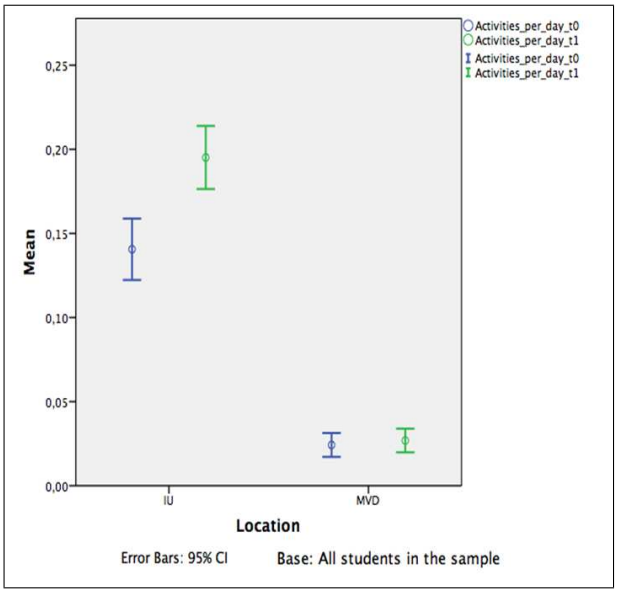


Figure 3. Intensity of PAM use vs. location

The comparison by location highlights that interior urban usage is one order of magnitude bigger than metropolitan use. On the other hand, there is a significant difference between t_0 and t_1 in the urban areas of the provinces whereas in Montevideo there is no significant variation. This can be seen in Figure 3.

The comparison by social-cultural context demonstrates first that unfavorable contexts have approximately one order of magnitude lower intensity of use. Secondly, when compared t_0 and t_1 there is statistical significance in unfavorable contexts only.

VI. DISCUSSION AND FURTHER RESEARCH

From a general perspective, there is a clear need for a more comprehensive information cartography of all the available information resources, as well as a better understanding of how the integration of different data sets can help Plan Ceibal to create better learning experiences, services, tools and public goods specialized in education. The main questions that arise for further research are:

- How to recognize the critical indicators that are more strongly correlated with student's performance?
- How to improve the data traceability of beneficiaries, as well as a better understanding of how the different technologies are being used?

 ISBN: 978-1-61208-423-7

- Will these new forms of tracking students behavior transform how we study the relationship between learning and human-computer interaction, or narrow the palette of research options and alter what '#edtech research' means?
- Will the deployment of these analytic techniques usher in a new wave of privacy incursions and invasive research initiatives? [12]
- How to provide channels and platforms that provide effective feedback for the beneficiaries, i.e., teachers, students, parents. etc.?
- What are the possible strategies to transit from holistic and statistical approaches, like location and socio-cultural context, to a student's level approach?

Given the rise of Big Data as a socio-technical phenomenon, we argue that it is necessary to reflect on these matters and consolidate good practices in the emerging field of learning analytics in order to monitor the influence of technologies in the education ecosystem.

The ongoing research shows some preliminary results about correlation between High Network Performance and PAM intensity of use as well as correlation to independent variables like location and socio-cultural contexts.

The next steps will delve deeper into the challenges faced by Plan Ceibal during it's process of transforming and exploiting the vast amounts of data generated by the organization. We highlight once again, the development of technical and institutional capabilities will allow Plan Ceibal to provide an improved framework towards learning analytics, evidence-based decision making and personalized learning.

ACKNOWLEDGMENTS

The authors would like to thank Daniel Castelo, Claudia Brovetto, Leonardo Castelluccio, Fiorella Haim and Juan Pablo González for their valuable comments. They also want to thank Plan Ceibal for the support of this initiative.

REFERENCES

[1] J. Manyika et al., "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, Tech. Rep., 2011.

[2] S. Lohr, The age of big data. New York Times., 2012, no. 11.

[3] Sharples, M. et al., "Learning design informed by analytics," Open University, Tech. Rep., 2014.

[4] "7 Cool Things to Help You Understand Big Data for Education," 2015, URL: http://tabletsforschools.org.uk/7-cool-things-to-help-you-understand-big-data-for-education/ [accessed: 2015-06-04].

[5] R. Ferguson, "Learning analytics: drivers, developments and challenges," International Journal of Technology Enhanced Learning, vol. 4(5), 2012, pp. 304–317.

[6] C. Bach, Learning analytics: Targeting instruction, curricula and student support. Office of the Provost, Drexel University., 2010.

[7] Plan Ceibal and ANEP, Plan Ceibal in Uruguay. From Pedagogical reality to an ICT roadmap for the future. UNESCO, 2011.

[8] E. Severin and C. Capota, 1 to 1 Models in Latin America and the Caribe, Panorama and Perspectives. BID, 2011.

[9] M. Fullan, N. Watson, and S. Anderson, Ceibal: Next steps. Michael Fullan Enterprises, 2013.

[10] A. L. Rivoir, Ed., Plan Ceibal and Social Inclusion. Interdisciplinary perspectives. Plan Ceibal and University of the Republic, 2011.

[11] N. S. I. INE, "Uruguay's national census, 2011," 2015, URL: http://www.ine.gub.uy/censos2011/index.html/ [accessed: 2015-06-19].

[12] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," Information, communication and society, vol. 5, no. 15, 2012, pp. 662–679.

 ISBN: 978-1-61208-423-7

Clustering Analysis of Academic Courses Based on LMS Usage Levels and Patterns:

Gaussian Mixture Model, K-Means clustering and Hierarchical clustering

Il-Hyun Jo
Department of Educational Technology
Ewha Womans University
Seoul, Korea
ijo@ewha.ac.kr

Jongwoo Song
Department of Statistics
Ewha Womans University
Seoul, Korea
josong@ewha.ac.kr

Yeonjeong Park
Department of Educational Technology
Ewha Womans University
Seoul, Korea
ypark78@ewha.ac.kr

Hyeyun Lee
Department of Educational Technology
Ewha Womans University
Seoul, Korea
hyeyun521@naver.com

Suyeon Kang
Department of Statistics
Ewha Womans University
Seoul, Korea
korea92721@naver.com

Abstract - **This study tried to find a group of academic courses based on the usage levels and patterns of Learning Management System (LMS) utilized in higher education using clustering techniques. LMS is an essential technology to support virtual learning environment where students have access to the learning materials that their instructors provide, submit the deliverables, and participate in various learning activities involving group projects, discussion forums, quizzes, and Wikis. However, the returns on large investment have not systematically performed in terms of what extent of students and instructors have utilized the system for their teaching and learning. In this study, 2,639 courses opened during 2013 fall semester in a large private university located in South Korea were analyzed with 13 observation variables that represent the characteristics of academic courses. Three clustering methods including Gaussian Mixture Model, K-Means clustering, and Hierarchical clustering contributed to (1) identifying large number of courses that show inactive and no usage of LMS, (2) disclosing the dramatically imbalanced clusters, and (3) identifying several clusters that present different usage patterns of LMS. The results of such *academic analytics* provide meaningful implications for academic leaders and university staff to make strategic decisions on the development of LMS.**

Keywords-Learning Management System; Clustering analysis; Gaussian Mixture Model; K-Means clustering; Hierarchical clustering.

I. INTRODUCTION

Currently, universities are increasing incorporating a Learning Management System (LMS) to support effective teaching and learning [1]. Whether focusing on campus-based learning in higher institute or distance learning, LMS is considered as an essential technology for virtual learning environment on e-learning systems where instructors provide various learning materials such as text, images, URL links and video clips to learners.

A common goal of LMS is to organize and manage different courses within an integrated system [1]. The integrated systems collect each learner's online behavior data in every class. Based on this data educational researchers and practitioners can analyze and interpret learners' status during the semester. University staff or decision makers can leverage such LMS usage trends analytics to derive proper treatment and policies to current learners.

Such a data-driven approach has been attempted in the field of higher education recently with the term of *academic analytics*. It has emerged after the widespread of data mining practices by the influence of business intelligence [2, 3]. This approach has been evaluated as a new tool to respond to increased concerns for accountability in higher education and to develop actionable intelligence to improve student success and learning environment [4]. For example, instructors and academic consultants are better able to understand the learner's learning behavior and performance, even their thoughts based on the rich data. Further, the academic analytics can help more strategic investment and development in a way to fulfill the needs of students and instructors based on the informed analytic results via the pattern-recognition, classification, and prediction algorithms of [5].

The data analytics in education has helped to develop prediction models for academic success of learners based on their behaviors and participation or identifying at-risk students for special guidance from their faculty and advisors [6, 7]. However, the previous applications of analytics have disclosed a further research to apply the elaborated analysis and develop more precise prediction models to prevent the drawbacks from the wrong feedbacks to students [8]. Therefore, as a preliminary research, this study highlights the need of the examinations of current usages and patterns of LMS. Instead of analyzing the individual student level data, the academic course data as a unit of analysis was utilized. We argue that without the thorough analysis on LMS usages and patterns and accurate clustering of the courses, it would not be able to build elaborated prediction models to estimate students' success and failure based on the online behavior records in LMS.

The data sets utilized for academic analytics can be diverse depending on the characteristics of institutions [5]. Not only the aforementioned LMS but also course management system (CMS), audience response systems, library systems are the examples. In this study, we utilized LMS dataset to analyze students' virtual learning behaviors and CMS data to collect the academic course's general information. By using both LMS and CMS data, the

 ISBN: 978-1-61208-423-7

clustering analysis of academic courses on the basis of virtual learning environment usage levels and patterns were synergistically performed. For the rigorousness and thoroughness on data analytics we employed multiple methods of clustering analysis: Gaussian Mixture Model, K-Means clustering and Hierarchical clustering. The specific research questions were as follows:

RQ1) To what extent have instructors and students utilized LMS for their teaching and learning?

RQ2) What clusters are formed as the patterns of LMS usages?

RQ3) How does clustering analysis detect academic courses that present inactive usage of LMS or unique LMS usage patterns?

II. METHODOLOGY

This study aimed to find a group of classes that are as homogenous as possible within group (cluster) and as inhomogeneous as possible between groups (clusters) based on their online activities and class sizes.

A. Research Context

The context of this study was a private university located in Seoul, Korea. With the supports of institution for teaching and learning in the university, we collected academic course data of the year of 2013 fall semester. All courses were opened using Moodle-based virtual learning environment regardless of the course type such as offline and online. Consequently, total 4,416 courses were analyzed at the initial data analysis step. However, since it was revealed that many courses did not use online campus, the exclusion of such non-active courses were performed. Finally, 2,639 courses were observed for this study with 13 variables.

A data set for analysis was prepared by combining two databases: CMS and LMS. CMS dataset contained course-related information indicating each student's hierarchical categorizations (i.e., graduate VS. undergraduate, mandatory VS. selective, affiliated colleges and department) and LMS dataset included online behavior tracks (i.e., total number of resources, notices, lecture notes, submissions, group works etc.). We integrated CMS and LMS dataset, and these data were divided in general indicator and activity-based indicator. Table I shows a total of 13 variables.

TABLE I. VARIABLE SUMMARY

	No.	Variable name	Variable explanation
General Indicator	1	**MEM**	Number of members
	2	**FRE**	Average log-in frequency per person
	3	**ACT**	Number of activity items
Activity-based Indicator	4	**RES**	Number of resources
	5	**NOT**	Number of notice
	6	**QNA**	Number of questions and answers
	7	**LEC**	Number of lecture notes
	8	**SUB**	Number of task submissions
	9	**GRO**	Number of group works
	10	**LIN**	Number of links
	11	**POS**	Number of discussion forum postings
	12	**QUI**	Number of Quiz
	13	**WIK**	Number of Wikis

B. Clustering Methods

1) Gaussian Mixture Model

GMM is a probabilistic model that assumes all data are from the mixture of normal distributions. The variables must be numeric since we assume that the data are from the multivariate normal distribution. The parameters (the proportion of each group, mean vectors, and variance matrix) are estimated by EM algorithm. In general, the number of clusters is very hard to estimate in the clustering analysis. However, we can estimate the optimal number of clusters in GMM using the Bayesian Information Criterion (BIC). We use the R-package "mclust" for GMM. The mclust package in R can estimate not only the number of clusters but also the optimal form of variance matrix. We will use the number of clusters from the GMM for the K-means and the hierarchical clustering, too.

2) K-means clustering

K-means clustering is one of the most popular clustering method because it is very fast to find clusters and very easy to understand. The objective function of K-means clustering is to minimize the sum of within scatters. Basically, it tries to find the k group that minimizes within-cluster sum of squares; therefore it maximizes the between-cluster sum of squares. Since it uses the squared Euclidean distances among the objects and the cluster centers are defined as the means of objects in each cluster, all variables must be numeric. We use K-means function in R for our analysis.

3) Hierarchical clustering

Hierarchical clustering method is used for building a hierarchy of clusters from data. Strategies for this clustering fall into two types: agglomerative for "bottom-up" approach and divisive for "top-down" approach. We use a bottom-up approach in this article. The algorithm finds the nearest two objects and merges them. It repeats this process until all objects are in one cluster. The final results are usually represented by the dendrogram. The hierarchical clustering methods can give different results depending on which distance metric we use between groups. There are several distance metrics between groups and we use the "complete-linkage" in our analysis. The "complete-linkage" is the maximum distance between two groups and it is known that the "complete-linkage" can find the compact clusters. We use "hclust" function in R for our analysis.

 ISBN: 978-1-61208-423-7

III. RESULTS

A. Descriptive Statistics

Before going to the clustering analysis, we examined descriptive statistics to find out the distribution of observations.

TABLE II. DESCRIPTIVE STATISTICS OF 2,639 COURSES

Name	Min	Max	Mean	SD	Skewness	Kurtosis
MEM	2	301	33.22	33.66	2.97	13.00
FRE	2	375	39.75	33.01	2.50	11.05
ACT	1	8	2.49	1.30	0.93	0.78
RES	0	**596**	11.87	21.49	12.22	263.56
NOT	0	132	6.64	9.26	3.21	20.15
QNA	0	**280**	2.95	14.25	12.09	183.95
LEC	0	176	3.74	9.69	5.16	51.87
SUB	0	36	0.95	2.82	4.97	32.73
GRO	0	**1612**	17.52	88.42	8.23	91.71
LIN	0	72	0.32	2.57	14.97	312.54
POS	0	**2810**	6.45	75.32	24.44	788.77
QUI	0	**215**	0.61	8.34	17.82	366.00
WIK	0	15	0.01	0.31	42.92	2005.64

As shown in Table 2, most variables have extremely high values. For example, the maximum values of variables indicate that 596 resources (RES), 176 lecture notes (LEC), 1612 board postings of group works (GRO), 2,810 discussion postings (POS), and 215 Q&A postings (QNA). These values present extremely high utilization level of few courses.

On closer inspection, one course which posted 2,810 forum discussion postings was big-sized basic requirement course and there were over a hundred students who signed up for class. There were 11 groups and they discussed enthusiastically with each other, so such very high postings were possible. Next, the other course which had 1,612 group works was the major course of educational technology and the instructor assigned team project during the semester. There were 10 groups and they used group board for team-based learning. Because they uploaded all the related materials for project, opinions and chatting messages in group board, so this high value was also possible. These cases looked as errors but it tells the 'real aspects' of unique courses.

Furthermore, the data were sparse by showing many observations with zero values. The variables from QNA to WIK have zero values for more than 50% of data. We can predict that there will be a single one big cluster with a lot of 'zero' observations. This one big cluster will have all the classes with minimal online activities. This cluster was not our interest but we were more interested in other clusters of small size and how different they are.

B. Gaussian Mixture Model

As Figure 1 indicates, Mclust finds the best model is three clusters with EEV (ellipsoidal, equal volume and shape covariance). In the point of three components (clusters), the increase of BIC starts decrease. However, four-cluster model is also close. Thus, we decided to investigate both three-cluster and four-cluster model.

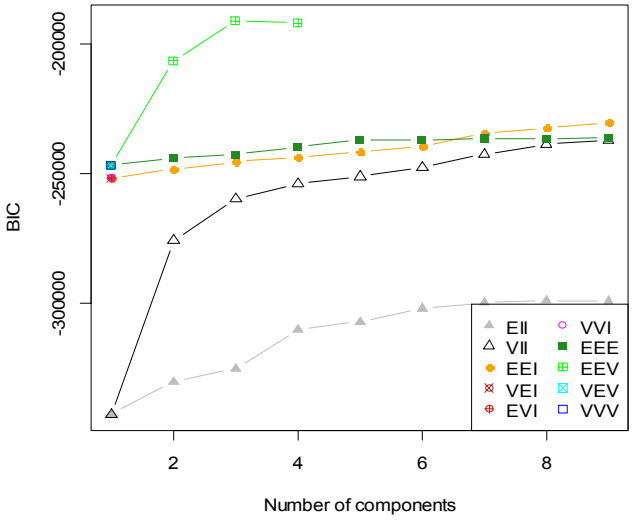


Figure 1.Results of EEV in Mclust

1) GMM with three clusters

The size of three clusters were 212, 2,360, and 67 respectively as seen in Table III.

TABLE III. CLUSTERING TABLE WITH THREE CLUSTERS

	Cluster 1	Cluster 2	Cluster 3
Number of Class	212	2360	67
Mixing Probability	0.08068	0.89393	0.02539

We checked the mean vectors (cluster centers) of three clusters. As Figure 2 indicates, cluster 3 (size 67, green line) has the higher mean values (more online activities) and cluster 1 (size 212, black line) is in the middle, and cluster 2 (size 2,360, red line) has the least online activities. On a closer view, cluster 3 has greatly high value of POS and cluster 1 has high GRO value.

 ISBN: 978-1-61208-423-7

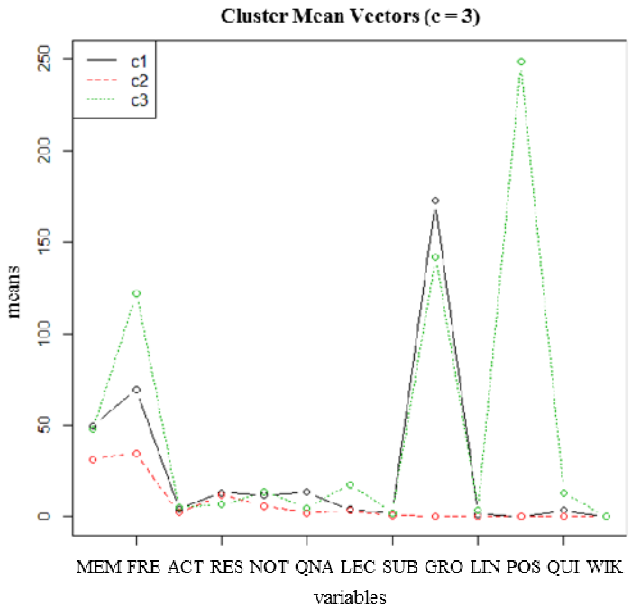


Figure 2. Mean vector plot of three clusters

Look inside the clustering table, 2,360 out of total 2,639 classes were included in cluster 2 where having at least online activities. They were inactive classes. Approximately 89% of total class did marginal performance at online campus. On the other hand, cluster 3 was the most active online classes. We can guess that these classes were actively discussed about their topic since both number of forum discussion postings and average log-in frequency per person are quite high. The rest courses in cluster 1 also participated in group work much but the average frequency mean is in-between cluster 2 and 3. This cluster is specialized in team project.

2) GMM with four clusters

We divided total classes into four clusters this time. The size of four clusters were 71, 2,322, 230, and 16 as sheen in Table IV.

TABLE IV. CLUSTERING TABLE WITH FOUR CLUSTERS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of Class	71	2322	230	16
Mixing Probability	0.02705	0.87962	0.08727	0.00606

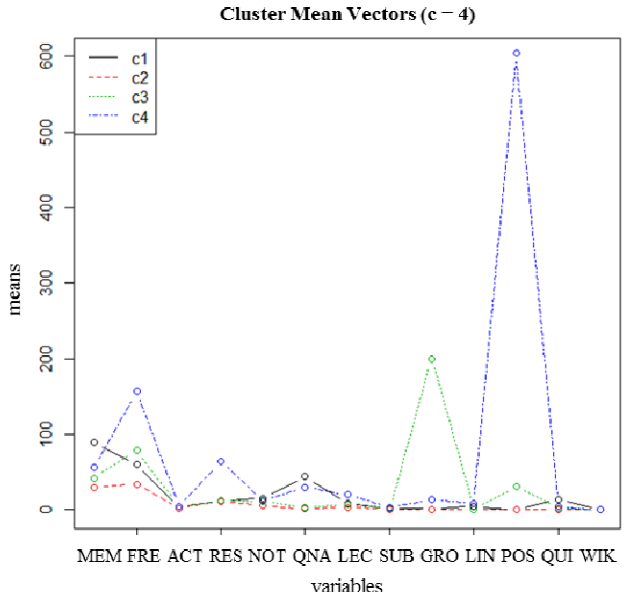


Figure 3. Mean vector plot of four clusters

When reviewing the cluster mean vector plot in Figure 3, cluster 4 (size 16, blue line) has extremely high mean values in POS while cluster 2 (size 2,322, red line) has low mean values in general. Cluster 4 shows the equal appearance with the cluster 3 in GMM with three clustering analysis. Moreover, in common with GMM with three clustering results, 9th variable (GRO) is shown the highest value in cluster 3 (size 230, green line), not the cluster 4 which has higher values in the gross. Courses which involved in cluster 3 were inactive in most of online activities except group works. Newly-drawn cluster 1 (size 71, black line) has the highest MEM value and it represents number of members including an instructor, teaching assistant and students. We are able to call its name, 'big-sized courses'.

The last thing we should observe carefully is that when we clustered total courses into four clusters using GMM, number of courses with highly active in online activities such as forum discussion postings and log-in frequency were decreased from 67 (see Table III) to 16(see Table IV).

C. K-means clustering

In addition to GMM, we also performed a clustering analysis using K-Means. As a first step, we analyzed with non-standardized dataset to see overall clusters and compare the results with GMM. However, due to the large scale differences among variables, we also conducted clustering with standardized dataset because we like to see the clustering results when all variables have the similar contributions in distances.

 ISBN: 978-1-61208-423-7

1) Using non-standardized data

The results of K-means clustering with non-standardized data showed similar results with GMM analysis. But this process was meaningful because the results identified fewer active online courses.

a) K-means clustering with three clusters

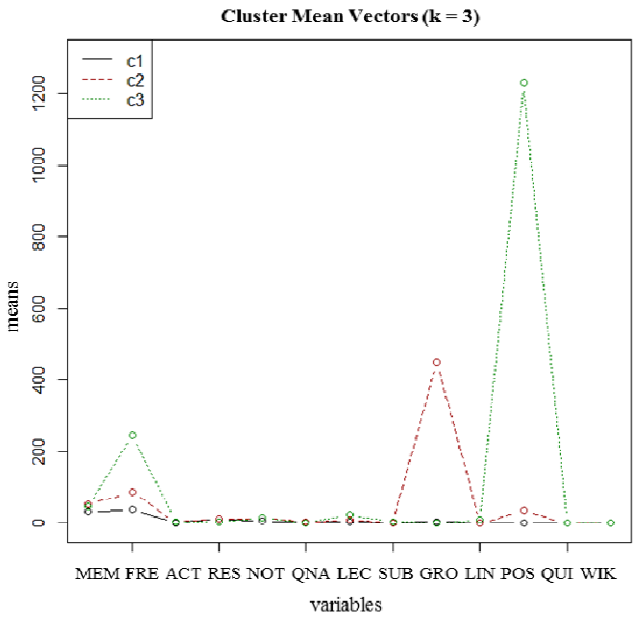


Figure 4. Mean vector plot of three clusters

Most of mean vector values about learners' online behavior were quite similar, similarly low but FRE, GRO, POS variables were distinguished among clusters. Learners who were included in cluster 3 (size 6, green line) classes logged LMS in the most frequently and wrote up the postings on the forum very much. Cluster 2 (size 71, red line) has high value of GRO which means group works. Cluster 1 (size 2,562, black line) which the most of classes were in has less online action.

TABLE V. CLUSTERING TABLE WITH THREE CLUSTERS

	Cluster 1	Cluster 2	Cluster 3
Number of Class	2562	71	6
Mixing Probability	0.97082	0.02690	0.00227

Six courses included in cluster 3 are listed on Table VI. They were super active classes in university. As shown in mean vector plot on Figure 4, these courses have high value of log-in frequency (FRE) and forum discussion postings (POS).

TABLE VI. DETAILED VARIABLE VALUES OF CLUSTER 3 COURSES

No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
255	103	**167**	7	8	71	7	37	3	0	0	**2810**	0	0
894	43	**264**	4	0	0	0	7	14	0	16	**991**	0	0
1299	30	**375**	3	0	0	0	27	0	0	14	**944**	0	0
1403	37	**204**	4	0	0	0	62	1	0	23	**715**	0	0
1630	46	**217**	8	2	22	1	13	12	0	1	**1297**	0	3
2049	18	**245**	4	13	5	1	0	0	0	0	**638**	0	0
M	46	**245**	5	4	16	2	24	5	0	9	**1233**	0	0

b) K-means clustering with four clusters

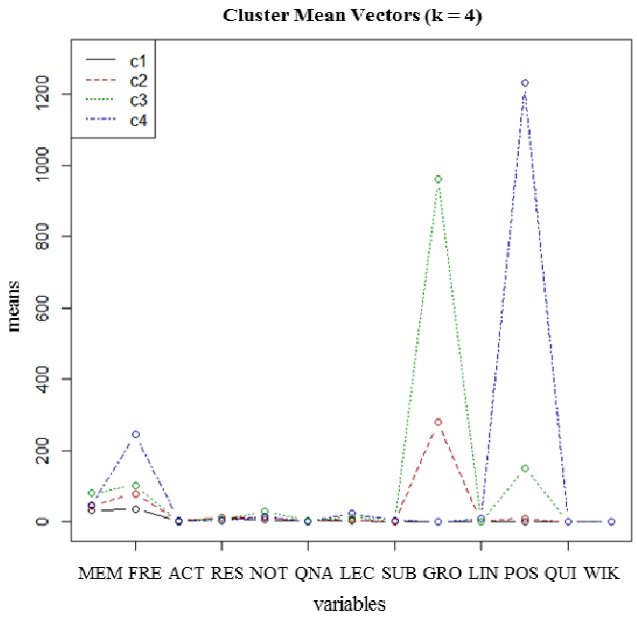


Figure 5. Mean vector plot of four clusters

When we were partitioning total courses into four clusters, cluster 3 and 4 were somewhat unique. Cluster 3 (size 11, green line) has high value of GRO variable and cluster 4 (size 6, blue line) is shown much online action in FRE and POS variables. As mentioned earlier, students in cluster 4 courses discussed with one another constantly and this fact can be proved by FRE and POS. Like the preceding, cluster 3 performed intensive group works. Newly created cluster 2 (size 109) compared to previous results was shown the middle activeness in LMS.

TABLE VII. CLUSTERING TABLE WITH FOUR CLUSTERS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of Class	2513	109	11	6
Mixing Probability	0.95225	0.04130	0.00417	0.00227

Six classes included in cluster 4 are exactly the same courses with the cluster 3 in K-means clustering with three clusters analysis (see Table VI).

 ISBN: 978-1-61208-423-7

2) Using standardized data

Prior to clustering data, we rescaled variables for comparability. So standardized data was utilized in this step. It showed quite different figures in mean vector plots for the plot of non-standardized dataset. Since K-means uses the squared Euclidean distance, the outliers can affect the clustering results significantly. However, if we use the standardized dataset, then the effect of outliers will be reduced, therefore it is unlikely to see very small sized clusters.

a) K-means with three clusters

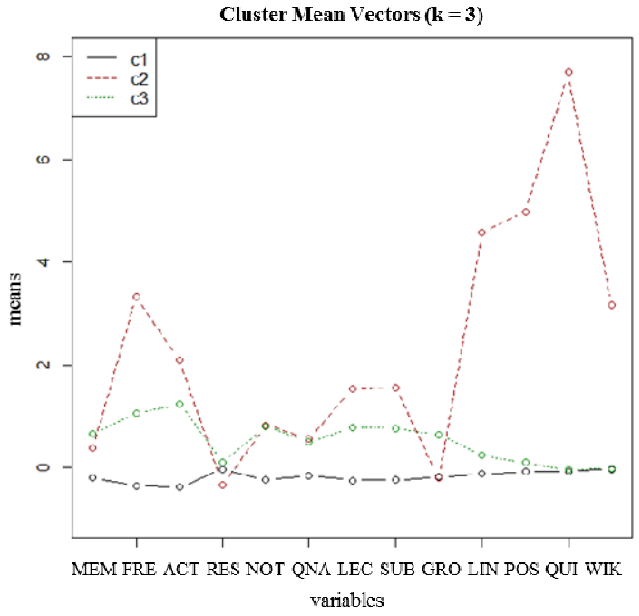


Figure 6. Mean vector plot of three clusters

As shown in Figure 6, cluster 2 (size 22, red line) has high mean vector value on the whole. FRE, ACT, LEC, SUB, LIN, POS, QUI and WIK values of cluster 2 were high. Among these, those courses used quiz function very frequently, so QUI was shown excessive activity log in comparison with other clusters.

TABLE VIII. CLUSTERING TABLE WITH THREE CLUSTERS

	Cluster 1	Cluster 2	Cluster 3
Number of Class	2030	22	587
Mixing Probability	0.76923	0.00834	0.22243

Cluster 1 (size 2,030, black line), about 77% of courses contained, was shown the low activeness in general without exception. However, cluster 2 was generally active. Cluster 3 (size 587, green line) was middle-active according to the LMS usage levels, but it had top-of-the-line value in MEM, RES and GRO. In contrast with non-standardized clustering results, these clusters were distinguished by the level of usage, not the unique extreme values.

b) K-means with four clusters

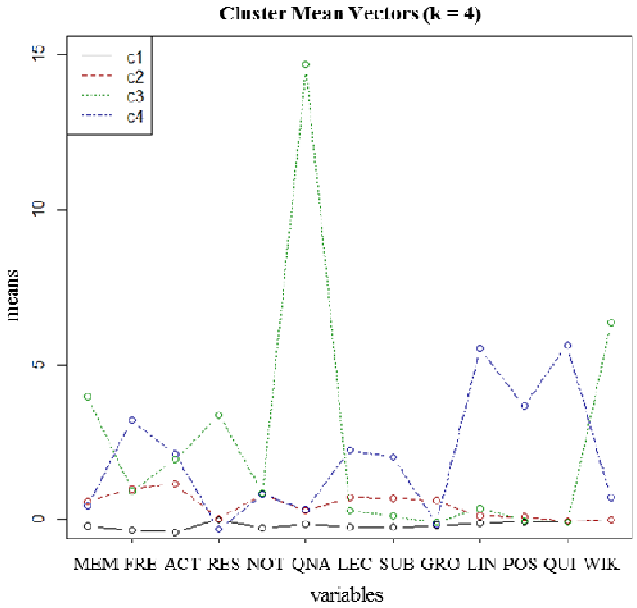


Figure 7. Mean vector plot of four clusters

Cluster 3 (size 8, green line) had unusually high mean vector values in QNA, compared to other clusters. Moreover, such MEM, RES and WIK values were also high. We can interpret this situation that there were many members in class, so lots of questions came out together. Another cluster 4 (size 30, blue line) has high action value in FRE, ACT, LEC, SUB, LIN, POS and QUI. In other words, students eagerly participated in LMS in average since the average log-in frequency per person value was the biggest among other cluster. Furthermore, we could assume that the courses provided both a great deal of course-related materials and the grade-related assignment. High values of SUB (number of task submission) and QUI (number of quiz) as well as LEC (number of lecture notes) and LIN (number of URL links) are the evidences. However, cluster 1's (size 1,979, black line) action was minor despite it took most of virtual learning environment courses. Likewise the previous analytic results of cluster 3 (size 587, green line), cluster 2 (size 622, red line) was shown the middle activeness.

TABLE IX. CLUSTERING TABLE WITH FOUR CLUSTERS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of Class	1979	622	8	30
Mixing Probability	0.74991	0.23570	0.00303	0.01137

 ISBN: 978-1-61208-423-7

D. Hierarchical clustering

Lastly, we analyzed academic courses with hierarchical clustering method. Standardized dataset was used to clustering.

1) Hierarchical clustering with three clusters

TABLE X. CLUSTERING TABLE WITH THREE CLUSTERS

	Cluster 1	Cluster 2	Cluster 3
Number of Class	2637	1	1
Mixing Probability	0.99924	0.00038	0.00038

The result of hierarchical clustering displays an unprecedented appearance. The only 158th class in Table XI came under cluster 2 (size 1) and a 255th class in Table XII was included in cluster 3 (size 1). Except those two certain classes, the rest of courses were clustered together in cluster 1 (size 2,637).

TABLE XI. DETAILED VARIABLE VALUES OF CLUSTER 2 COURSE

Class No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
158	144	93	7	19	8	108	29	0	0	9	30	0	**15**

158th course utilized many activity items (ACT = 7) in moderate way and interestingly used Wiki function in its course. It was the course of economics department. Actually, 15 times was not that huge usage number but as almost the whole courses had not used Wiki (M = .01, SD = .31), this class was chosen for the sole course in cluster 2 because of WIK.

TABLE XII. DETAILED VARIABLE VALUES OF CLUSTER 3 COURSE

Class No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
255	103	167	7	8	71	7	37	3	0	0	**2810**	0	0

255th class represents extremely high value of forum discussion postings. This course also utilized many activity items (ACT = 7) and specifically in POS, it showed unparalleled usage. It was possible because there were lots of members in class. Every person uploaded 27.28 postings averagely and it would be an acceptable number.

2) Hierarchical clustering with four clusters

TABLE XIII. CLUSTERING TABLE WITH FOUR CLUSTERS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of Class	2634	1	1	3
Mixing Probability	0.99811	0.00038	0.00038	0.00114

Four clusters analytic result was pretty similar with those *three* clusters hierarchical clustering. Cluster 2 and 3 courses (158th and 255th class) were the same with the previous result. However, newly created cluster 4 (size 3) differed from the previous one. Three classes out of 2,637 courses had high mean value in RES (number of resources).

TABLE XIV. DETAILED VARIABLE VALUES OF CLUSTER 4 COURSES

Class No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
514	46	49	2	**596**	0	227	0	0	0	0	0	0	0
1151	84	58	4	**276**	44	19	0	1	0	0	0	0	0
1557	52	83	4	**401**	5	1	29	0	0	0	0	0	0
Mean	60.67	63.33	3.33	**424.33**	16.33	82.33	9.67	0.33	0.00	0.00	0.00	0.00	0.00

These three courses did not utilized many activities so the variables from SUB to WIK got almost zero value. Specifically, courses had the highest RES values. We can interpret that instructors in these courses chose the resource application instructional method and provided many useful resources for the subject.

3) Hierarchical clustering with five clusters

TABLE XV. CLUSTERING TABLE WITH FIVE CLUSTERS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Number of Class	2629	5	1	1	3
Mixing Probability	0.99621	0.00189	0.00038	0.00038	0.00114

Cluster 3 (size 1), 4 (size 1) and 5 (size 3) were same with cluster 2, 3 and 4 in hierarchical clustering with *four* clusters results. Cluster 2 (size 5) was broken loose from cluster 1 (size 2,629) and five classes in Table XVI were included.

TABLE XVI. DETAILED VARIABLE VALUES OF CLUSTER 2 COURSES

Class No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
31	183	49	4	0	10	6	22	0	0	**33**	0	0	0
571	103	59	5	12	6	12	0	2	0	**31**	0	0	0
594	101	52	5	21	24	18	0	1	0	**30**	0	0	0
1243	46	163	7	0	8	8	41	5	0	**48**	201	28	0
1694	55	52	4	0	12	0	33	1	0	**72**	0	0	0
Mean	97.6	75.0	5.0	6.6	12.0	8.8	19.2	1.8	0.0	**42.8**	40.2	5.6	0.0

These five classes had actively shared useful URL links during the semester as a course material. Mainly, instructors provided references from the web in big-sized courses.

4) Hierarchical clustering with six clusters

TABLE XVII. CLUSTERING TABLE WITH SIX CLUSTERS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Number of Class	2620	5	9	1	1	3
Mixing Probability	0.99280	0.00189	0.00341	0.00038	0.00038	0.00114

 ISBN: 978-1-61208-423-7

This case, cluster 2 (size 5), 4 (size 1), 5 (size 1) and 6 (size 3) took on an exactly the same aspect with *five* clusters hierarchical clustering. All the courses which were included in each cluster were the same. Some courses in previous cluster 1 (size 2,620) were divided into two clusters in here as cluster 1 and 3 (size 9).

IV. DISCUSSION AND CONCLUSION

The purpose of study was to cluster academic courses in higher education in accordance with virtual learning environment usage levels and patterns. For this goal, we employed three methodologies: Gaussian mixture model, K-Means clustering and hierarchical clustering and could draw several implications.

The results of this study found that clusters were considerably imbalanced. Descriptive statistics revealed that outliers from dataset were so abnormally high in some variables. On the other hand, most of values were quite low and most of them were zero. This initial data condition led to disproportionate result and this would be the reason that some enthusiastic courses continuously came up. It may not be the cluster in which decision makers wanted to see from LMS usage patterns in higher education institute. However, this real combined data and the results emphasize the true status quo. We can infer that instructors do not know much how to use virtual learning environment well and they might have a hard time to facilitate the LMS use. It is time for academic leaders and university decision makers to form a practical plan which can improve utilization in a balanced way.

Nevertheless, this study revealed that certain enthusiastic courses were drawn out repeatedly when using these three clustering methods. Since such courses had unique characteristics distinguishing from other courses, this study suggests a further in-depth study to examine remarkable instructional methods and challenges in aspect of teaching and learning.

Three methodologies (Gaussian Mixture Model, K-means clustering and Hierarchical clustering) for clustering analysis of academics was meaningful respectively. GMM, as an initial step, was essential to check overall clusters of 2,639 academic courses opened during one semester. As classic and the most popular algorithm, K-means with both non-standardized and standardized dataset contributed to identify prototypical LMS usage patterns by revealing clusters of course utilized *forum-based online instruction*, *quiz-based online instruction*, and *wiki-based instruction*. Hierarchical clustering method was also valuable for the detection of extreme outlier courses that revealed *resource-based online instruction*. Because of hierarchical analytic approach, few outlier could not be included in other cluster naturally but it was left in isolation. This study confirmed that the different strengths of three methodologies leveraged to escalate the effectiveness and robustness of clustering analysis.

Finally, this study represented that online learning activity was fairly marginal despite the advance of information and communication technology (ICT) and its applications for promoting blended learning policy in higher education. We conclude that offline courses were central to most of higher education. We found too many courses did not incorporate a variety of activity items. LMS such as Moodle and Blackboard provides lots of meaningful activity opportunity like discussion, group works, quiz and Wiki. Although we consider that there might be some cultural characteristics of university in South Korea, we believe the analytics methods and approaches incorporated in this study contribute to the area of academic analytics in the field of higher education.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2013S1A5A2A03044140).

REFERENCES

[1] C. Dalsgaard, "Social software: E-learning beyond learning management systems," *European Journal of Open, Distance and E-Learning,* vol. 2006, 2006.

[2] P. Baepler and C. J. Murdoch, "Academic analytics and data mining in higher education," *International Journal for the Scholarship of Teaching and Learning,* vol. 4, p. 17, 2010.

[3] P. J. Goldstein and R. N. Katz, *Academic analytics: The uses of management information and technology in higher education*: Educause, 2005.

[4] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, "Academic analytics: A new tool for a new era," *Educause Review,* vol. 42, p. 40, 2007.

[5] K. E. Arnold, "Signals: Applying Academic Analytics," *Educause Quarterly,* vol. 33, p. n1, 2010.

[6] K. E. Arnold and M. D. Pistilli, "Course signals at Purdue: Using learning analytics to increase student success," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 2012, pp. 267-270.

[7] A. Essa and H. Ayad, "Student success system: risk analytics and data visualization using ensembles of predictive models," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 2012, pp. 158-161.

[8] A. Kruse and R. Pongsajapan, "Student-centered learning analytics," *CNDLS Thought Papers,* pp. 1-9, 2012.

 ISBN: 978-1-61208-423-7

Projective Adaptive Resonance Theory Revisited with Applications to Clustering Influence Spread in Online Social Networks

Jianhong Wu
LIAM, Department of Mathematics and Statistics, York University,
Toronto, Canada, M3J1P3.
Email: wujh@yorku.ca

***Abstract*— We revisit the theory and applications of the Projective Adaptive Resonance Theory (PART) neural network architecture for clustering high dimensional data in low dimensional subspaces and nonlinear manifolds. We put a number of PhD theses, research publications and projects of the York University's Laboratory for Industrial and Applied Mathematics (LIAM) in a coherent framework about information processing delay, high dimension data clustering, and nonlinear neural dynamics. The objective is to develop both mathematical foundation and effective techniques/tools for pattern recognition in high dimensional data.**

Keywords-projective clustering; nonlinear dynamics in processing high dimensional data; influence soread in online social network.

I. INTRODUCTION

In a series of studies starting in the papers [1] [2] and the thesis [3], the Laboratory for Industrial and Applied Mathematics (LIAM) at York University (Toronto, Canada) has been developing a comprehensive framework about information processing delay, high dimension data clustering, and nonlinear neural dynamics. The objective is to develop both mathematical foundation and effective techniques/tools for pattern recognition in high dimensional data. Some of the earlier developments have been reported in the monograph [4] and the survey [5]. The survey also provided a heuristic description of the philosophy about how the modern nonlinear dynamic systems theory (invariant manifolds, domain attractions, global convergence, Lyapunov functions etc.) provides some theoretical principles based on recent biological evidences for novel neural network based clustering architectures to speed up information processing to assist decision making. Here, we briefly describe the current status (Section II) and then summarize the interdisciplinary nature (Section III) of a high dimensional data projective-clustering driven academic-industrial collaboration based on nonlinear dynamics and neural networks.

II. CURRENT STATUS

Under the framework "Projective Adaptive Resonance Theory" (PART), we *developed a novel neural network architecture and algorithm to detect low dimensional patterns in a high dimensional data set.* These are the patterns characterized by the so-called projective clusters, in nonlinear subspaces or nonlinear manifolds. PART *has received much attention by data science researcher and end-user community,* and has formed the core data analytics tools of three Collaborative Research Development projects (CRD), funded by the Natural Science and Engineering Research Council of Canada (NSERC). In particular, the NSERC CRD project Enterprize Software for Data Analytics in collaboration with InferSystems Inc. is based on the application of PART to analyzing odd bidding behaviors in a real-time bidding auction; while the project An Online Integrated Health Risk Assessment Tool brings together a team of investigators with expertise for an interdisciplinary and multi-institutional collaboration to develop systematic analyses converging on a single number, modeled after the single-number Credit Score, to inform chronic disease decision making, both at the population and individual levels. These are also part of a newly funded NSERC Collaborative Research and Training Experience Program in Data Analytics & Visualization.

The PART algorithm has since been used in a number of applications. It was used to develop a powerful gene filtering and cancer diagnosis method in [6][7][8], which shows that *"the results have proven that PART was superior for gene screening"*. PART was also used for clustering neural spiking trains [9], ontology construction [10], stock associations [11], and information propagation in online social networks [12][13]. The PART algorithm has also been extended to deal with categorical data in the thesis [14].

The PART architecture is based on the well known ART developed by Carpenter and Grossberg, with a selective output signaling (SOS) mechanism to deal with the inherent sparsity in the full space of the data points in order to focus on dimensions where information can be found. The key feature of the PART network is a hidden layer of neurons which incorporates SOS to calculate the dissimilarity between the output of a given input neuron with the corresponding component of the template (statistical mean) of a candidate cluster neuron and to allow the signal to be transmitted to the cluster neuron only when the similarity measure is sufficiently large. *Recently discovered physiological properties* of the nervous system, the adaptability of transmission time delays and the signal losses that necessarily arises in the presence of transmission delay, enabled us to interpret SOS as a plausible mechanism from the self-organized adaptation of transmission delays driven by the aforementioned dissimilarity. The result is a

 ISBN: 978-1-61208-423-7

novel clustering network, termed PART–D, *with physiological evidence from living neural network and rigorous mathematical proof of exceptional computational performance* [15].

Such an adaptation can be regarded as a consequence of the Hebbian learning law, and the dynamic adaptation can be modeled by a nonlinear differential equation using dissimilarity driven delay in signal processing. This links to the PhD thesis [16], which proposed an alternative neural network formulation of the Fitts' law for the speed-accuracy trade-off of information processing, and its subsequent publications including [17][18][19][20]. When the delay adaption rates are in certain ranges, we observe nonlinear oscillatory behaviors (clustering switching) and this oscillation slows down the convergence of the clustering algorithm. How to detect and prevent these oscillations is the focus of the thesis [21] the studies [22][23].

III. SUMMARY

In summary, there have been increasing physiological evidences to support the idea of projective clustering using neural networks with delay adaption, there has been some theoretical analysis to show why such a network architecture works well for high dimensional data, and there have been sufficient applications to illustrate PART clustering algorithms are efficient. An interdisciplinary approach for high dimensional data clustering clearly shows the potential to develop a dynamical system framework for pattern recognition in high dimensional data.

REFERENCES

[1] Y. Cao & J. Wu, "Projective ART for clustering data sets in high dimensional spaces", Neural Networks, 15, 105-120, 2002.

[2] Y. Cao & J. Wu, "Dynamics of projective resonance theory model: the foundation of PART algorithm", IEEE Trans Neural Networks, 15, 245-260, 2004.

[3] Y. Cao, "Neural networks for clustering: theory, architecture, algorithm, and neural dynamics", York University, 2002.

[4] G. Gan, C. Ma, and J. Wu, Data Clustering: Theory, Algorithms and Applications, ASA-SIAM Series on Statistics and Applied Probability, 466 pages, 2007.

[5] J. Wu, "High dimensional data clustering from a dynamical systems point of view," in Bifurcation Theory and Spatio-Temporal pattern Formation, The Fields Institute Communication, vol. 49, W. Nagata and N. Sri. Namachchivaya eds. Philadophia: American Mathematical Society, vol. 49, pp. 117-150, 2006.

[6] H. Takahashi, T. Kobayashi, and H. Honda, "Construction of robust progonostic predictors by using projective adaptive resonance theory as a gene filtering method", Bioinformatics, 21:179, 2005.

[7] H. Takahashi, Y. Murase, T. Kobayashi, and H. Honda, "New cancer diagnosis modeling using boosting and projective adaptive resonance theory with improved reliable index", Biochemical Engineering Journal, 33:100, 2007.

[8] H. Takahashi, T. Nemoto, T. Yoshida, H. Honda, and T. Hasegawa, "Cancer diagnosis marker extraction for soft tissue sarcomas based on gene expression profiling data by using projective adaptive resonance theory (PART) filtering method", BMC Bioinformatics, 7:1, 2006.

[9] J. Hunter, J. Wu, and J. Milton, "Clustering neural spike trains with transient", Proc.the 47th IEEE Conferenceon Decision and Control, (December 9-11, 2008), pp. 2000-2005. DOI: 10.1109/CDC.2008.47387292008.

[10] R.-C. Chen and C.-H. Chuang, "Automating construction of a domain ontoloogy using a projective adaptive resonance theory neural network and bayesian network", Expert Systems, 25(4):414-430, 2008.

[11] L. Liu, L. Huang, M. Lai, and C. Ma, "Projective ART with bu ers for the high dimensional space clustering and an application to discover stock associations", Neurocomputing, 72:1283-1295, 2009.

[12] M. Freeman, J. McVittie, I. Sivak, and J. Wu, "Viral information propagation in the Digg online social network", Physica A: Statistical Mechanics and its Applications. 415:1, 87-94, 2014.

[13] G. Gan, J. Yin, Y. Wang, and J. Wu, "Complex data clustering: from neural network architecture to theory and applications of nonlinear dynamics of pattern recognition", Proc. 13th International Symposium on Mathematical and Computational Biology (The Fields Institute, November 4-8, 2013), pp. 85-106.

[14] G. Gan, "Subspace clustering for high dimensional categorical data". Master's thesis, York University, Toronto, Canada, October 2003.

[15] J. Wu, H. ZivariPiran, J. Hunter, and J. Milton, "Projective clustering using neural networks with adaptive delay and signal transmission loss", Neural Computation, 23(6):1568-1604, 2011.

[16] D. Beamish, "50 years later: a neurodynamic explanation of Fitts' law", PhD thesis, York University, 2004.

[17] D. Beamish, M. Bhatti, S. MacKenzie, and J. Wu, "Fifty years later: a neurodynamic explanation of Ftts' law", J. Royal Society Interface, 3(10):649-654, 2006.

[18] D. Beamish, S.A. Bhatti, C.S. Chubbs, I.S. MacKenzie, J. Wu, and Z. Jing, "Estimation of psychomotor delay from the Ftts' law coefficients", Biological Cybernetics, 101(4):279{296, 2009.

[19] D. Beamish, S.A. Bhatti, J.Wu, and Z. Jing, "Performance limitation from delay in human and mechanical motor control", Biological Cybernetics, 99(1):43-61, 2008.

[20] D. Beamish, S. MacKenize, and J. Wu, "Speed-accuracy trade-o in planned arm movements with delayed feedback", Neural Networks, 19(5):582-599, 2006.

[21] Q. Hu, "Differential equations with state-dependent delay: global Hopf bifurcation and smoothness dependence on parameters", PhD thesis, York University, 2008.

[22] Q. Hu & J. Wu, "Global continua of rapidly oscillating periodic solutions of state-dependent delay differential equations", Journal of Dynamics and Differential Equations, 22(2):253-284, 2010.

[23] Q. Hu, J. Wu, and X. Zou, "Estimates of periods and global continua of periodic solutions for state-dependent delay equations", SIAM Journal on Mathematical Analysis, 44(4):2401-2427, 2012.

 ISBN: 978-1-61208-423-7

Accelerated Mean Shift For Static And Streaming Environments

Daniel van der Ende*, Jean Marc Thiery†, and Elmar Eisemann‡
Delft University of Technology
Delft, The Netherlands
Email: *daniel.vanderende@gmail.com, †j.thiery@tudelft.nl, ‡e.eisemann@tudelft.nl

***Abstract*—Mean Shift is a well-known clustering algorithm that has attractive properties such as the ability to find non convex and local clusters even in high dimensional spaces, while remaining relatively insensitive to outliers. However, due to its poor computational performance, real-world applications are limited. In this article, we propose a novel acceleration strategy for the traditional Mean Shift algorithm, along with a two-layer strategy, resulting in a considerable performance increase, while maintaining high cluster quality. We also show how to to find clusters in a streaming environment with bounded memory, in which queries need to be answered at interactive rates, and for which no mean shift-based algorithm currently exists. Our online structure is updated at very minimal cost and as infrequently as possible, and we show how to detect the time at which an update needs to be triggered. Our technique is validated extensively in both static and streaming environments.**

Keywords–Data stream clustering; Mean Shift

I. INTRODUCTION

Although streams of data have been generated for a considerable amount of time, the analysis of these streams is a relatively young research field. Data streams present additional challenges for the field of data mining. Traditional data mining algorithms cannot be directly applied to data streams, due to the additional data stream constraints, which has led to considerable research into new methods of analyzing such high-speed data streams [1], [2].

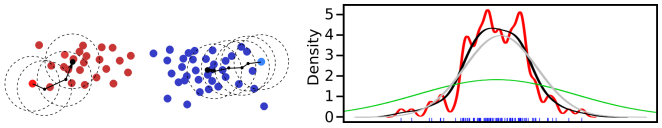
Mean Shift, initially proposed by Fukunaga et al. [3] and later generalized by Cheng et al. [4] and Comaniciu et al. [5], is a well-known clustering algorithm that has a number of attractive properties, such as its ability to find non-convex clusters. However, its performance has always been a concern, and it is because of this that, we believe, Mean Shift has never been applied in a streaming environment.

In this article, we present a modification of Mean Shift that can be used in both static and stream clustering environment. In the static environment, execution time of Mean Shift is reduced while a high level of cluster performance is maintained (Section III-A). Our main contribution concentrates on streaming environments, and we derive an efficient triggering mechanism, used to determine when a reclustering of the structure is necessary (Section III-B). We provide extensive experimental validation of our two contributions.

II. BACKGROUND

A. Mean Shift

1) Overview: Mean Shift is a mode-seeking, density-based clustering technique, with as main parameter a *kernel bandwidth* h describing the scale at which clusters are expected. In this regard, Mean Shift can be seen as a natural multi-scale clustering strategy.



(a) Each point performs an iterative gradient ascent of the estimated density towards a local maximum.

(b) Estimated density (red, grey, black) using various bandwidths. Blue points are distributed according to the green density.

Figure 1. Mean Shift overview process

Considering an input data set $\mathcal{P} = \{\mathbf{p_i}\}$ in dimension d and a density kernel K, a Mean Shift clustering of $\mathcal{P}$ is obtained as follows: For every point $\mathbf{p_i}$, initialize $p_i^0 = \mathbf{p_i}$ and iteratively compute p_i^{k+1} from p_i^k by performing a gradient ascent of the density kernel. Upon convergence $p_i^\infty = \bar{p}_i$, where $\bar{p}_i$ is a local maximum of the density kernel. Points of $\mathcal{P}$, which converge towards the same local maximum are then clustered together. Figure 1(a) gives a schematic overview of this process.

It should be noted, that the underlying geometric structure of the clusters is of course dependent on the kernel that is used. In particular, changing the bandwidth of the kernel results in more or fewer local maxima of the resulting density kernel. Figure 1(b) illustrates this fact.

A variety of kernels has been used in the literature, the most common of which is the traditional Gaussian kernel (with *bandwidth* $h \in \mathbb{R}$):

$$K(x,p) = \pi^{-d/2} \exp(-||x-p||^2/h^2) \qquad (1)$$

Note that this isotropic kernel is sometimes replaced by an anisotropic Gaussian kernel described by a symmetric positive matrix H (i. e., $\exp(-||x-p||^2/h^2)$ is replaced by $\exp(-(x-p)^{\mathrm{T}} \cdot H \cdot (x-p))$). However, if the anistropy of the kernel is uniform (i. e., $H(x) = H$ regardless of the location x), then these two approaches are completely equivalent.

Indeed, because H is symmetric definite positive, it can be decomposed as $H = U^{\mathrm{T}} \cdot \Sigma \cdot U$, U being a rotation matrix and Σ being a diagonal matrix with positive entries, which describe the different anisotropic scales of the kernel. Then, by noting $\sqrt{\Sigma}$ the diagonal matrix with squared scales ($\sqrt{\Sigma}^{\mathrm{T}} \cdot \sqrt{\Sigma} = \Sigma$), it is easy to verify that $(x-p)^{\mathrm{T}} \cdot H \cdot (x-p) = ||(x'-p')||^2$, with $y' := \sqrt{\Sigma} \cdot U \cdot y$ for every point y.

It is therefore entirely equivalent to use a uniform anisotropic kernel on the input data and use a uniform isotropic

 ISBN: 978-1-61208-423-7

kernel on data that has been globally transformed through the rigid transformation $y' := \sqrt{\Sigma} \cdot U \cdot y$. Note that traditionally, principal component analysis (PCA) [6] is a common strategy to first transform the input data before applying Mean Shift.

In our work, we will thus only focus on the case of isotropic Gaussian kernels.

2) Bandwidth estimation: The user may not always have an idea of what bandwidth to use, in the context of data which is difficult to explore, visualize and understand, such as high-dimensional data. There are a great number of bandwidth estimation techniques [7]–[10] providing results commonly accepted by the scientific community as *intrinsic* to the data. In this respect, Mean Shift can then be seen as a non-parametric clustering method. In our work, for datasets with non-provided bandwidth, we will use Silverman's rule of thumb [9].

3) Performance: Although Mean Shift has many attractive properties, such as its ability to find non-convex clusters and its multiscale nature, it also has some limitations and issues. The most important limitation is its performance. As Fashing et al. [11] have shown, Mean Shift is a quadratic bound maximization algorithm whose performance can be characterized as being $O(kN^2)$, where N is the number of points, and k is the number of iterations per point.

Many modifications to Mean Shift have been proposed [12]–[17]. Carreira-Perpiñán [12] identify two ways in which Mean Shift can be modified to improve performance: 1) Reduce the number of iterations, k, used for each point, 2) Reduce the cost per iteration. As Carreira-Perpiñán demonstrates, both of these techniques have their own merits and issues. Another class of Mean Shift modifications is that of data summarization, followed by traditional Mean Shift on a summary of the original data. Our algorithm falls into this category. This is an approach that other acceleration strategies have also applied [14], [16]. Further details on this will be given in Section III.

B. Data Stream Clustering

A number of authors have assessed the complexity of mining data streams [18], [19]. Barbara [19] focused on data stream clustering, listing a number of requirements: 1) Compactness of representation, 2) Fast, incremental processing of new data points, 3) Clear and fast identification of outliers. Due to the nature of streams, time is very limited. Because of this, data stream clustering algorithms need to be able to respond extremely quickly to the changes that occur over time in the dataset, often called concept drift. Moreover, because of the often huge datasets, memory is also constrained. Our approach has both attractive time and memory use characteristics, as will be discussed in Section III.

Many stream clustering algorithms use a two-phase approach. The approach centers on an online phase, which summarizes the data as it is streamed in, and an offline phase, which executes a given clustering algorithm on the summaries produced. The summaries are generally referred to as micro-clusters, and due to a number of attractive properties can be updated as time progresses and new data is streamed in (and old data is streamed out). CluStream [20] maintains q micro-clusters online, followed by a modified k-means algorithm that is executed when a clustering query arrives. DenStream [21] is similar, exchanging the k-means algorithm for DBSCAN, and distinguishing various quality levels of micro-clusters. Finally, D-Stream [22] uses a sparse grid approach. All these three algorithms have complex parameters. Moreover, when a clustering query arrives, a clustering is always executed.

The Massive Online Analysis (MOA) [23] framework offers a sandbox environment for easy comparison of several stream clustering algorithms. Among those, D-Stream [22] is the most related to our approach. Even though, D-Stream is not a Mean Shift algorithm. It is a density-based approach, and it partitions the space by computing the connected components of the set for which the local density is higher than a given threshold. Other parts of the space are considered outliers (Mean Shift offers a soft characterization of outliers, through the number of points in clusters and the value of the density at the clusters' center). It integrates a particular kind of density decaying mechanism, whereas our approach allows for various windowing strategies. Further, our main contribution is a new triggering mechanism, which detects events for when the clustering needs to be updated. Although not investigated, our triggering mechanism could be used for D-Stream as well. Note that the purpose of this paper is not to claim the superiority of Mean Shift over other existing clustering algorithms. Each algorithm has its advantages and drawbacks.

III. METHOD

A. Static Clustering

As discussed in Section II-A3, there are several ways in which previous work has improved Mean Shift. Our algorithm aims to reduce the number of input points for the Mean Shift. This is achieved by first discretizing the data space using a sparse d-dimensional regular grid, with a cell size of the order of the bandwidth (coarser discretizations lead to artifacts). For each grid cell C_i, the number of points n_i assigned to it is maintained, along with the sum of these points S_i. This enables the computation of an average position of the points within the cell, denoted $\overline{C_i} = S_i/n_i$. We then simply cluster the cells $\{C_i\}$ by applying the Mean Shift algorithm over them, using $K_{\overline{C}}(p, C_i) = n_i K(p, \overline{C_i})$ as the underlying kernel (see Algorithm 1). This is equivalent to computing the Mean Shift over all input points, after having set each point to the center of its cell. Although extremely simple, this strategy proved robust and efficient during our experiments. It also allows us to run Mean Shift over infinitely growing datasets with bounded memory, as long as the range of the data remains bounded, which is a required property in streaming environments.

Algorithm 1 Update clustering of cells $\{C_i\}$.

for all cell C_i **do**
  $\overline{C_i} = S_i/n_i$
end for
kdt = computeKdTree($\{\overline{C_i}\}$)
for all cell C_i **do**
  $\hat{c}_i = \overline{C_i}$
  for $it < ItMax$ **do**
    NN=kdt.nearestNeighbors($\hat{c}_i$)
    $\hat{c}_i = \sum_{k \in NN} n_k K(\hat{c}_i, \overline{C_k}) \overline{C_k} / \sum_{k \in NN} n_k K(\hat{c}_i, \overline{C_k})$
  end for
end for
cluster $\{C_i\}$ based on proximity of $\{\hat{c}_i\}$.

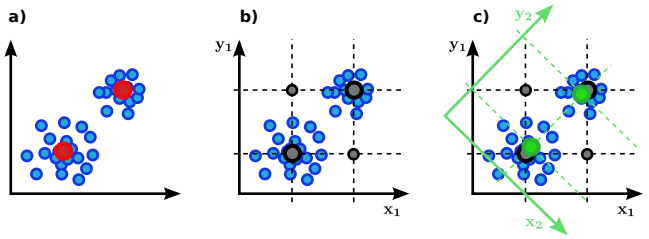


Figure 2. **a):** Two clusters in $2D$, with their centers in red. **b):** Clusters in each dimension (dot lines), whose product badly approximate the $2D$ clusters. **c):** Consensus over additional axes helps to identify the $2D$ clusters from the product of the $1D$ clusters of the projected data.

B. Stream Clustering

For stream clustering, the most common methodology is a two-phase approach, as discussed in Section II-B. A major disadvantage of this approach is that when a user query arrives, the offline clustering algorithm is executed over the data summaries, regardless of whether the dataset has changed significantly since the previous execution of the offline phase. This leads to unnecessary and expensive computations.

Our algorithm aims to avoid such needless clustering algorithm execution by accurately detecting when the data has changed sufficiently to warrant a new clustering. This is achieved by fast, effective analysis of the data currently being considered. It should be noted that this approach can be used, regardless of the type of window used. In this article, we have applied both landmark and sliding windows of various sizes.

First, when the stream clustering is initialized, a static clustering, as described in Section III-A is performed. This clustering will serve as our initial reference clustering. On each stream iteration, we evaluate whether the data distribution has changed sufficiently compared to this reference clustering to require a reclustering. If so, a clustering is executed and the result is considered the reference clustering for future iterations. By only executing the clustering algorithm, Mean Shift in our case, when necessary, a great amount of execution time is saved. Querying the cluster for a point is then done by finding the closest cell average and retrieving its cluster index (this requires an acceleration structure such as a Kd-Tree, which was already computed at the Mean Shift step).

We base our trigger mechanism on the monotonicity lemma defined by Agrawal et al. [24] as:

Lemma 3.1 (Agrawal): If a collection of points S is a cluster in a k-dimensional space, then S is also part of a cluster in any $(k-1)$-dimensional projections of this space.

Following this lemma, if any of the k-dimensional clusters change, a $(k-1)$-dimensional subcluster should also change. We make use of this point and set up a collection of low-dimensional *data observers* (in our case, $1D$), which we can update efficiently when adding or removing points from the structure, and which will trigger a reclustering of the structure when necessary. Our algorithm is parametrized by the chosen distribution of observers as well as by their *sensitivity*.

Each observer i is defined as an histogram $\mathbf{H_i}$ of the data projected onto an axis $\mathbf{a_i}$. Because high-dimensional data usually overlaps in separate dimensions (see Figure 2), we consider not only the canonical axes $\{\mathbf{e_k}\}$, but also randomly distributed axes in $\mathbb{R}^d$, and we will define the final decision for the reclustering as a consensus over the observers. Since we cannot make any assumption on the upcoming data (e. g., align data using PCA), we create a random set of pairs of indices $(k_1, k_2) \in [1, d]^2$ and define the additional axes as $(\mathbf{e_{k_1}} + \mathbf{e_{k_2}})/\sqrt{2}$ and $(\mathbf{e_{k_1}} - \mathbf{e_{k_2}})/\sqrt{2}$ (we thus *intricate* the canonical dimensions (k_1, k_2), see Figure 2**(c)**). All histograms are treated equally throughout.

When a clustering is performed, each histogram is saved as $\bar{\mathbf{H}}_\mathbf{i}$. On each subsequent stream iteration, data points are added to the grid (or removed from it if a time-dependent window is used), all histograms are updated, and we determine if the stream iteration has significantly altered the data distribution, in which case we need to update the clustering.

We define the measure between histograms $\bar{\mathbf{H}}_\mathbf{i}$ and $\mathbf{H_i}$ as their Jensen-Shannon divergence:

$$D_{JS}(\bar{\mathbf{H}}_\mathbf{i} \parallel \mathbf{H_i}) = \frac{1}{2} D_{KL}(\bar{\mathbf{H}}_\mathbf{i} \parallel M) + \frac{1}{2} D_{KL}(\mathbf{H_i} \parallel M) \quad (2)$$

where $M = \frac{1}{2}(\bar{\mathbf{H}}_\mathbf{i} + \mathbf{H_i})$, and $D_{KL}(P \parallel Q)$ is the Kullback-Leibler divergence between histograms P and Q:

$$D_{KL}(P \parallel Q) = \sum_k P(k) \ln \frac{P(k)}{Q(k)} \quad (3)$$

This measure is a distance, which is (symmetric and) always defined. Note that the direct use of the Kullback-Leibler divergence between $\bar{\mathbf{H}}_\mathbf{i}$ and $\mathbf{H_i}$ results in $+\infty$ in cases where points are removed from a cell, i. e., $Q(k) = 0$ in (3).

A histogram i *votes* for a reclustering if $D_{JS}(\bar{\mathbf{H}}_\mathbf{i} \parallel \mathbf{H_i}) > \epsilon$ (ϵ defines the *sensitivity*, which is our main input parameter).

A reclustering is then decided if the proportion of histogram voting for a reclustering is larger than a random variable, which we take between 0 and 1. This procedure is a standard Monte Carlo voting scheme, which will never (resp. always) trigger a reclustering if no (resp. all) histograms vote for it, and which will trigger a reclustering with probability defined by the consensus among the observers.

The procedure described above (for which pseudo-code is given in Algorithm 2) is easily maintainable in a streaming environment, as it only requires removal and addition of points to histograms, which can take place very quickly. Moreover, the discretization of the data space bounds the memory use in such a way that very large datasets and data streams can succinctly, but accurately be stored and used.

Algorithm 2 Add (remove) p during streaming.

Require: saved histograms $\{\bar{\mathbf{H}}_\mathbf{i}\}$, sensitivity ϵ
  grid $\leftarrow (\rightarrow)\ p$ ▷ update grid
  $n^{\text{vote}}_{\text{histo}} = 0$
  for all histogram $\mathbf{H_i}$ with axis $\mathbf{a_i}$ **do**
    $\mathbf{H_i} \leftarrow (\rightarrow)\ \mathbf{a_i}^{\mathsf{T}} \cdot p$ ▷ update $\mathbf{H_i}$
    $n^{\text{vote}}_{\text{histo}} += D_{JS}(\bar{\mathbf{H}}_\mathbf{i} \parallel \mathbf{H_i}) > \epsilon\ ?\ 1 : 0$ ▷ get vote of $\mathbf{H_i}$
  end for
  if $n^{\text{vote}}_{\text{histo}} > rand() * N^{\text{total}}_{\text{histo}}$ **then**
    for all histogram $\mathbf{H_i}$ **do** $\bar{\mathbf{H}}_\mathbf{i} = \mathbf{H_i}$ ▷ save $\mathbf{H_i}$ in $\bar{\mathbf{H}}_\mathbf{i}$
    end for
    require update of Mean Shift
  end if

 ISBN: 978-1-61208-423-7

IV. Empirical Results

A. Metrics

We have computed the following cluster validation metrics: Jaccard Index, Rand Index, Fowlkes-Mallows Index, Precision, Recall, F-Measure (see the work of Meila et al. [25]). These metrics are based on pair-wise comparison of points of a reference clustering A and a comparison clustering A'. All metrics assess whether A' correctly classified the relation between the points in each pair. We use these 6 metrics to quantify our results instead of simply picking one, because there is no real consensus on what is the correct metric between clusterings. Furthemore, the metrics we chose are common in the data clustering scientific community and will hopefully provide a real insight into the behaviour of our algorithms to the reader. A value of 0 indicates completely different clusterings whereas 1 indicates identical ones.

We implemented our method in Python, since it is cross-platform and integrates well with real-world systems.

For the static clustering experiments, we compared the traditional Mean Shift and our modified algorithms on the input data points (with the same input bandwidth).

For the stream clustering experiments, the metrics show the deviation between the clustering of the cell averages $\{\overline{C_i}\}$, when using the triggering mechanism or instead updating the clustering every time.

The reason for this choice (comparing clusterings of the averages instead of the original points) is simply practical: we could not run the computation of these metrics on huge datasets for every stream step in a reasonable amount of time. Fortunately, the depicted errors are over-conservative: the true errors are actually lower than the ones we show. Indeed, consider the case of false classification of a new point in a clustering of $100k$ points: it will have a minor impact on the metrics as it is an outlier in the data, however it will create a new grid average in our coarse summarization grid (e.g., summarized by 100 cells) and will therefore result in computed errors (based on the averages), which are much higher.

B. Static Clustering

In order to evaluate our algorithm's performance, a large number of datasets were used. For each dataset, we compare our method with the traditional Mean Shift. Figure 3 shows a comparison between our approach and traditional Mean Shift. Most metric values have a value of the order of 0.99. While there are some minor differences, these regard points which are at the boundaries between visible clusters or outliers. In general, higher errors occur for datasets presenting a high variability of the range over its various dimensions (see the remark on the equivalence between isotropic and anisotropic Mean Shift in Section II-A1).

We have conducted experiments in higher dimension. Although visually comparable, it is difficult to even assess the correctness of the Mean Shift clustering by projecting the data on a 2D space, due to overlap in the visualization. Table I summarizes our results on various datasets commonly found in the scientific literature.

While we experienced a reasonable gain in performance for small to reasonably big datasets, this is of small importance. Rather, we emphasize that our approach produces results which

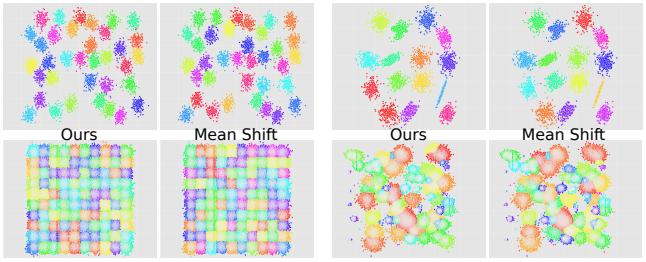


Figure 3. Comparison with Mean Shift on 2D data.

TABLE I. Summary of static clustering results. d: dimension. N: number of points. M1: Jaccard Index. M2: Fowlkes-Mallows Index. M3: Rand Index. M4: Precision. M5: Recall. M6: F-Measure

	d	N	M1	M2	M3	M4	M5	M6
A1	2	3000	0.95	0.97	0.99	0.97	0.97	0.97
A2	2	5250	0.96	0.98	0.99	0.98	0.98	0.98
A3	2	7500	0.96	0.98	0.99	0.98	0.98	0.98
S1	2	5000	0.99	0.99	0.99	0.99	0.99	0.99
S2	2	5000	0.95	0.98	0.99	0.98	0.97	0.98
S3	2	5000	0.87	0.93	0.99	0.95	0.91	0.93
S4	2	5000	0.85	0.92	0.99	0.94	0.90	0.92
Birch 1	2	100000	0.91	0.95	0.99	0.95	0.95	0.95
Birch 2	2	100000	0.64	0.78	0.95	0.76	0.99	0.78
Birch 3	2	100000	0.95	0.97	0.99	0.97	0.97	0.97
Dim 3	3	2026	0.99	0.99	0.99	0.99	0.99	0.99
Dim 4	4	2701	0.99	0.99	0.99	0.99	0.99	0.99
Dim 5	5	3376	0.99	0.99	0.99	0.99	0.99	0.99
D5	5	100000	0.99	0.99	0.99	0.99	0.99	0.99
Abalone	8	4177	0.99	0.99	0.99	0.99	0.99	0.99
D10	10	30000	0.99	0.99	0.99	0.99	0.99	0.99
D15	15	30000	0.99	0.99	0.99	0.99	0.99	0.99

are consistent with the traditional Mean Shift. This is the most important part of the validation, as it indicates that our approach can be used for Mean Shift clustering in a streaming environment, with potentially infinitely growing data. Note that, by construction, the error which is introduced by our approximation decreases with the size of the datasets on which it is used, while its efficiency obviously increases drastically.

C. Stream Clustering

For the stream clustering validation, we compare the results obtained when running our approach with the reclustering trigger enabled and disabled (i. e., reclustering on every stream iteration, regardless of lack of changes in the data distribution).

We show results on datasets of dimension 2 and 7, with a varying value of ϵ, with a fixed time window (with removal of old points) or not (adding points only), and with time-coherent or time-incoherent streaming of the data (i. e., whether the data is streamed in a structured way). Please note that "time-coherency" refers to the fact that points which are streamed in successively are roughly expected to belong to the same cluster. Table II summarizes the statistics of the conducted experiments, and the metrics plots for these test runs are presented in Figure 4. One interesting aspect with regards to our reclustering triggering is the value of ϵ which is used to threshold the Jensen-Shannon divergence value. For the CoverType dataset (dimension 7, $581k$ points), two test runs with identical configurations were performed, with $\epsilon = 0.001$ (Figure 4 (a)), and with $\epsilon = 0.003$ (Figure 4 (b)). The initial clustering was both times performed with $25k$ points, which is

TABLE II. Statistics of the experiments conducted in streaming environments. N: number of points. d: dimension. n: number of points for the initial clustering. δn: number of points added at each stream step. Window: number of points of the sliding window (if used). **TC**: time-coherency of the data stream.

	N (tot.)	d	ϵ	n (init.)	δn	Window	**TC**
a) Cov	$581k$	7	0.001	$25k$	$1k$	$25k$	NO
b) Cov	$581k$	7	0.003	$25k$	$1k$	$25k$	NO
c) Synth.1	$1M$	2	0.003	$50k$	$1k$	$50k$	NO
d) Synth.2	$1M$	2	0.003	$50k$	$1k$	$50k$	YES
e) Synth.3	$1M$	2	0.003	$50k$	$1k$	NO	YES

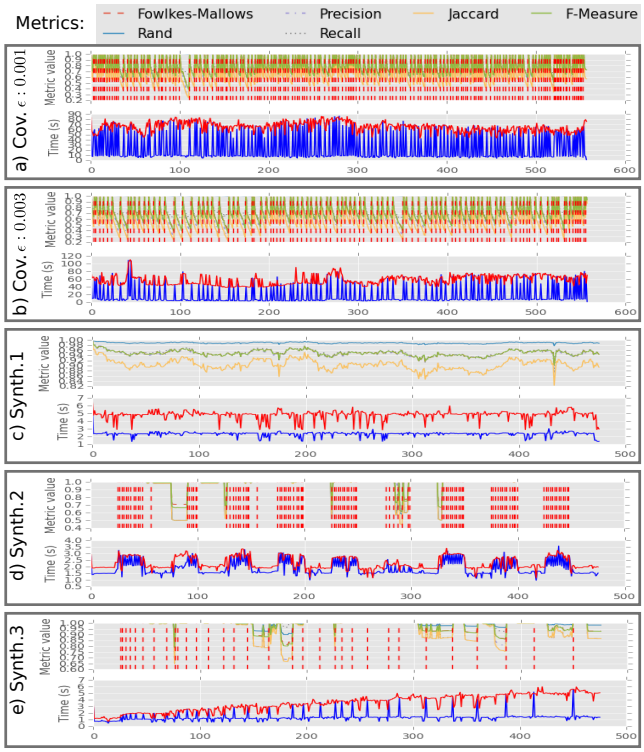


Figure 4. Comparison between our triggered clustering and constantly updated clustering on various datasets. For the timings, the red curve indicates the computation time per stream iteration when no triggering is used, and the blue curve indicates the one when our triggering mechanism is used. The vertical, dashed red lines indicate triggering events.

also the length of the sliding window that was used, and data points were streamed by sets of $1k$ points.

A lower value for ϵ should result in more frequent reclustering triggers, in an attempt to maintain a higher level of clustering quality. Although hard to see due to the high number of reclusterings, it is clear from these images that the lower ϵ affects the triggering mechanism. Reclusterings are more frequent and generally cluster quality is kept at a higher level. It should be noted that this dataset is also an example of a failure case of our algorithm, but also of the Mean Shift algorithm and any bandwidth-dependent algorithm. From the clustering results it is clear that the bandwidth estimate is completely incorrect. The bandwidth for this dataset was computed to be approximately 105.39. However, the CoverType dataset is not normally distributed, and the range of the data over the various dimensions varies from 1 to 109. Note for example that, on this dataset, a state-of-the-art Mean Shift grid approach implemented in Scikit-learn [26] provided results with metric values of 0.2 (with the same grid parameters).

For other experiments, the value of ϵ was set to $\epsilon = 0.003$.

Experiment Synth.1 (Figure 4 (c)) was done with a dataset of $1M$ points in dimension 2, with a sliding window of $25k$ points, with $1k$ points added at each stream and for time-incoherent streamed data. We observe that no reclustering is ever performed for this experiment. However, the metrics we obtain over time consistently remain over 0.7, which indicates that the initial clustering we had was good enough for the whole streaming session. Note that 0.7 is roughly the metrics values for which a reclustering was decided in previous experiments under similar conditions ($\epsilon = 0.003$), which indicates that the a-posteriori errors resulting from a given value of ϵ are consistent over the experiments.

Experiment Synth.2 (Figure 4 (d)), which was conducted under similar conditions as experiment Synth.1 with the sole difference of streaming time-coherent data, presents highly structured reclustering events. The reclustering events correspond to the appearance and disappearance of complete clusters, Our triggering mechanism visibly adapts in a non-trivial way to the structure of the underlying data.

Note that, for real-life datasets, the reality corresponds probably to a mix of these two behaviours (i. e., there are several levels of consistency in data, e. g., for visited websites during the day or in various places over the world, etc.). The strength of our approach is that we make no assumption on the structure of the data which is going to be streamed in, and that it adapts automatically to its underlying structure.

Finally, experiment Synth.3 was performed on a dataset of $1M$ points, without window (i. e., no points are removed). It is visible that the time for updating the structure grows almost linearly over time, while the frequence of the triggering events is actually inversely linear over time, which is the behaviour which is to be expected in order to provide timely-bounded analysis of growing data. Of course, there is a limit to this, and it is impossible to guarantee this behaviour for arbitrarily distributed data (over space and/or time).

V. Discussion

The experiments performed have shown that our algorithm produces accurate clusterings, at reduced cost, and only when necessary to maintain cluster quality. Moreover, our triggering mechanism allows Mean Shift to be applied in a streaming environment, which, to our knowledge, has not been achieved before. We now discuss possible extensions of our method.

First, it may be possible to apply a divide-and-conquer approach to the overall data space using the information contained in the histograms. In some cases, the clusters in the data space are clearly separated. It could then be useful to split the space, based on this information, into separate areas using cutting hyperplanes, and run our method on these distinct subspaces. This would allow for greater parallelism and avoid unnecessary work being done on clusters that do not change. Hinneburg et al. [27] developed a clustering algorithm based only on such cutting hyperplanes. Their technique for finding the optimal cutting hyperplanes could be applied to the sparse grid used in the approach discussed in this article.

 ISBN: 978-1-61208-423-7

Second, data sparseness can reduce the performance improvement over Mean Shift to some extent. In these extreme cases, if each point is placed in a grid cell of its own, running the Mean Shift on these cell averages $\overline{C_i}$ is effectively the same as running on the input data. This is also dependent on the bandwidth value used. Note however, that this effect appears mostly for "small" datasets. For very big datasets, which we target, it becomes improbable to keep a high degree of sparsity.

Third, as was discussed in Section II-A2, the bandwidth value h to a large extent determines the final clustering result. Thus, the quality of the results are also dependent on the quality of the bandwidth estimate or the value provided by the user. This sensitivity to the bandwidth parameter is an inherent problem for all approaches based on a kernel density estimate. An interesting avenue of research could be to maintain clusterings for various bandwidth values, and to use this information to derive a continuous clustering as the interpolation of the computed ones. Currently, if the bandwidth is reset by the user during streaming, our approach cannot update the clustering efficiently.

VI. Conclusion and Future Work

In this paper, we have presented an effective and efficient modification of the Mean Shift. The modification allows for faster execution when applied in a static context, and makes it possible to use Mean Shift in a streaming environment. The application of Mean Shift in a streaming environment is based on a two-phase approach, where a memory-efficient structure is maintained online, and Mean Shift is executed on this summary structure offline. Our triggering mechanism ensures that the expensive process of executing Mean Shift happens as infrequently as possible and only when necessary to ensure a high clustering quality. Only if the data distribution has changed strongly Mean Shift is executed again. Our approach is extensively validated in both the static and streaming environments, and shows good performance in both. We believe that our triggering mechanism might be usable for other stream algorithms. However, designing optimal mechanisms for specific stream algorithms as well as for specific error measures is an interesting lead for future work.

Acknowledgments

This work was partly funded by Mobile Professionals BV and EU FET Project Harvest4D.

References

[1] C. C. Aggarwal, Ed., Data Streams - Models and Algorithms, ser. Advances in Database Systems. Springer, 2007, vol. 31.

[2] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. de Carvalho, and J. Gama, "Data stream clustering: A survey," ACM Computing Surveys (CSUR), vol. 46, no. 1, 2013, p. 13.

[3] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," Information Theory, IEEE Transactions on, vol. 21, no. 1, 1975, pp. 32–40.

[4] Y. Cheng, "Mean shift, mode seeking, and clustering," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 17, no. 8, 1995, pp. 790–799.

[5] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24, no. 5, 2002, pp. 603–619.

[6] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no. 11, 1901, pp. 559–572.

[7] D. Comaniciu, V. Ramesh, and P. Meer, "The variable bandwidth mean shift and data-driven scale selection," in Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, vol. 1. IEEE, 2001, pp. 438–445.

[8] D. Comaniciu, "An algorithm for data-driven bandwidth selection," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 25, no. 2, 2003, pp. 281–288.

[9] M. C. Jones, J. S. Marron, and S. J. Sheather, "A brief survey of bandwidth selection for density estimation," Journal of the American Statistical Association, vol. 91, no. 433, 1996, pp. 401–407.

[10] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," Journal of the Royal Statistical Society. Series B (Methodological), 1991, pp. 683–690.

[11] M. Fashing and C. Tomasi, "Mean shift is a bound optimization," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, 2005, pp. 471–474.

[12] M. A. Carreira-Perpinan, "Acceleration strategies for gaussian mean-shift image segmentation," in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 1. IEEE, 2006, pp. 1160–1167.

[13] D. DeMenthon and R. Megret, Spatio-temporal segmentation of video by hierarchical mean shift analysis. Computer Vision Laboratory, Center for Automation Research, University of Maryland, 2002.

[14] D. Freedman and P. Kisilev, "Fast mean shift by compact density representation," in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 1818–1825.

[15] B. Georgescu, I. Shimshoni, and P. Meer, "Mean shift based clustering in high dimensions: A texture classification example," in Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. IEEE, 2003, pp. 456–463.

[16] H. Guo, P. Guo, and H. Lu, "A fast mean shift procedure with new iteration strategy and re-sampling," in Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on, vol. 3. IEEE, 2006, pp. 2385–2389.

[17] X.-T. Yuan, B.-G. Hu, and R. He, "Agglomerative mean-shift clustering," Knowledge and Data Engineering, IEEE Transactions on, vol. 24, no. 2, 2012, pp. 209–219.

[18] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream systems," in Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2002, pp. 1–16.

[19] D. Barbará, "Requirements for clustering data streams," ACM SIGKDD Explorations Newsletter, vol. 3, no. 2, 2002, pp. 23–27.

[20] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in Proceedings of the 29th international conference on Very large data bases-Volume 29. VLDB Endowment, 2003, pp. 81–92.

[21] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise." in SDM, vol. 6. SIAM, 2006, pp. 326–337.

[22] Y. Chen and L. Tu, "Density-based clustering for real-time stream data," in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007, pp. 133–142.

[23] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "Moa: Massive online analysis," The Journal of Machine Learning Research, vol. 11, 2010, pp. 1601–1604.

[24] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, Automatic subspace clustering of high dimensional data for data mining applications. ACM, 1998, vol. 27.

[25] M. Meilă, "Comparing clusteringsan information based distance," Journal of Multivariate Analysis, vol. 98, no. 5, 2007, pp. 873–895.

[26] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.

[27] A. Hinneburg and D. A. Keim, "Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering," in Proceedings of the 25th International Conference on Very Large Data Bases, ser. VLDB '99, 1999.

 ISBN: 978-1-61208-423-7

Fast and Unsupervised Classification of Radio Frequency Data Sets Utilizing Machine Learning Algorithms

Phil Romero
Los Alamos National Laboratory
Los Alamos, New Mexico 87545
Email: prr@lanl.gov

Kalpak Dighe
Los Alamos National Laboratory
Los Alamos, New Mexico 87545
Email: kdighe@lanl.gov

***Abstract*—Collection of Radio Frequency data can overwhelm even the largest data storage capacities very quickly due to high sampling frequencies. There are many sources of possible error in maintaining an accurate record of the captured signals. These issues can be solved, in large part, through an automatic classification of data sets gathered that eliminates the possibility of human error and assures that the proper type of signals were captured in a timely fashion. In this paper, we will describe the process used to produce a classification system. The goal is to identify and use measures produced from the raw signal information and/or the spectrograms for input into an algorithm that produces clusters based on similarity that will classify the data into subsets with the least amount of computational complexity. K-means clustering and principal component analysis are utilized in a two step process to perform the classification of the data sets. Minimal amounts of measures have been found to produce satisfactory results in separating the raw signal data into dissimilar signal types based on a 32768 sample size. This minimizes computational complexity while still producing output used in the second stage of the process to classify the data sets. A method of classification was found that produces minimal false positive errors while selecting the proper number of clusters without resorting to more computationally complex methods thereby decreasing the time spent classifying.**

***Keywords*—Digital Signal Processing; Machine Learning; Radio Frequency.**

I. INTRODUCTION

Collection of Radio Frequency data can overwhelm even the largest data storage capacities very quickly due to high sampling frequencies. The sampling frequencies can range up to two or even five billion samples per second with many channels collecting the data simultaneously. Data rates can exceed 200 GB per second[1] and it is prohibitively expensive to store large samples in real time. Adding to the problem is the time required to verify the desired signals were recorded in the data collection and properly annotating the data for convenient retrieval at subsequent times. Also noteworthy is the problem created by both expected[2] and unexpected sources of radio frequency signals that can diminish the value of the data collected[3]. Human error can also lead to incorrect annotation of data whose consequences can be difficult to mitigate. These issues can be solved, in large part, through an automatic classification of data sets gathered that eliminates the possibility of human error, assures that the proper type of signals were captured in a timely fashion and eliminates the need for storage of uninteresting data sets. A methodology for signal discovery is proposed in Section II and is compared with a currently used alternative. Results are presented for an optimum sample size and input parameters in Section III. Results of the first clustering process are presented in Section IV, this process considers each data set independently. The results of the second clustering process, where data sets are compared, is discussed in section V. Finally, a conclusion is presented in Section VI.

II. METHODOLOGY

The radio frequency spectrum ranges from around 3kHz to 300GHz and is, in part, utilized to carry communication signals. These communications signals vary widely including AM radio broadcast signals, television broadcast signals, FM radio broadcast signals, Cell phone signals, GPS, and wireless computer networks. All of these signal sources can produce produce a significant amount of background noise in the RF spectrum. Typically, the background noise must be considered when capturing signals in the RF spectrum and the proper adjustments must be made to ensure they do not interfere with signals of interest, examples of which are shown in Figure 1 and in Figure 2.

A frequency analysis can be performed on the discrete-time signal by converting the time-domain sequence to an equivalent frequency-domain representation. This can be accomplished with the Fourier Transform of the discrete-time signal. Further processing can produce a spectrogram which shows the power level at given frequencies for the timespan in question as shown in Figure 3. The goal is to use measures/features produced from the raw signal information and/or the spectrograms for input into an algorithm that produces clusters based on similarity that will classify the data into subsets with the least amount of computational complexity. This would eliminate a time consuming process that must be undertaken by an expert in Digital Signal Processing that is prone to error. In order to discover signals within the data sets, spectrograms[4] would need to be produced for each segment of data, the known signals would need to be removed through the application of digital filters[5] without eliminating any part of the signal we may be interested in and the images produced would then have to be examined. Given the large numbers of data segments to examine and the possibility of a digital filter eliminating a signal of interest, this method vastly improves throughput in identifying signals within the data. Several measures were computed from both the raw signal and the spectrogram to

 ISBN: 978-1-61208-423-7

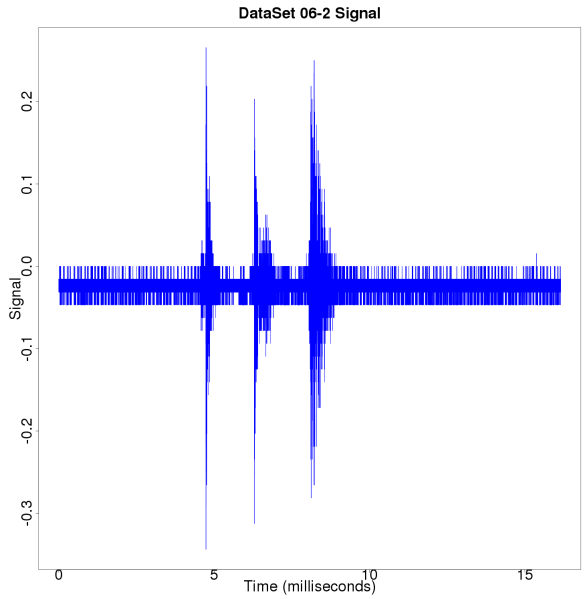


Figure 1. This figure shows a time-discrete signal waveform collected on a regular time interval in the RF spectrum.

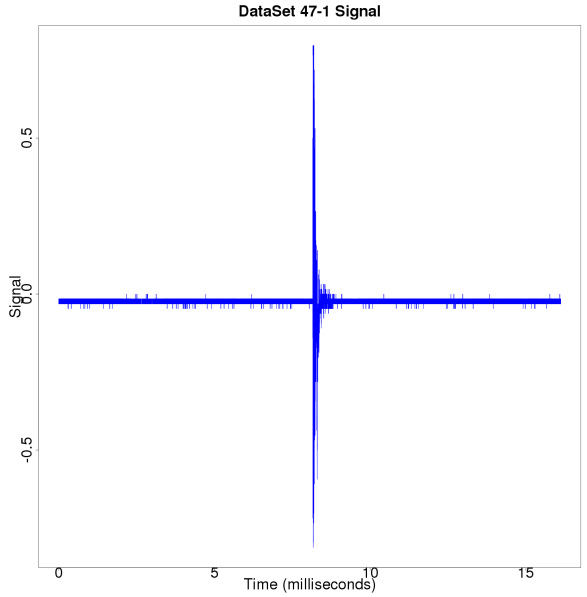


Figure 2. This figure shows another example of a time-discrete signal waveform collected on a regular time interval in the RF spectrum.

be input into the clustering algorithm. Among these features were:

1) The maximum number of frequencies identified above a given power threshold at every time processed from the spectrogram.
2) The maximum number of continuous frequencies above a given power threshold at every time processed from the spectrogram.
3) The mean power produced over the entire timespan of each spectrogram.
4) The standard deviation of power produced over the entire timespan of each spectrogram.
5) The minimum power produced at every time processed from each spectrogram.
6) The maximum power produced at every time processed from each spectrogram.
7) The median power produced over the entire timespan of each spectrogram.
8) The mode of power produced over the entire timespan of each spectrogram.
9) The mean value of the covariance of power produced over the entire timespan of each spectrogram.
10) The mean of the unprocessed signal data of a given timespan.
11) The standard deviation of the unprocessed signal data of a given timespan.
12) The minimum of the unprocessed signal data of a given timespan.
13) The maximum of the unprocessed signal data of a given timespan.
14) The absolute value of the minimum of the unprocessed signal data of a given timespan.
15) The median value of the unprocesssed signal data of a given timespan.
16) The mode of the unprocessed signal data of a given timespan.
17) The absolute value of the mean value of the unprocesed signal data of a given timespan.

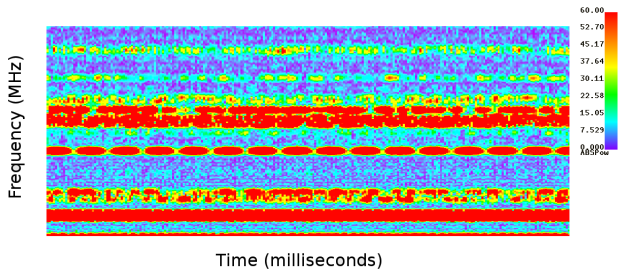


Figure 3. This figure shows an example spectrogram in the radio frequency range. The red and yellow lines are background noise caused by communication signals.

The process of identifying the timespan to process the data with was determined by processing with several different numbers of sample sizes including 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768 and 65536. The sample sizes were restricted to powers of two due to considerations of applying a discrete fourier transform to calculate the spectrograms for each timespan. This is essentially an optimization problem where the sample size needs to be able to resolve complex signal information into repeatable patterns while minimizing computational complexity. The k-means clustering algorithm[6] works by dividing a large sets of points into any number of ”neighborhoods” requested by the user. In our case, the points are all the above measurements for computed for every chosen timespan for a given sample size within a data set. It is important to note that three separate data set file lengths were utilized, consequently, the result has to be independent of the size of the data set processed. Formally, the k-means algorithm is used to solve the following problem:

Given: a set of observation (x_1, x_2, .. x_n) where each

 ISBN: 978-1-61208-423-7

observation is a y-dimension vector.

Task: Partition the n observations into k sets ("neighborhoods") to minimize the within-cluster sum of squares.

The output of the k-means algorithm is an assignment of each observation into one of the k clusters, the sum of squares within each cluster, the distance between clusters, along with other statistical measures. A technique called Principal Components Analysis (PCA) is also used to simplify a complex multivariate data set to expose the underlying sources of variation in the data. A full description of Principal Components Analysis is extremely complicated and better left to more authoritative resources[9]. A challenge with the k-means algorithm arises from the fact that it can produce different data clusters in subsequent runs because it may have found a local minima rather than the global minima. Another problem with the algorithm is that a user must select the k (or number of data clusters) prior to starting the algorithm. There are several options to guard against picking the wrong value for k including the elbow method[8], using the X-means algorithm[10], using the Gmeans algorithm[11], and a proposed manipulation of the k-means output parameters are also investigated in an attempt to minimize computational complexity.

It was unknown if the features/measures needed to have equal or unequal weightings before the methodology was implemented, however, k-means allows for changing the weightings should the need arise. Agglomerative heirarchical clustering methods were eliminated from consideration due to the added complexity deemed unecessary. There are certainly other clustering approaches that could also have been considered such as k-medoids[12] or DBSCAN[13] that are considered more robust than k-means, however, we have found in extensive use of k-means that we have never encountered any instability issues. Therefore, k-means was selected over other methods for it's flexibility and simplicity as well as it's low computational complexity.

The data was collected in an RF laboratory environment with a commercial, programmable broadband signal generator. Repeatability of the experiment is not an issue as the signal generating codes are archived.

We investigated the output of the k-means/PCA algorithms after they are applied to each of the 96 sample data sets in order to classify each of the data sets by the patterns found within them. This allows for the possibility that a given data set has more than one type of signal within the data set. It also means that this is a two stage process whereby the initial k-means/PCA process serves as input into another k-means/PCA process to classify each data set from the combinations of data found by the first process. It should also be noted that combinations of patterns found in the data set are important to find thereby rendering the two step process as necessary.

III. Results Optimizing the Sample Size and Input Parameters

Data was processed from all 96 data sets in sample sizes of 128, then doubling in size until 65,536 was reached. This produced 10 different complete sets of data that were analyzed for suitability. The smaller sample sizes produced larger amounts of clusters and longer processing times than the larger sample sizes. The combinations of larger number of clusters, when combined in the second clustering processes, would produce a more complex classification set, consequently the small sample sizes were eliminated from consideration. The larger sample size of 65,536 was thought to produce too few samples from the clustering process with smaller data sets, thereby diminishing the value of the clustering precision. An optimal sample size of 32,768 was decided upon as the proper balance between precision, output complexity, and computational complexity. The input parameters were compared to determine whether it was necessary to perform the more computationally complex calculations necessary to produce spectrograms. The R statistical packages provides output showing the importance of variables in producing the principal components analysis, from these results it was clear that it was not necessary to perform the more computationally complex work required to produce the spectrograms since input produced from processing only the raw signal data could be produced with less work (in less time) without any significant loss in clustering precision. A subset of the variables calculated from only the raw signal data were further reduced due to two factors. The first factor was that in order to produce a clustering output, all input variables must have a non-zero variance for all data sets, this eliminated many of the variables from consideration into the final optimal method. The next factor that eliminated variables for input into the clustering was the significance upon the clustering, again as determined by the principal components analysis. This process left only the following three variables that need to be calculated on the raw signal for the clustering process:

1) The mean of the unprocessed signal data of a given timespan.
2) The standard deviation of the unprocessed signal data of a given timespan.
3) The absolute value of the mean value of the unprocessed signal data of a given timespan.

IV. Results of the First Clustering/PCA Process

The R program that was written to produce the clustering/pca analysis for the first step in the process also creates several plots. A small sample of the plots produced are shown starting with the cluster plot for the first selected data set in Figure 4. The unique signal that exists in cluster number eight can be found on the first data line and is shown in Figure 5. It should be noted that a principal components analysis plot would look exactly like the cluster plot without the ovals and with the green numbers shown as symbols or labels indicating the data set identifiers.

Another set of plots for a selected data set is shown in Figure 6 and in Figure 7. This time there are two signals that are separated from all other data points in the file, the line shows that there are two members in cluster number 12. The third set of plots for yet another selected data set is shown in Figure 8 and in Figure 9. This time there are three signals that are separated from all other data points in the file, the ellipse shows that there are three members in cluster number 12. More sets of plots can be shown that show similar patterns with most of the identified clusters being very near each other on the plot and a few small clusters very isolated from the rest. However, there is another pattern shown in some plots where there are no clear outliers amongst the clusters, this occurs when no signal has been found in the data set (and when only

 ISBN: 978-1-61208-423-7

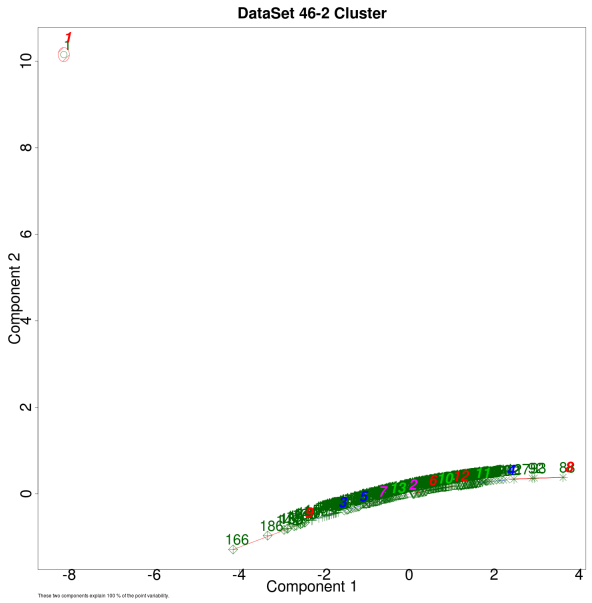


Figure 4. This figure shows a cluster plot shown with the axes being the first two principal component axes. The data point lines are the thinner font dark green labels and the clusters are identified with the thicker font.

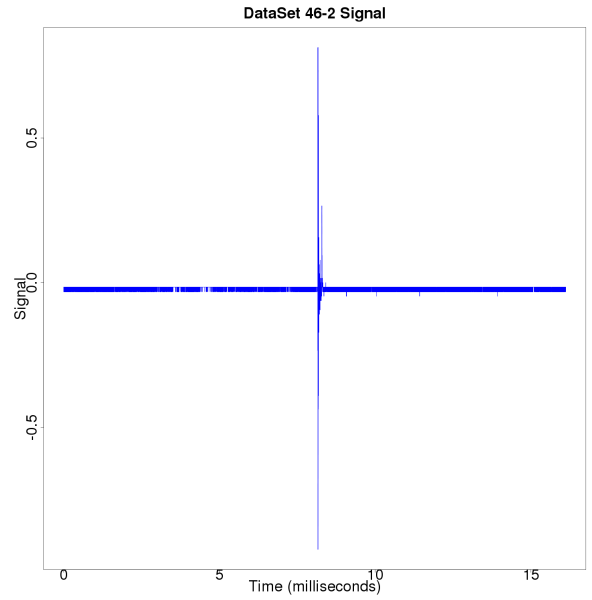


Figure 5. This figure shows the unique signal identified in the cluster plot in Figure 4.

one type of signal occurs in the data set) and the differences in the data are small througout the data set.

An additional comment should be made here noting that with very large data sets, larger than approximately 32GB, it may be better to switch to the k-medoids algorithm instead of the k-means algorithm since it is more robust to noise and outliers. This is because k-medoids minimizes a sum of pairwise dissimilarities rather than the k-means algorithm which minimizes a square of Euclidean distances[12].

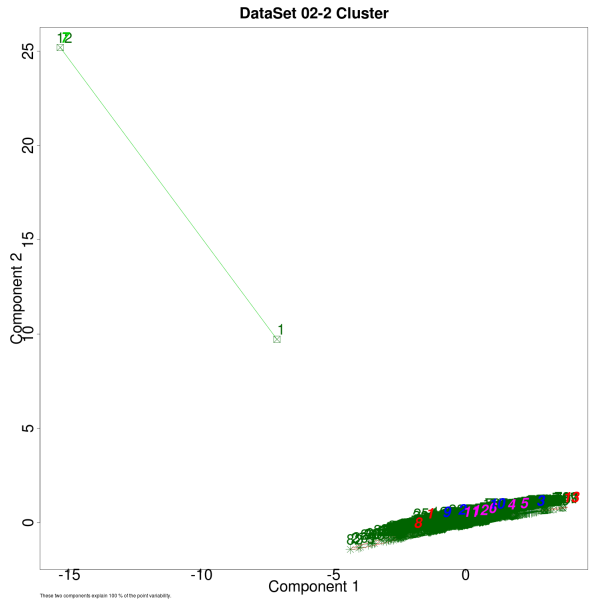


Figure 6. This figure shows a cluster plot shown with the axes being the first two principal component axes.

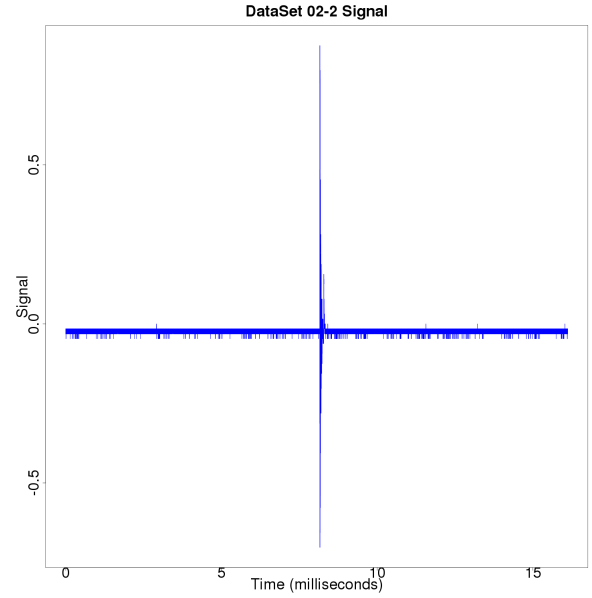


Figure 7. This figure shows the unique signal identified in the cluster plot in Figure 6.

V. Results of the Second Clustering/PCA Process

With promising results found from similarities of the principal component plots, we hypothesize that information contained in the clustering/PCA analysis might be enough to classify all the data sets into subsets. Data for each clustering of all 96 data sets were gathered to provide as much statistical data as possible. This was done in two differing techniques, the first was an attempt to characterize the data set by cluster size alone and yielded 22 columns of data for each of the data sets. The hypothesis here is that the distribution of cluster sizes

 ISBN: 978-1-61208-423-7

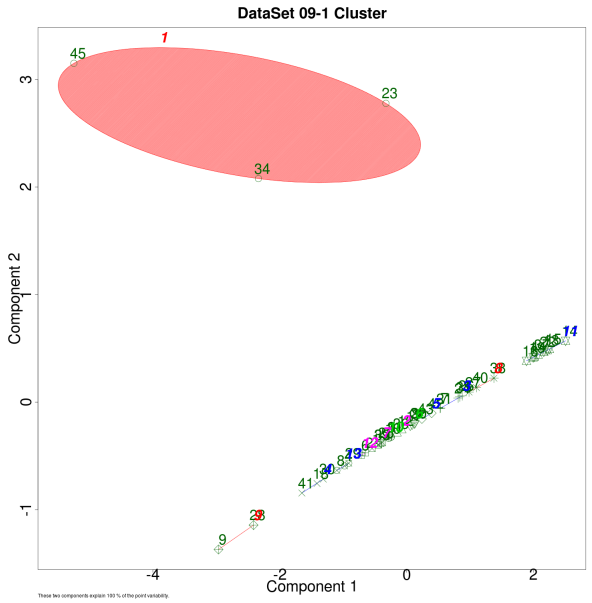


Figure 8. This figure shows a cluster plot shown with the axes being the first two principal component axes.

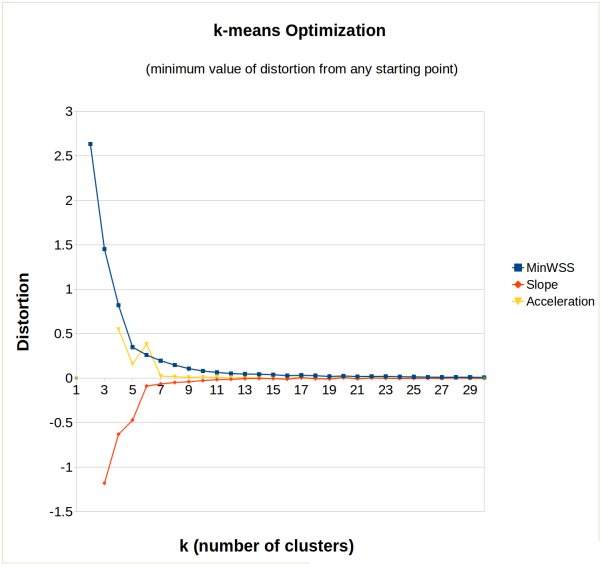


Figure 10. This figure shows a plot of the minimum value of distortion derived from any choice of starting points of the clustering.

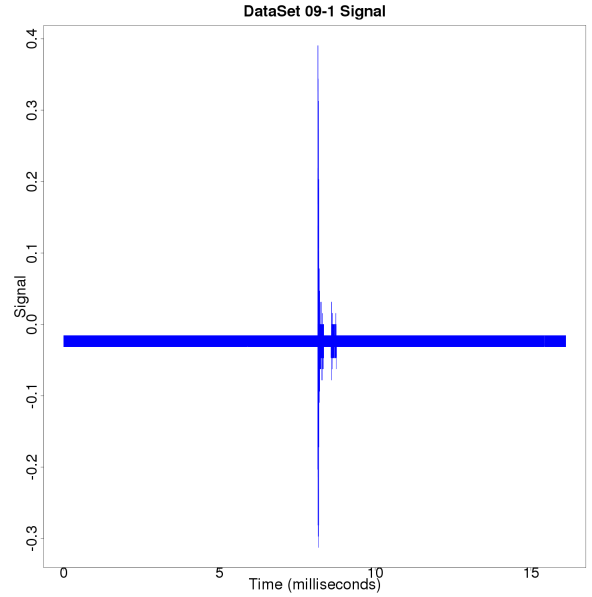


Figure 9. This figure shows the unique signal identified in the cluster plot in Figure 8.

might prove to be enough to classify the data sets. The second technique attempts to capture information from the clustering analysis about the geometry of the clusters by separating them into three parts, the cluster with the smallest number of members, the cluster with the next smallest number of members and by all the number remaining clusters combined. Centers of mass for all three input factors were then calculated for all three inputs and distances were calculated between them. The sizes of the members were normalized and added to this data producing 24 columns of data. The data was processed by cluster size alone to determine the total within sum of squares metric, also known as distortion, this yielded the plot in Figure 10. This plot also includes the first and second derivatives to clarify how the distortion curve moves with increasing numbers of clusters. This shows that there is no clear choice that stems from minimizing distortion. The value of distortion continues to decline as more clusters are added and it is clear that the limit will be reached when there are 96 clusters (the same number as the data points). This is due to the fact that there is no penalty for creating new clusters. A penalty for creating new clusters was added to the calculation by dividing the distortion value by the median number of members in the clusters, this is shown in Figure 14. This works well for geometrically processed data sets (but not for the cluster size processed data sets) and shows a clear minimum of seven as a choice for k.

The geometrical processing of the data sets produced better results that were better defined as evidenced by lesser distortion numbers output from the clustering. The x-means algorithm chose a value for k of 14 clusters whereas, the G-means algorithm chose a value for k of only 7 clusters. Verification of the clusters was done by noting if the data set had signals, what kind, and if there were high degrees of background noise in the data sets. A cluster plot shown in Figure 12 shows clusters 4 and 6 isolated on the left hand side, these clusters have no signals present in any of the data sets enclosed in these clusters. Cluster 1 has the most background noise of the clusters that have signals in them. This cluster also had three data sets in which we haven't found any signals, consequently they are thought to be misclassified by this method. All three of these data sets labeled 22-2, 23-2 and 24-2 in Figure 13 are shown as the points furthest from the center of cluster 1 and closest to cluster 4 by the cluster map plot in Figure 14. This result shows that the 3 misclassified data sets are closest to the boundary of the the data sets that have no signals in them raising the possibility of further improvement by adding another measure to the clustering to eliminate this error. However, due to the assymetric cost of missing a signal of interest over including false positives, we are confident that

 ISBN: 978-1-61208-423-7

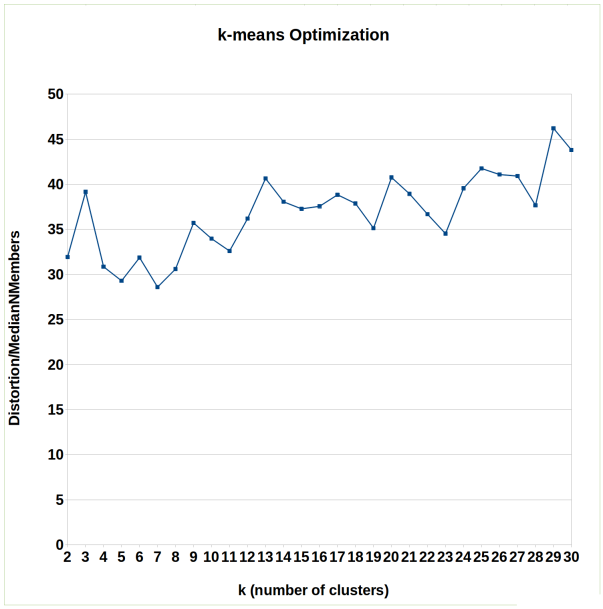


Figure 11. This figure shows a plot of the minimum value of distortion derived from any choice of starting points of the clustering, this time the distortion value is divided by the median number of members in each cluster.

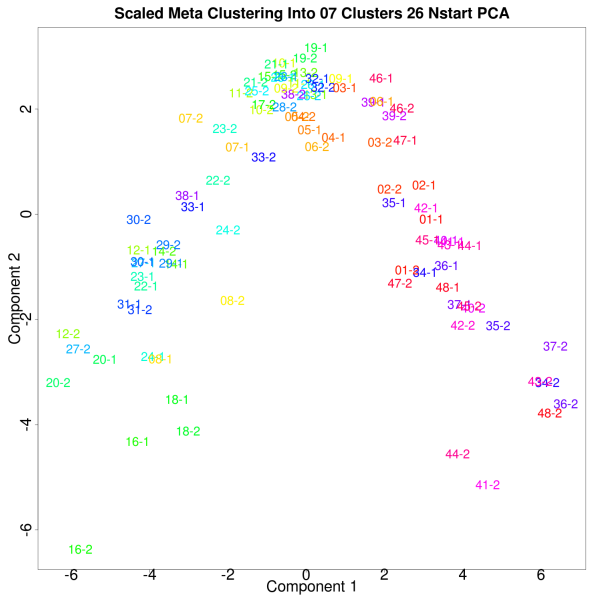


Figure 13. This figure shows a principal components analysis plot with the labels showing the data set identifiers.

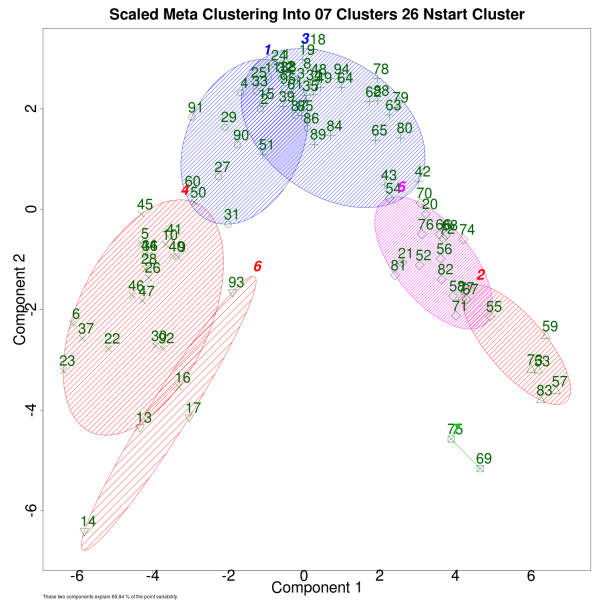


Figure 12. This figure shows a cluster plot of the the lowest distortion plot derived for seven clusters with geometrical processing of the data sets.

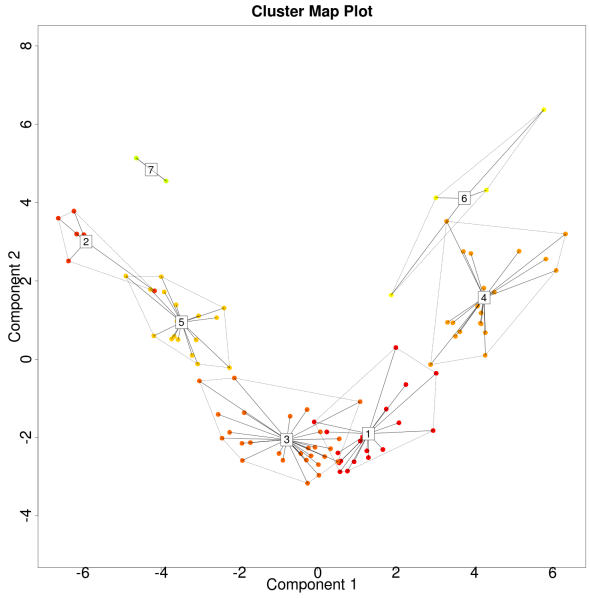


Figure 14. This figure shows a principal components analysis plot with the labels showing the data set identifiers.

method is exceptionally suited to our needs.

VI. Conclusion

A process has been described here that uses Machine Learning algorithms to classify data sets composed of RF signals. Only three measures/features have been found to produce satisfactory results in separating the raw signal data samples into classifications based on a sample size of 32,768 and the necessity of producing spectrograms was eliminated. This minimizes computational complexity while still producing output used in the second stage of the process to classify the data sets. A fast method of classification was found that produces minimal false positive errors while selecting the proper number of clusters without resorting to more computationally complex methods, thereby, decreasing the time spent classifying.

Acknowledgment

The authors would like to thank our colleague Daryl Grunau who was instrumental in making necessary resources available.

 ISBN: 978-1-61208-423-7

REFERENCES

[1] Teledyne Lecroy, "LabMaster 10-100Zi 100 GHz Oscilloscope" url=http://teledynelecroy.com/100ghz/

[2] R. W. Zeng and Y. Chen Li. "Design of Digital Multiple Frequency Notch Filter Based on Free Search Algorithm". Computer Engineering, vol. 40, number 12, 2014, pp. 209-213.

[3] R. G. Lyons. Understanding Digital Signal Processing, Third Edition. Prentice Hall. 2010, ISBN: 9780137027415.

[4] S. K. Mitra. Digital Signal Processing: A Computer-Based Approach, Fourth Edition. McGraw-Hill. 2010, ISBN: 9780077366766.

[5] J. G. Proakis and D. G. Manolakis. Digital Signal Processing: Principles, Algorithms, and Applications, Fourth Edition. Pearson Education Limited. 2013, ISBN: 9781292025735.

[6] A. Coates and A. Y. Ng. "Learning Feature Representations with K-means". Neural Networks: Tricks of the Trade, vol. 7700, 2012, pp. 561-580.

[7] T. W. Liao. "Clustering of time series data–a survey". Pattern Recognition, vol. 38, 2005, pp. 1857-1874.

[8] D. J. Ketchen, Jr and C. L. Shook. "The application of cluster analysis in Strategic Management Research: An analysis and critique" Strategic Management Journal, vol. 17, 1996, pp. 441-458.

[9] H. Abdi and L. J. Williams. "Principal Component Analysis" Wiley Interdisciplinary Reviews: Computational Statistics. 2010.

[10] D. Pelleg and A. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters" url=http://pelleg.org/shared/hp/download/xmeans.pdf

[11] G. Hamerly and C. Elkan, "Learning the k in k-means" Proceedings of the 17th Annual Conference on Neural Information Processing Systems (NIPS), 2003, pp. 281-288

[12] S. M. Tagaram. "Comparison between K-Means and K-Medoids Clustering Algorithms" International Journal of Advanced Computing. April 2011.

[13] M. Ester, H. Kriegel, J. Sander and X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise". Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press., 1996, pp. 226231.

 ISBN: 978-1-61208-423-7

Dynamical Behavior of Communicability Structures in Complex Networks

Kyungsik Kim
Department of Physics
Pukyong National University
Busan, South Korea
Email: kskim@pknu.ac.kr

Seungsik Min
Department of Natural Science
Korea Naval Academy
Changwon, South Korea
Email: fieldsmin@hanmail.net

***Abstract*—We investigate the microscopic community structure of the Korean meteorological society in the author network. Through oscillator networks, we simulate and analyze the averaged communicability functions. After constructing networks triggered an equally contributed weight between the first author and other authors in one published paper, we mainly treat these structures of communicability after constructing networks triggered an equally contributed weight between the first author and other authors in an author network. Our results support the development of the adaptability and the stability of social organization in the social networks.**

Keywords-Communicability function; Oscillator network; Community structure; Author network.

I. INTRODUCTION

Network science has emerged and been utilized as one of the important frameworks when each researcher studies complex systems [1-4]. An important property of networks is the existence of modules or communities, and the communicability between a pair of nodes in a network is concerned with the shortest path connecting both nodes. Estrada et. al. [5] proposed a generalization of the communicability by elucidating both for the shortest paths communicating between two nodes and for all the other walks travelling between two distances. The communicability detection allows one to determine potentially the unaware and hidden relationships between nodes and also allows one to reduce a large complex network into smaller and smaller groups. Presently, the community detection within networks is an open subject of great interest.

Complex networks are also ubiquitous in many biological, ecological, technological, informational, and infrastructural systems [6–12]. It is clear that the atomic, oscillating, and social systems display network-like structures using the tools of statistical mechanics. These methods and techniques were contributed to shed light on the structure and dynamics of social, economic, biological, technological, and medical systems [13-15]. It is actually recognized that the analogy functions that describe the properties depend mainly on the structural properties of the system in networks as well.

In this paper, we study the community structure of the Korean meteorological society in the author network. The data we used are the published papers of 676 authors from the Korean meteorological society publications in the author network, from March 2008 to November 2013. We simulate and analyze four other kinds of averaged communicability.

II. COMMUNICABILITY IN NETWORKS

We mainly consider the theoretical methods of microscopic communicability in networks. First of all, let us introduce the concept of communicability in networks by describing a community structure. The communicability structure can invoke the concept of walks in networks. A walk of length k is a sequence of nodes $n_0, n_1, \ldots, n_{k-1}, n_k$ such there is a link from n_{i-1} to n_i for each $i = 1, 2, \ldots, k$ [16]. Using the concept of walk we can define the communicability between two nodes. The communicability function [4] is represented in terms of $G_{pq} = \sum_{k=0}^{\infty} c_k (A^k)_{pq}$. Here, A is the adjacency matrix, which has unity if the nodes p and q are linked to each other, but has zero otherwise. The adjacency matrix $(A^k)_{pq}$ gives the number of walks of length k starting at the node p and ending at the node q [17,18]. The two novel communicability functions are calculated as

$$G_{pq}^{EA} = \sum_{k=0}^{\infty} e \frac{(A^k)_{pq}}{k!} = (e^A)_{pq} \cdot \quad (1)$$

where e^A is a matrix function that can be defined using the following Taylor series [19]. The communicability function G_{pq} is obtained by using the weighted adjacency matrix $W = (W_{ij})_{n \times n}$. Centrality measures were originally introduced in social sciences [20,21] and are now widely used in the whole field of complex network analysis [9]. We can derive the communicability function as

$$G_{pq}^{RA} = \beta K m \omega^2 G_{pq}(\beta) \quad (2)$$

 ISBN: 978-1-61208-423-7

with the identification $\alpha = 1/K$.

From the fact that the Laplacian matrix of a connected network has a nondegenerate zero eigenvalue, we can calculate another correlation function as

$$G^{D}_{pq}(\beta) = \frac{1}{\beta K m \omega^2}(L^{+})_{pq}, \quad (3)$$

where L^{+} is the Moore–Penrose generalized inverse of the Laplacian.

In a network of quantum oscillators, we start by considering the quantum-mechanical counterpart of the Hamiltonian H_A. After arranging several equations, we can see that

$$G^{EA}_{pq} = \exp(\beta\hbar\Omega) G^{A}_{pq}(\beta). \quad (4)$$

The diagonal thermal Green's function is given in the framework of quantum mechanics, and we can compute the off-diagonal thermal Green's function as

$$G^{A}_{pq}(\beta) = \exp(-\beta\hbar\Omega)(\exp[\frac{\beta\hbar\omega^2}{2\Omega} A])_{pq}. \quad (5)$$

Note that when the temperature tends to infinity or $\beta \to 0$, there is absolutely no communicability between any pair of nodes. That is, $G^{EA}_{pq}(\beta \to 0) = 0$. If we consider the case when the temperature tends to zero or $\beta \to \infty$, then there is an infinite communicability between every pair of nodes, i.e., $G^{A}_{pq}(\beta \to \infty) = \infty$. Furthermore, the communicability function is represented in terms of

$$G^{EL}_{pq}(\beta) = G^{L}_{pq}(\beta) - 1, \quad (6)$$

where the same quantum-mechanical calculation by using the Hamiltonian H_L in Eq. (4) is calculated as

$$G^{L}_{pq}(\beta) = (\exp[-\frac{\beta\hbar\omega^2}{2\Omega} L])_{pq}. \quad (7)$$

From Eqs. (6) and (7), the communicability function G^{EL}_{p} gives $G^{L}_{pq}(\beta) - 1$ upon setting $\beta\, \hbar\omega^2 = 2\Omega$ [4]. Lastly, we simulate and analyze the averaged communicability function for a given node defined as

$$G_p = \frac{1}{n-1}\sum_{p \neq q}^{n} G_{pq}. \quad (8)$$

Consequently, the communicability functions G^{RA}_{pq} and G^{D}_{pq} become the types of the thermal Green's function of classical harmonic oscillators in networks of the community structure. The communicability functions $G^{EA}_{pq}(\beta)$ and $G^{EL}_{pq}(\beta)$ also become the types of the thermal Green's function in quantum harmonic oscillators.

III. NUMERICAL CALCULATIONS AND RESULTS

In order to simulate and analyze the averaged communicability functions, the data are the published papers for 676 authors of the Korean meteorological society publications in the author network from March 2008 to November 2013. We assume that it only takes an equally contributed weight between all authors in one published paper.

We implement the computer-simulation of the four communicability functions. Figure 1 shows the color-map diagram of the communicability function matrices as G^{RA}_{p}, $1/G^{D}_{p}$, G^{EA}_{p}, and G^{EL}_{p} for 676 authors of the Korean meteorological society publications, among four averaged communicability functions [28]. If two members are highly correlated, the representation approaches the color red. If they are weakly correlated, the representation approaches dark blue.

We can simulate four averaged communicability functions constituting a number of published papers for 676 members of the author network. The weight of community means the value (that is, 1/the number of authors) that all authors are bestowed the same weight upon one published paper. Then, we assume that the weight of community and the weight of published papers for the 1-st author is one for the 1-st author. We now speculate that the phase transition among these functions may exist near 200-th authors. In next time, we will aim to find it through networks of other societies.

Table 1 summarizes the values of the averaged communicability functions, the weight of community, and a number of published papers for 100-th, 300-th, and 600-th authors, respectively. These values are normalized values divided by the maximum value of each factors. For the value of G^{EA}_{p} between two authors, the 600-th author approaches to zero. We find that the G^{EL}_{p} relatively correlates highly when this value is compared to other ones.

 ISBN: 978-1-61208-423-7

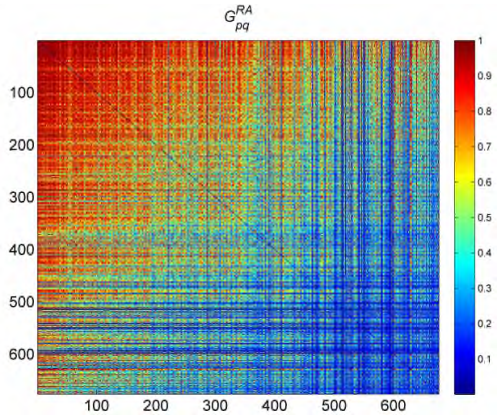


Figure 1. Color map diagram of relative communicability function matrices as G_p^{RA} from top to down for 676 members of the author network.

TABLE I. Values of the weight of published papers Pp, the weight of community W, and the averaged communicability functions.

Sequent order of authors	P_c	W_c	G_p^{EA}	G_p^{RA}	G_p^{EL}	$1/G_p^D$
100	0.247	0.059	0.057	0.065	0.998	0.149
300	0.082	0.015	0.016	0.017	0.946	0.042
600	0.030	0.005	0.0	0.004	0.554	0.026

IV. CONCLUSIONS

We have studied the community structure of Korean meteorology fields in the 676 author networks of all Korean meteorological society publications from March 2008 to November 2013. We mainly implemented the computer-simulation of the four communicability functions.

To compare the four averaged communicability functions, it was shown that the G_p^{EL} constructs a stronger community structure rather than the other three. The function G_p^{EA} finds the community structure weaker than the other three as well. We can make use of the four averaged communicability functions to compute the measures of a community structure, and it is hoped that our method and technique will lead us to more general results in the future.

It is not trustworthy now, but we anticipate that the phase transition among the averaged communicability functions may exist at one value near 200-th authors. Our results cannot yet be compared to that of other social networks, but we hope to compare to our results to other successful results in social networks that have been prominently produced and published. Next time, we hope to discuss the phase transition of the averaged communicability functions, with network systems of other societies. In the future, we will apply the community structure to the cases of different contributed weight between authors. Therefore, further work is needed for the case with societies of more than the author and citation networks. The formalism of our analysis can be extended to both the discrimination and the characterization of communicability functions in other various societies.

ACKNOWLEDGMENT

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2013R1A1A2008558) and by Center for Atmospheric Sciences and Earthquake Research (CATER 2012-6110).

REFERENCES

[1] M.J.E. Newman, Networks. An Introduction, Oxford University Press, Oxford, 2010.

[2] C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics, Rev. Modern Phys. 81 (2009) 591.

[3] Y.S. Cho, S. Hwang, H.J. Herrmann, B. Kahng, Science 339 (2013) 1185.

[4] E. Estrada, N. Hatano, M. Benzi, Phys. Rep. 514 (2012) 89.

[5] E. Estrada and N. Hatano, Phys. Rev. E 77 (2008) 036111.

[6] G. Caldarelli, *Scale-Free Networks*, Complex Webs in Nature and Technology, Oxford University Press, Oxford, 2007.

[7] L. da Fontoura Costa, O.N. Oliveira Jr., G. Travieso, F.A. Rodrigues, P.R. Villas Boas, L. Antiqueira, M.P. Viana, L.E. Correa Rocha, Adv. Phys. 60 (2011) 329.

[8] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Phys. Rep. 424 (2006) 175.

[9] E.J. Newman, SIAM Rev. 45 (2003) 167.

[10] M.J.E. Newman, *Networks*, Oxford University Press, Oxford, 2010.

[11] S.H. Strogatz, Nature 419 (2001) 268.

[12] D.J. Watts, *Small Worlds: The Dynamics of Networks Between Order and Randomness*, Princeton University Press, Princeton, 2003.

[13] M. Buchanan, *The Social Atom*, Cyan Books and Marshall Cavendish, 2007.

[14] R. N. Mantegna and E. H. Stanley, *Introduction to Econophysics: Correlations and Complexity in Finance*, Cambridge University Press, Cambridge, 1999.

[15] B.K. Chakrabarti, A. Chakraborti, A. Chatterjee, *Econophysics and Sociophysics: Trends and Perspectives*, Wiley VCH, Berlin, 2006.

[16] D. Cvetković, P. Rowlinson, S. Simić, *Eigenspaces of Graphs*, Cambridge University Press, Cambridge, 1997.

[17] F. Harary, A.J. Schwenk, Pacific J. Math. 80 (1979) 443.

[18] E. Estrada, J.A. Rodriguez-Velazquez, Phys. Rev. E 71 (2005) 056103.

[19] N. Higham, *Function of Matrices*, Philadelphia, PA, 2008.

[20] L.C. Freeman, Social Netw. 1 (1979) 215.

[21] S. Wasserman, K. Faust, *Social Network Analysis*, Cambridge University Press, Cambridge, 1994.

 ISBN: 978-1-61208-423-7