ICDS 2015

The Ninth International Conference on Digital Society

ISBN: 978-1-61208-381-0

February 22 - 27, 2015

Lisbon, Portugal

ICDS 2015 Editors

Lasse Berntzen, HBV, Norway
Åsa Smedberg, Stockholm University, Sweden

ICDS 2015

Forward

The ninth edition of The International Conference on Digital Society (ICDS 2015) was held in Lisbon, Portugal, February 22 - 27, 2015.

Nowadays, most of the economic activities and business models are driven by the unprecedented evolution of theories and technologies. The impregnation of these achievements into our society is present everywhere, and it is only question of user education and business models optimization towards a digital society.

Progress in cognitive science, knowledge acquisition, representation, and processing helped to deal with imprecise, uncertain or incomplete information. Management of geographical and temporal information becomes a challenge, in terms of volume, speed, semantic, decision, and delivery.

Information technologies allow optimization in searching an interpreting data, yet special constraints imposed by the digital society require on-demand, ethics, and legal aspects, as well as user privacy and safety.

The event was very competitive in its selection process and very well perceived by the international scientific and industrial communities. As such, it is attracting excellent contributions and active participation from all over the world. We were very pleased to receive a large amount of top quality contributions.

The accepted papers covered a large spectrum of topics related to advanced networking, applications, social networking, security and protection, and systems technologies in a digital society. We believe that the ICDS 2015 contributions offered a panel of solutions to key problems in all areas of digital needs of today's society.

We take here the opportunity to warmly thank all the members of the ICDS 2015 technical program committee as well as the numerous reviewers. The creation of such a broad and 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 efforts to contribute to the ICDS 2015. We truly believe that thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations and sponsors. In addition, we also gratefully thank the members of the ICDS 2015 organizing committee for their help in handling the logistics and for their work that is making this professional meeting a success.

We hope the ICDS 2015 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress on the topics of the conference.

We also hope that Lisbon provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

ICDS 2015 Chairs

Lasse Berntzen, HBV, Norway
Åsa Smedberg, DSV, Stockholm University/KTH, Sweden
Freimut Bodendorf, University of Erlangen, Germany
A.V. Senthil Kumar, Hindusthan College of Arts and Science, India
Charalampos Konstantopoulos, University of Piraeus, Greece
Andranik Tangian, Wirtschafts- und Sozialwissenschaftliches Institut - Düsseldorf | Karlsruhe Institute of Technology, Germany

ICDS 2015

COMMITTEE

ICDS Advisory Committee

Lasse Berntzen, HBV, Norway
Åsa Smedberg, Stockholm University, Sweden
Freimut Bodendorf, University of Erlangen, Germany
A.V. Senthil Kumar, Hindusthan College of Arts and Science, India
Charalampos Konstantopoulos, University of Piraeus, Greece
Andranik Tangian, Wirtschafts- und Sozialwissenschaftliches Institut - Düsseldorf | Karlsruhe Institute of Technology, Germany

ICDS 2015 Technical Program Committee

Habtamu Abie, Norwegian Computing Center, Norway
Mir Abolfazl Mostafavi, Université Laval - Québec, Canada
Witold Abramowicz, The Poznan University of Economics, Poland
Gil Ad Ariely, Interdisciplinary Center Herzliya (IDC), Israel
Mutaz M. Al-Debei, University of Jordan, Jordan
Ali Ahmad Alawneh, Philadelphia University, Jordan
Adolfo Albaladejo Blázquez, Universidad de Alicante, Spain
Cristina Alcaraz, University of Malaga, Spain
Salvador Alcaraz Carrasco, Universidad Miguel Hernández, Spain
Eugenia Alexandropoulou, University of Macedonia, Greece
Shadi Aljawarneh, Isra University - Amman, Jordan
Giner Alor Hernández, Instituto Tecnológico de Orizaba-Veracruz, México
Aini Aman, Universiti Kebangsaan Malaysia, Malaysia
Pasquale Ardimento, University of Bari, Italy
Masrah Azrifah Azmi Murad, Universiti Putra Malaysia, Malaysia
Saïd Assar, Institut Mines-Telecom - Ecole de Management, France
Charles K. Ayo, Covenant University, Nigeria
Gilbert Babin, HEC Montréal, Canada
Kambiz Badie, Research Institute for ICT, Iran
Ilija Basicevic, University of Novi Sad, Serbia
Farid E. Ben Amor, University of Southern California / DIRECTV, USA
Khalid Benali, LORIA -Université de Lorraine, France
Martina Benvenuti, University of Bologna, Italy
Eleni Berki, University of Tampere, Finland
Lasse Berntzen, HBV, Norway
Aljoša Jerman Blažič, SETCCE - Ljubljana, Slovenia
Marco Block-Berlitz, Hochschule für Technik und Wirtschaft Dresden, Germany
Freimut Bodendorf, University of Erlangen, Germany
Nicola Boffoli, University of Bari, Italy
Sabrina Bonomi, e-Campus University, Italy
Mahmoud Boufaida, Mentouri University of Constantine, Algeria

Danielle Boulanger, University of Lyon-Jean Moulin, France
Mahmoud Brahimi, University of Msila, Algeria
Diana Bri Molinero, Universitat Politècnica de València, Spain
Anna Brunstrom, Karlstad University, Sweden
Alberto Caballero Martínez, Universidad Católica San Antonio de Murcia, Spain
Joseph Cabrera, Marywood University, USA
Luis M. Camarinha-Matos, New University of Lisbon, Portugal
Omar Andres Carmona Cortes, Instituto Federal do Maranhão, Brazil
Maria Chiara Caschera, IRPPS-CNR - Rome, Italy
Oscar Castillo, Tijuana Institute of Technology, Mexico
Walter Castelnovo, University of Insubria, Italy
Sudip Chakraborty, Valdosta State University, USA
Ramaswamy Chandramouli, NIST, USA
Monica Chis, Frequentis A.G Austria, Romania
Sung-Bae Cho, Yonsei University, Korea
Kim-Kwang Raymond Choo, University of South Australia, Australia
Kalloniatis Christos, University of the Aegean, Greece
Yul Chu, University of Texas Pan American, USA
Arthur Csetenyi, Budapest Corvinus University, Hungary
Glenn S. Dardick, Longwood University, USA
David Day, Sheffield Hallam University, UK
Peter Day, University of Brighton, UK
Marco De Marco, Università Marconi, Italy
Gert-Jan de Vreede, University of Nebraska at Omaha, USA
Jana Dittmann, University of Magdeburg, Germany
Jerome Donet, Université de Lorraine, France
Prokopios Drogkaris, University of the Aegean - Karlovasi, Greece
Mohamed Dafir El Kettani, ENSIAS - University Mohammed V-Souissi – Rabat, Morocco
Gerard De Leoz, University of Nebraska at Omaha, USA
Pedro Felipe do Prado, University of São Paulo, Brazil
Joël Dumoulin, HumanTech Institute, Switzerland
Martin Ebner, Graz University of Technology, Austria
Noella Edelmann, Danube University Krems, Austria
Ahmed El Oualkadi, Abdelmalek Essaadi University, Morocco
El-Sayed Mohamed El-Alfy, King Fahd University of Petroleum and Minerals, Saudi Arabia
Jacques Fayolle, Télécom Saint-Etienne | Université Jean Monnet, France
Matthias Finger, SwissFederal Institute of Technology, Switzerland
Karla Felix Navarro, University of Technology, Sydney
Adina Magda Florea, University Politehnica of Bucharest, Romania
Robert Forster, Edgemount Solutions, USA
Roberto Fragale, Universidade Federal Fluminense (UFF), Brazil
Marco Furini, University of Modena and Reggio Emilia, Italy
Shauneen Furlong, Territorial Communications Ltd.-Ottawa, Canada / University of Liverpool, UK
Amparo Fúster Sabater, Information Security Institute (CSIC) – Madrid, Spain
Daniel Gallego, Universidad Politécnica de Madrid, Spain
Matjaz Gams, Jozef Stefan Institute, Slovenia
Jean-Gabriel Ganascia, University Pierre et Marie Curie, France
Miguel Garcia, University of Valencia, Spain

Copyright Information

Table of Contents

Scott Friedman, David Musliner, and Peter Keller

Visual Data Mining Framework Based on a Peer-to-Peer Architecture

Hamadi Abdelkrim, Nader Fahima
Ecole nationale Supérieure d'Informatique, ESI
Algiers, Algeria
e-mail : {a_hamadi, f_nader@}esi.dz

Abstract—Nowadays, the continuous hardware evolution is driving computer systems to be able to store large amounts of data; therefore, exploring and analyzing this huge volume of data is becoming more and more difficult to manage. Peer-to-Peer (P2P) network produce more computing power in terms of processing, communication systems and large storage. Visual Data Mining (VDM) using P2P infrastructure can be more beneficial to improve both performance and quality of the selected data. This paper aims at describing a prototype of our framework Visual Data Mining Distributed (VDMD) based on VDM algorithms in a P2P architecture.

Keywords-Visual Data Mining; Peer-to-Peer Architecture; JXTA specification

I. INTRODUCTION

Internet is a network consisting of millions of computers connected at any given time. All the computers are theoretically connected to one another, and information stored on any of these systems can be accessed. The topology of computers on the Internet is a group of machines spread out in various locations. Computers within each group or subnet are visible to each other and sometimes visible to other subnets on the Internet.

Advances in computing and communication over networks, such as Internet, intranets, and wireless networks, have resulted in various pervasive distributed environments. Many of these environments have to deal with massive data collections in terabyte scale maintained over geographically distributed sites. The data is collected as potential source of valuable information, providing a new competitive advantage. However, finding the valuable hidden information is a difficult task. Thus, some VDM algorithms are applied to find classifiers, associations, clusters and other patterns in large and distributed data sets. The purpose of VDM is either to help explain the past, or try to predict the future based on past data. Data mining techniques help identify patterns in a vast data store, and then build models that concisely represent those patterns [1].

Distributed computing plays an important role in the VDM process for several reasons. First, VDM often requires a huge amount of resources in terms of storage space and computation time. Second, data are often inherently distributed into several databases, making the centralized processing of this data not very inefficient and prone to security risks [1].

In this paper, we describe our efforts to create a framework to implement VDM in P2P architecture. The remainder of this paper is organized as follows. Section II presents a review of VDM. Section III introduces the P2P architecture. Section IV discusses the proposed VDMD framework and concludes with future work.

II. VISUAL DATA MINING

For data mining to be effective, it is important to include the human in the data exploration process and combine mainly creativity and pattern recognition abilities of human with the enormous storage capacity and the computational power of today's computers. The basic idea of data visualization is to present the data in some visual form, allowing the human to get insight into data. Visualization becomes useful as soon as the data analysis starts and the exploration goals are still vague.

Data Exploration usually follows a three step process: Overview first, zoom and filter, and then details-on-demand [2]. The Data Mining (DM) expert always needs to get an overview of the data what helps him to identify interesting patterns in this data. This corresponds to Data understanding phase of CRISP-DM methodology [3].

The techniques of VDM can be classified based on three criteria (see Figure 1): the data to be visualized, the visualization technique, and the interaction and distortion technique used [4].

The data, usually, consists of a large number of records each consisting of a list of values calling in data mining attributes or in visualization dimensions. We call the number of variables the dimensionality of the data set. Data sets may be one-dimensional, such as temporal data; two-dimensional, such as geographical maps; multidimensional, such as tables from relational database.

The next data types are text/hypertext or hierarchies/graphs. Text data type is distinguished by the fact that it cannot be easily described directly by numbers and therefore text has to be firstly transformed into describing vectors, for example word counts. Graphs types are widely used to represent relations between data, not only data alone. The last types are algorithms and software.

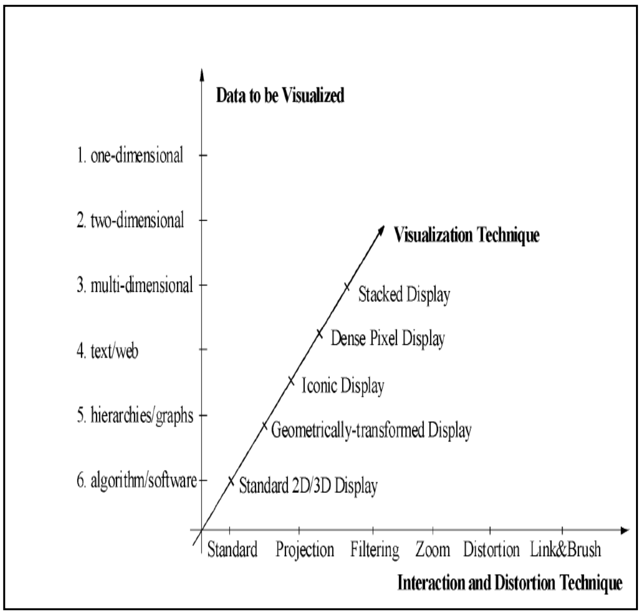


Figure 1. Classification of information visualization techniques.

Note that our framework is designed to support different data types and that it can use a combination of multiple visualization and interaction techniques.

III. DISTRIBUTED ARCHITECTURES

The client-server architecture, P2P architecture and hybrid architecture try to achieve scalability through various means. Scalability can be achieved either by increasing the resources or by reducing the consumption.

A. The Client-Server architecture

The client–server model of computing [5] is a distributed application structure that partitions tasks between the providers of a resource or service, called servers, and service requesters, called clients. Often clients and servers communicate over a computer network on separate hardware, but both client and server may reside in the same system. A server host runs one or more server programs which share their resources with clients. A client does not share any of its resources, but requests a server's content or service function. Clients, therefore, initiate communication sessions with servers which await incoming requests.

B. P2P architecture

P2P computing or networking [5] is a distributed application architecture that partitions tasks between peers. Peers are equally privileged, equipotent participants in the application. Peers make a portion of their resources, such as processing power, disk storage or network bandwidth, directly available to other network participants, without the need for central coordination by servers or stable hosts. Peers are both suppliers and consumers of resources, in contrast to the traditional client-server model in which the consumption and supply of resources is divided. Emerging collaborative P2P systems are going beyond the era of peers doing similar things while sharing resources, and are looking for diverse peers that can bring in unique resources and capabilities to a virtual community thereby empowering it to engage in greater tasks beyond those that can be accomplished by individual peers, yet that are beneficial to all the peers.

C. The Hybrid architecture

It is possible to combine P2P architecture with a server-based architecture. Hybrid models are a combination of peer-to-peer and client-server models. A common hybrid model is to have a central server that helps peers find each other. There are a variety of hybrid models, all of which make trade-offs between the centralized functionality provided by a structured server/client network and the node equality afforded by the pure peer-to-peer unstructured networks. Currently, hybrid models have better performance than either pure unstructured networks or pure structured networks because certain functions, such as searching, do require a centralized functionality but benefit from the decentralized aggregation of nodes provided by unstructured networks.

IV. OUR FRAMEWORK VDMD

Guedes et al. [6] described a service-oriented architecture (SOA) that offers simple abstractions for users and supports computationally intensive applications for data mining. Zhang et al. [7] adopted Web Services Description Language (WSDL) for specifying the interfacing of the data mining components, and Business Process Execution Language for Web Services (BPEL2WS) for specifying the execution flow.

The proposed VDMD framework is based on the hybrid topology where computers can be both client and server. The VDMD framework offers the following features:

- Huge amount of resources in storage space and computation time.
- Make the information from mass of data discovery and display to be better.
- A fragmentation of the databases.

Figure 2 shows the VDMD framework architecture. It is composed of nine main components described as follows:

- **DatabasePeer**: This peer is responsible for all database access and control. Data is received by the peer and placed in one database associated with the peer. Depending on the needs, the database can be either on the same node as the peer or on a different one as: VDM Server A, VDM Server B, Server A, Server B, Server C. The tables and indexes of the databases can be partitioned.
- **DateminingPeer**: This peer is charged to execute the data mining programs.

- **GatheringPeer**: This peer is responsible for gathering any data result from "DM Peer Group" and saving that data to a businessPeer. This peer could be a spider that looks at the result of data mining programs for data.
- **OverviewPeer:** This peer is responsible to get the user a global overview of the data.
- **FilterPeer:** This peer is responsible to execute all filters on the data.
- **DetailPeer:** This peer is responsible for executing the "on-demand details" requests.
- **ZoomPeer:** This peer can display more information about the selected data.
- **BusinessPeer:** This peer is responsible for acting as a buffer between the GUIClientPeer and the "VDM Peer Group" (OverviewPeer, FilterPeer, DetailPeer and ZoomPeer). This peer simply receives a packet and forwards it to the "VDM Peer Group", but additional logic could be incorporated.
- **GUIClientPeer:** This peer is responsible for requesting an image from the database to be displayed. The GUIClientPeer will typically be a GUI-based application that a person will use to request data from the data. The peer will interact with a BusinessPeer, which will in turn attempt to communicate with "VDM Peer Group".

Our protytype can support any VDM technique. Bellow, a working flow:

- We supposed that the datasource are fragmented on three servers : VDMServerA and ServerB and ServerC.
- Implementing the program to access to data on three machines "DatabasePeer".
- Choosing the circle segments technique [8] from the dense pixel techniques, it maps each dimension value to a coloured pixel and group these pixels belonging to each dimension into adjacent areas. To arrange the pixels on the screen, there are two techniques: recursive pattern technique and circle segments technique.
- Implementing the program of circle segments technique on three machines "Datamining Peer" so each machine can perform this program on the data of the three servers: ServerA, ServerB and ServerC.
- Imlplementing the interaction and distorition techniques(Dynamic projections, interactive filtering, intercative zomming and interactive distortion) on four machines as following: OverviewPeer, FilterPeer, DetailPeer and ZoomPeer.
- Implementing a GUI application which offers the services and displays the result to the final users.

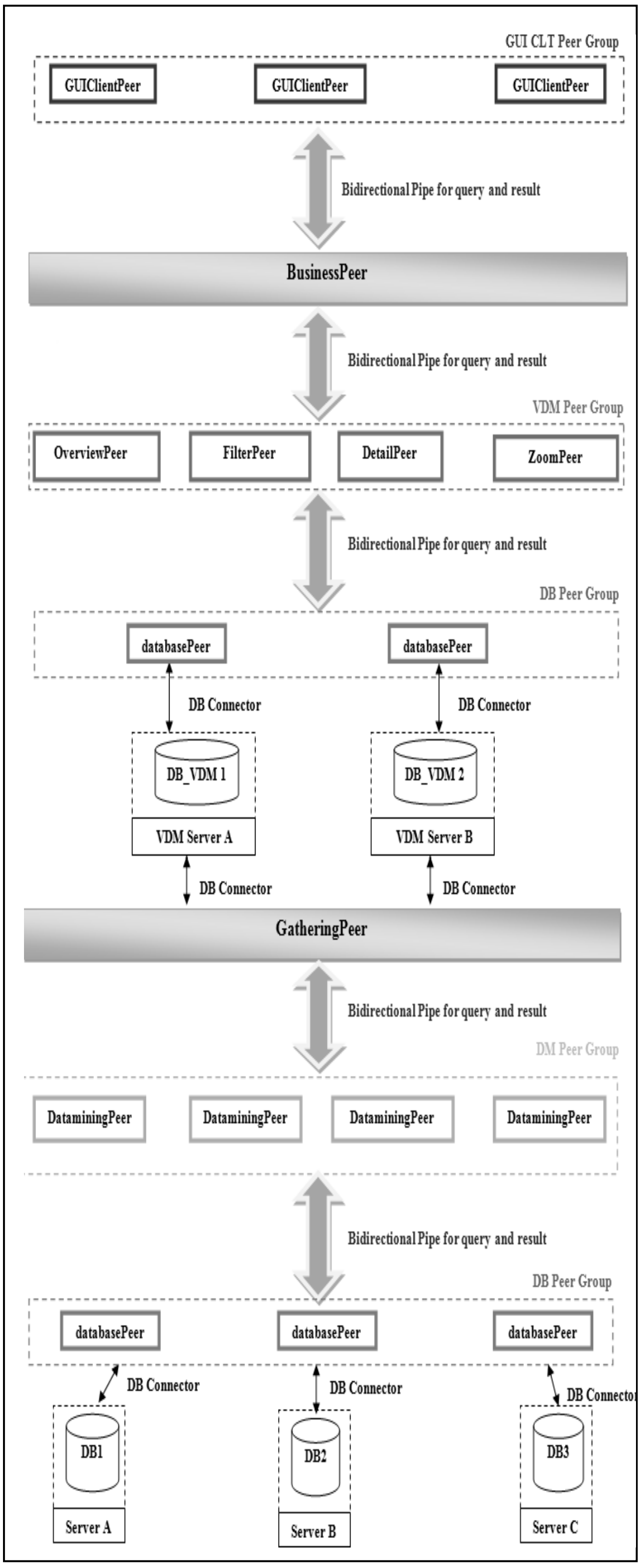


Figure 2. The basic architecture of VDMD framework.

Until now, we did not implement the proposed framework. Our plan is to test it with different data sets and data mining techniques then compare it to client-server and grid architectures.

V. CONCLUSIONS

The proposed Visual Data Mining framework based on Peer-to-Peer architecture has a modular, extensible design. Our architecture can use various data mining programs [9] and handle a huge amount of data [10]. In our future work, we plan to continue our research as bellow:

- Test and evaluate the P2P architecture with Peersim simulator [11] then compare it with a Client-Server architecture and a grid architecture.
- Implement the P2P architecture with JXTA specifications.

There are major futures challenges developing this framework :

- Handle various structures of data.
- Compare the performances of our architecture with the grid architecture.
- Improve the displaying of the data to the final user.
- Introduce other peer group.

REFERENCES

[1] I. Niskanen and J. Kantorovitch,"Ontology driven data mining and information visualization for the networked home", Research Challenges Infromation Science (RCIS), Fourth International Conference, May 2010, pp. 147-156.

[2] B. Shneiderman, “The Eye Have It: A Task by Data Type Taxonomy for Information Visualizations”, Visual Languages, 1996.

[3] CRISP-DM: Crosss Industry Standard Process for Data-Mining, 1999.http://www.crisp-dm.org/.

[4] A. Keim Daniel, ” Information Visualization and Visual Data Mining”, IEEE Transactions on Visualization and Computer Graphics, January-March 2002, vol. 8, no. 1.

[5] A. Yahyavi and B. Kemme, “Peer-to-Peer Architectures for Massively Multiplayer Online Games: A Survey”, ACM Computing Surveys (CSUR), vol 46, iss. 1, October 2013.

[6] D. Guedes, W. Meira and R. Ferreira, "A Service-Oriented Architecture for High-Performance DataMining", IEEE Internet Computing , , July-August 2006, vol 10, No 4, pp. 36 – 43.

[7] X. Zhang H-F. Wong, and W.Cheung, "A Privacy-Aware Service-oriented Platform for Distributed Data Mining", Proceedings of 3rd IEEE Conference on Entreprise Computing, E-Commerce and E-Services (EEE’06), San Francisco, California, June 26-29 2006, pp. 44-48

[8] M. Ankerst, D.A. Keim, and H. Kriegel, “‘Circle Segments’: A Technique for Visually Exploring Large Multidimensional Data Sets”, Proc. Visualization ‘96, Hot Topic Session, San Francisco, CA, 1996.

[9] E. Kandogan, “Visualizing Multi-Dimensional Clusters, Trends, and Outliers using Star Coordinates”, Proc. ACM SIGKDD ’01, 2001, pp. 107-116.

[10] U. Fayyad, U. Piatetsky-Shapir, and G. Smyth, "The KDD process for extracting useful knowledge from volumes of data”, Communications of the ACM, vol. 39, iss. 11, Nov 1996, pp. 27-34.

[11] http://peersim.sourceforge.net/, retrieved: October, 2014.

Dynamic Systems State Model for E-Readiness Estimation

An evaluation of the e-readiness layered model

Najib Belkhayat
LAMAI, Cadi Ayyad University
Management School: ENCG - M
Marrakech, Morocco
najib.belkhayat@gmail.com

Aziz Doukkali, Boubker Regragui
TIES, Mohamed V University
IT engineering School: ENSIAS
Rabat, Morocco
regragui@ensias.ma, doukkali@ensias.ma

***Abstract*—The E-readiness Layered Model (ELM) tends to cover the lack of indicators measures for the e-readines index calculation. This paper's aim is to test the effectiveness of this model through the comparison of an ELM estimted index and a calculated index. Indeed, European I2010 indicators' measurements database is used in this test and the comparison of data series shows a high correlation between the two indexes (Pearson and Spearman coefficients are above 0.95). Also, the two main differences between the data series are analysed in the light of two other international e-readiness initiatives (the Economist Intelligence initiative Unit one and the International Telecommunication Union one). This test shows that ELM could lead to estimate the e-readiness index using 6 indicators instead of 20 (70% less).**

Keywords-e-readiness; dynamic system modeling; state model; e-readiness layered model; Spearman coefficient.

I. INTRODUCTION

The world discovered in 2011 the importance of composed indexes through the powerful financial rating agencies, such as Standard & Poors and Moody's. Indeed, the famous "AAA" rating for countries and financial institutions solvency can influence their economic stability. Similarly, the Shanghai ranking participates in guiding students towards the highest-ranked universities; the backwardness of the French universities in this index led to the creation of a parliamentary commission for the development of education in France.

In this international context, the e-readiness indexes are an important approach for measuring and developing the integration of new technologies in countries and regions. These indexes are important as they contribute to the governments' e-Strategy assessment and the orientation of Foreign Direct Investment (FDI) to different countries (investment in offshore software development, in call centers, etc.). However, the availability of Information and Communication Technologies (ICT) indicators' measures is one of the main issues for the development and spread of this concept.

To address this issue, Belkhayat et al. [1] proposed an E-readiness Layered Model (ELM) to reduce the measurements' need in the e-readiness estimation process. Belkhayat et al. [11] proved the convergence of the ELM and tested its pediction accuracy for one indicator. This paper evaluates the ELM capacity to reduce the indicators measurements' need for the e-readiness index estimation.

Indeed, Section 2 of this paper introduces the main issues related to the excessive measures need to calculate e-readiness indexes. Section 3 gives a brief presentation of the layered approach proposed in the literature to overcome this problem. In Section 4, we precise the testing approach adopted and the results observed.

II. EXCESSIVE NEED OF INDICATORS MEASURES: MAIN OBSTACLE TO THE E-READINESS CONCEPT DEVELOPMENT

The e-readiness tends to measure and analyze the level of integration of ICT in countries' development. The measurement of e-readiness is a global index based on the compilation of a set of indicators correlated to ICTs. This concept of e-readiness has gained importance globally within last decade. Indeed, the action plans of the two World Summits on the Information Society (Geneva 2003 and Tunis 2005) called for the periodic evaluation and comparison of international performance in the field through a composite index comprising comparable statistical indicators [2]. To this end, several organizations, such as International Telecommunication Union (ITU), World Economic Forum (WEF), the Economist Intelligence Unit (EIU), have focused on this concept and have developed specific approaches for the e-readiness measurement and its exploitation in guiding the development of new technologies in different countries and regions. Two main objectives are reported for these approaches: ICT integration evaluation and different countries ranking according to a composite index [3].

Thus, each of these e-readiness approaches relies on a set of indicators that reflect a vision of the ICTs role in countries' development. These indicators can be related to the digital infrastructure, ICTs human resources, and also other areas that may be impacted by new technologies such as governance or economics. At this stage, the number of indicators considered by each approach may range from a dozen (11 approach ITU) to more than a hundred ("e-readiness ranking" developed by IBM and EIU [4][5][6][7].

Thus, the calculation of an e-readiness index requires thousands of indicators measurements (the number of indicators multiplied by the number of countries in the panel). Also, to get an accurate picture, these measures should be carried out synchronously and in a relatively

ISBN: 978-1-61208-381-0

small window of time. The UN report [7] presents the latter as a challenge: a window of 30 days was set in the approach used to measure the e-Government index, but the realization could not be done in less than 75 days due to the complexity of collecting and processing indicators measurements.

Faced with the lack of measures, regular e-readiness evaluation initiatives resort to reducing the indicators framework or forcing the measurements database completeness using measurements from previous years or estimating the missing measurement mainly by the hot deck method (expert comparison to other similar countries) [8]. However, the reduction of the indicators framework impacts the relevance of the index while the missing data estimation is difficult in the ICT field due to the unavailability of large historical databases and the frequent change of the indicators definition (necessary changes to take account of technological developments and societal phenomena related to ICT (4G-3G - DSL, social networks, e-learning, etc.) [9].

On the other hand, measurement quality differs from one country to another depending on the data source, the survey methodology or even the perception of people who respond to these inquiries.

Davidrajuh produced an attempt to address the first issue through fuzzy logic [10]. To cope with the problem related to the exponential number of inference rules needed, this attempt was based on the aggregation of e-readiness indicators in a smaller set of "key indicators". However, the author does not specify the validation process and the results of this approach tests. We have not found any other related work.

III. LAYERED APPROACH AND STATE MODEL FOR PREDICTING THE E-READINESS INDICATORS MEASURES

We proposed an approach for modeling the e-readiness indicators evolution considering them as measurable characteristics identifying a “virtual" dynamic system [1]. This paper focused on a layered modeling approach and cross indicators impactability concept leading to a dynamic system state model. Indeed, the indicators are classified into three layers (basic layer, intermediate layer and target layer) according to the importance they represent to the e-Strategy process owner which is the highest authority of the state or country:

Basic Layer: featuring the basic indicators that can be of two types: basic indicators on which it is possible to act by decision (example: Cellular phone network coverage or the rate of research and development (R&D) budget to Gross Domestic Product (GDP) can be directly impacted by government decision) or general prerequisite indicator part of a wider area than that studied (example: illiteracy rates).

Target layer: featuring indicators that represent a development goal, e.g., rate of e-business GDP to GDP.

Intermediate layer: having intermediate indicators that are neither basic indicators nor target ones. These indicators generally represent milestones that help ensure the smooth progress of projects but are not final goals in themselves, e.g., percentage of the population using the Internet.

Thus, this classification is based on an assimilation of e-readiness to a dynamic system whose state is characterized by the chosen indicators framework and where the basic indicators represent the system control levers. This design leads to a state model linking indicators variations over time:

$$EV_{k+1} = (RIM_k)^T * EV_k + C_k \qquad (1)$$

where EV_k is the indicators measures evolution vector for the period k (period between T_{k-1} and T_k); C_k is a constant vector corresponding to the basic indicators evolutions planned by the e-Strategy stakeholders for the next period and $(RIM_k)^T$ is the transition matrix calculated on the basis of previous measures and an indicators’ impactability matrix [11].

Belkhayat et al. used Kalman filtering for testing the ELM convergence and the indicators prediction accuracy [11].

Thus, the ELM state model suggests that the knowledge of current measures and the planed variations in basic indicators (C_k commands) allows us to predict the other indicators variations and then the reconstitution of the next system state vector.

ELM is a new indicators’ measurements prediction approach used in the e-readiness field where the unavailability of large historical databases and the frequent change of the indicators definition are limiting the use of other known prediction methods used in the conventional sectors [9]. We could estabilish the convergence of the state model and the evaluation of the accuracy of the indicators’ estimations in comparison with measured values [11]. In this paper, we evaluate the accuracy of the e-readiness index based on the prediction model in comparison with the other e-readiness indexes measured and estimated by other institutions. The methodology and results of this evaluation are the subject of the next section.

IV. ELM EVALUATION: LAYERED MODEL APPLIED TO THE I2010 INITIATIVE

A. Materials and methods

We used the I2010 database [12] to test the efficiency of ELM through the evaluation of its potential reduction of measures need to calculate e-readiness using the predictive model. This database fits to our test needs because it is containing regular yearly measurements for a significant number of countries and indicators (29 countries and 52 indicators).

The testing approach relies on the comparison of two e-readiness indexes: the one calculated on the basis of all 2010 measures and the one calculated on the basis of 2010 basic indicators measures and the ELM state model. Indeed, the use of 2008 and 2009 measures, as well as the 2010 basic indicators measures, leads to the prediction of the other indicators variations; then, one can calculate the 2010 index based on the predictive model.

ISBN: 978-1-61208-381-0

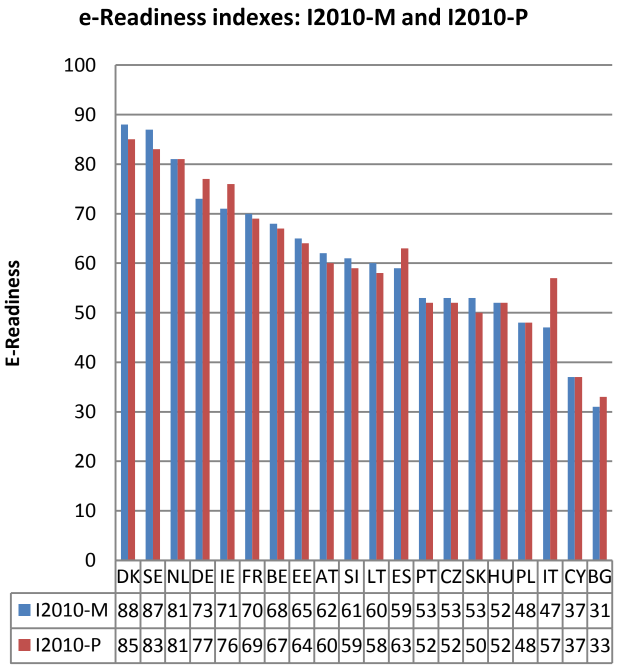
To execute this approach, we began by reducing the database to keep only the indicators and the countries for which we have all 2008-2009-2010 measures. The reduced database included 20 countries and 20 indicators.

The aim of the second step was to classify indicators according to the ELM three layers. The third step provided the necessary inputs to calculate the transition matrix of the state model (impactability matrix and calculation of indicators optimal values). Appendix 1 contains these elements that are the same used in [11] for the model convergence validation tests.

The last step concerned the estimation of the indicators variations and the calculation of the two indexes to compare. In the following, we note « I2010-P » the index calculated on the basis of predictions and « I2010-M » the index calculated on the basis of all 2010 measures.

B. *Results and discussion*

Figure 1 shows the two indexes for the 20 countries of the panel. We based the correlation analysis of these two data series on the Pearson correlation coefficient which “is a measure of the linear association between two variables using quantitative data”, and the Spearman correlation coefficient which is a “correlation measure of association between two variables when ordinal or rankordered data are available” [13]. The two data series correlation analysis gives us 97.5% for the Pearson correlation coefficient and 95.3% for the Spearman correlation coefficient (countries ranking correlation). These correlation coefficients values (close to 100%) denote that the two data series are highly correlated.



	DK	SE	NL	DE	IE	FR	BE	EE	AT	SI	LT	ES	PT	CZ	SK	HU	PL	IT	CY	BG
🟦 I2010-M	88	87	81	73	71	70	68	65	62	61	60	59	53	53	53	52	48	47	37	31
🟥 I2010-P	85	83	81	77	76	69	67	64	60	59	58	63	52	52	50	52	48	57	37	33

Figure 1: E-readiness 2010 – Measures # Predictions.

Also, the graphic analysis shows two main issues: the first is related to the measures and ranking deviations of Spain and Italy and the second is related to the tighter spreads between the five countries of the leading group: Denmark, Sweden, Netherlands, Germany and Ireland. To figure out these issues, we made comparisons of our two indexes among the other main e-readiness indexes measured in 2010: « e-readiness Ranking 2010» published by the EIU and IBM and IDI (ICT Development Index) published by the ITU (International Telecommunication Union).

1) *First issue: Italy and Spain indexes and ranking differences*

Italy and Spain indexes have significant differences that affect the ranking. Indeed, if Italy and Spain are removed from the data series, Figure 2 includes the two indexes ranking differences of the other 18 countries (Denmark being the leader).

Figure 2 shows that the countries ranking does not change except for Slovakia and Hungary, which had a small gap of about 1% and exchanged their position by passing Slovakia behind Hungary.

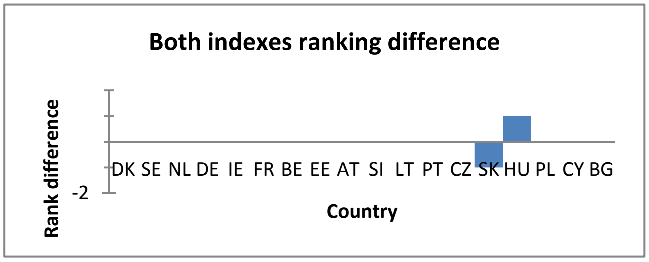


Figure 2: Countries ranking difference between I2010-P and I2010-M.

To understand this difference, we gathered the ranking of the two countries in the two other e-readiness indexes mentioned above. Hungary exceeds Slovakia of about 3% in the "e-readiness Ranking 2010 report”. In IDI, Hungary exceeds Slovakia of about 2%. These findings confirm that the index « I2010-P » complies with the two countries ranking in the other indexes measured in 2010.

Regarding Italy and Spain rankings, we find that Spain moves from the 12th position to the 9th position winning three ranks and Italy goes from the 18th to the 13th position winning five ranks. Also, we gathered their respective rankings (in the panel of studied countries) in the other published indexes. Table I shows the ranking of the two countries following the four indexes under comparison:

TABLE I. ITALY AND SPAIN RANKING IN THE FOUR INDEXES

	I2010-M	I2010-P	EIU-IBM	IDI (ITU)
Spain	12	9	9	10
Italy	18	13	11	12

Table I confirms that the index « I2010-P » complies with the two countries ranking in the other indexes measured in 2010. This allows us to estimate that the

ISBN: 978-1-61208-381-0

ranking of Spain and Italy according to I2010-P is more accurate than their ranking in I2010-M.

2) Second issue: The tighter spreads between the leading group countries

The distance between the first and the fifth of the ranking frontrunners is tighter in I2010-P. Indeed, the gap decreases from 18 points in I2010-M (more than 20% of the leader index) to 9.5 points in I2010-P (~ 12% of the leader index). In comparison with the other indexes, this difference is about 8% in the "e-readiness Ranking 2010" and 15% in IDI. This allows us to confirm that I2010-P is more accurate than in I2010-M regarding this issue.

The accuracy of I2010-P is clearer if we consider only Denmark, Sweden, Netherlands and Germany. Indeed, the amplitudes of the indexes for these four countries are in Table II as follows:

TABLE II. THE AMPLITUDE OF THE DISTANCE BETWEEN THE FOUR FRONTRUNNERS (BY INDEX)

	I2010-M	**I2010-P**	**EIU-IBM**	**IDI (ITU)**
amplitude	16,8%	9%	8,1%	11,6%

This table shows that the amplitude of I2010-P is within the same range than those of EIU and IDI while the amplitude of I2010-M is more than two times the EIU one.

V. CONCLUSION

The lack of ICT indicators measures is hindering the e-readiness (e-strategy) concept development (e-readiness is the first phase of the e-strategy process). ELM approach attempts to overcome this problem through a predictive system based on dynamic systems state models. This paper has attempted to assess the contribution of this approach by comparing an index calculated on the basis of 20 indicators measures and an index calculated on the basis of six basic indicators measures and the predictive model (used input tables for the ELM model are presented in the appendix 1).

The Pearson and Spearman correlation coefficients of the results exceed 95%. This ensures that the two indexes are highly correlated. On the other hand, the graphics exploratory analysis revealed two main differences between the two compared indexes. The results comparison with other e-readiness indexes measured in 2010 showed a greater consistency for the index based on the predictive model. Thus, this test assumes the robustness of the predictive model based on the ELM while allowing a significant advantage through the reduction of indicators' measures need of 70% (from 20 to 6 indicators).

However, it should be remembered that the predictive model requires the availability of the two previous periods' measurements. This leads us to conclude that the ELM approach can be very useful for initiatives that plan to last over time (annual rankings). Indeed, these initiatives can invest in exhaustive measures of the first two years and then use the ELM approach to decrease the need of indicators measurements. Also, the ELM can be used as a simulator to assist in the e-strategy fomulation (based on e-readiness diagnostic) through the prediction of the impact of planed actions on basic indicators. Indeed, the stakeholders can estimate the impact of their action plan before its validation.

Finally, we find that the ELM approach is not specific to e-readiness and can be exploited in other sectors like human development where indicators definitions can change over time.

REFERENCES

[1] N. Belkhayat, A. Doukkali, and B. Regragui, "E-Strategy formulation: a new approach based on a layered model", JATIT, ISSN: 1992-8645, vol. 36, 2012, pp. 101-112.

[2] International Telecommunications Unit, "Measuring the Information Society: The Information Development Index 2009", 2009.

[3] Bridges.org (2005), E-Ready for What? E-Readiness in Developing Countries: Current Status and Prospects toward the Millennium Development Goals, Report prepared for Infodev. http://www.infodev.org, [Retrieved: December, 2014].

[4] INSEAD – World Economic Forum, "The Global Information Technology Report 2013", 2013.

[5] International Telecommunications Unit, "Measuring the Information Society: The Information Development Index 2013", 2013.

[6] The Economist Intelligence Unit Limited and IBM Corporation, The 2010 e-readiness rankings. Report, 2010.

[7] United Nations, (2010), "United Nations E-Government Survey 2010", 2010.

[8] International Telecommunications Unit, "Measuring the Information Society: The Information Development Index 2011", 2011.

[9] O. Gärdin, "The New Economy New challenges for the statistical system", The International Association for Official Statisticians Conference, 2002, London, 2002.

[10] R. Davidrajuh, "Building a fuzzy logic based tool for e-readiness measurement", Electronic Government, An International Journal, vol. 5, no. 2, 2008.

[11] N. Belkhayat, A. Doukkali, and B. Regragui, "e-readiness: a novel approach for indicators measurements estimation and prediction", JATIT, vol. 69, n°3, 11.2014, pp. 617-631.

[12] http://ec.europa.eu/digital-agenda/download-data, [Retrieved: December, 2014]

[13] D. R. Anderson, D. J. Sweeney, T. A. Williams, Statistics for Business and Economics. Eleventh edition, 2011.

[14] M. Freudenberg, "Composite Indicators of Country Performance: A Critical Assessment", OECD Science, Technology and Industry Working Papers, 2003/16, OECD Publishing. 2003.

[15] European Union Commission - Commission of the European Communities, "Europe's Digital Competitiveness Report - ICT Country Profiles", 2010.

[16] European Union Commission - Commission of the European Communities, "Priorities for a new strategy for European information society 2010-2015", 2010.

APPENDIX 1

The optimal values in Table III are estimated on the basis of the measured maximum value and a potential evolution margin [11].

TABLE III. SELECTED SET OF INDICATORS, THEIR CLASSIFICATION AND OPTIMAL VALUES

N°	Indicator	Layer*	Opt V
1	% of population doing an online course (in any subject)	T	9
2	% of population interacting online with public authorities	T	75
3	% of enterprises interacting online with public authorities	T	100
4	Total electronic sales by enterprises, as a % of their total turnover	T	24
5	% of enterprises using any computer network for sales (at least 1%)	I	30
6	% of population who are regular internet users (at least once a week)	I	95
7	% of population looking for information about goods and services online	I	90
8	% of population looking online for a job or sending a job application	I	28
9	% of population looking online for information about education, training or course offers	I	40
10	% of enterprises submitting a proposal in a public e-tender (e-procurement)	I	23
11	% of population ordering goods or services online	I	74
12	% of population selling goods or services online (e.g. via auctions)	I	23
13	% of enterprises using any computer network for purchases (at least 1%)	I	57
14	Fixed broadband penetration	B	50
15	% of households having a broadband connection	B	100
16	% of enterprises having a fixed broadband connection	B	100
17	% of households with access to the Internet at home	B	100
18	% of population using online banking	B	100
19	% of basic public services for citizens, which are fully available online	B	100
20	% of basic public services for enterprises, which are fully available online	B	100

* T: Target; I: Intermediate; B: Basic

Table IV bellow contains the indicators Impact Matrix proposed in [11]. The firt line of this matrix assumes that the target indicator N°1 is impacted by the indicators N° 9, 11, 15 and 18. The value in a cell corresponds to the relative importance of the corresponding impact: the indicators N° 9 and 15 have a greater impact on the indicator N° 1 than the indicators N° 11 and 18.

ISBN: 978-1-61208-381-0

TABLE IV. SELECTED INDICATORS IMPACT MATRIX WITH RELATIVE WEIGHTS

N°	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
1	0	0	0	0	0	0	0	0	5	0	3	0	0	0	5	0	0	2	0	0
2	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	5	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	1	0	5
4	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	0	0	5	0	0
5	0	0	0	0	0	5	0	0	0	1	3	0	5	0	0	0	0	3	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	5	0	1	0
7	0	0	0	0	3	5	0	0	0	0	0	0	0	0	0	0	0	0	1	0
8	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	4	0	1	0
9	0	0	0	0	0	0	0	5	0	0	1	0	0	0	0	0	5	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	3	0	1
11	0	0	0	0	5	0	3	0	0	0	0	1	0	0	0	0	0	5	0	0
12	0	0	0	0	0	0	0	0	0	0	5	0	3	0	0	0	0	3	0	0
13	0	0	0	0	5	0	0	0	0	0	0	1	0	0	0	2	0	3	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...																				
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

ISBN: 978-1-61208-381-0

A Comparison of Data Mining Techniques for Anomaly Detection in Relational Databases

Charissa Ann Ronao, Sung-Bae Cho
Computer Science Department
Yonsei University
Seoul, South Korea
cvronao@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Abstract—**Data mining has gained a lot of attention in recent years especially with the advent of big data. In line with this, relational database management systems (RDBMS) have also become the ultimate layer in preventing malicious data access. However, despite the presence of traditional database security mechanisms, it is apparent that database intrusions still occur. Thus, there is an imminent need in developing a robust and efficient intrusion detection system (IDS) especially tailored for databases. Among the few studies that have been published with regards to the problem at hand, most researchers have proposed the use of data mining techniques to detect database anomalous behavior. However, up to this date, there has been no work aimed to objectively compare these various data mining techniques as applied to the field of database IDS. In this paper, we evaluate the state-of-the-art feature selection and data mining algorithms in the context of database IDS and provide a clear performance comparison of these techniques under common grounds. Experiments show that principal components analysis produces a reasonably compact and meaningful subset of features while graphical models like decision trees, random forest, and Bayesian networks yield a consistently high performance in detecting anomalies in databases.**

Keywords-intrusion detection; anomaly detection; database security; data mining; analysis.

I. Introduction

In today's information revolution era, data has become more and more indispensable to individuals, companies and organizations. This paved the way to developing relational database management systems (RDBMS), which can organize, contain, and protect these data from malicious threats. However, despite access controls and firewalls that are widely incorporated in these systems, it has been found that they are inadequate in defending against anomalous attacks. Moreover, network-based and host-based intrusion detection systems (IDS), although having been extensively researched and implemented in recent years, are awfully insufficient and unsuitable in detecting attacks specifically targeted to databases [1]. In particular, insider threats are as much of a concern as outsider threats, i.e., privileged users, if corrupt, can potentially cause more damage than average users. While many works have focused on how data can be protected from external attacks, there have been very few researches regarding the problem of protecting data from insider threats [2]. Because of this, there has been a growing awareness that a strong and effective IDS especially tailored for databases needs to be developed.

An efficient and robust intrusion detection mechanism is crucial in building a strong database security framework. In line with this, a number of data mining techniques have been proposed to perform this task [3]. Although previous works have integrated data mining algorithms in their IDS framework, to the best of our knowledge, none of these works have performed an in-depth evaluation and performance comparison of data mining algorithms in the context of database intrusion detection. To address this, this paper provides a clear comparison and parallel evaluation of state-of-the-art data mining methods in the application of database IDS. We mine SQL query logs and exploit the presence of role-based access control (RBAC) mechanism, which has already been adopted in various commercial RDBMS products [4], to detect anomalies. We model normal access behavior through these queries along with their corresponding role annotations, and detect anomalies by tagging queries that deviate from these normal access behaviors.

The rest of the paper is organized as follows: Section II describes the related work, followed by the discussion of system architecture, feature extraction, state-of-the-art feature selection methods, and data mining techniques in Section III. Section IV presents our experiment results, and finally, we draw our conclusion in Section V.

II. Related Work

IDS's are generally divided into two main categories: signature-based and anomaly-based. Signature-based or misuse-based systems make use of explicitly defined attack signatures and detect intrusions by blacklisting. This kind of system is ineffective in the face of new types of attacks, which, in turn, makes it susceptible to evasion methods that take advantage of the expressiveness of the SQL language [5]. On the other hand, anomaly-based systems model normal behavior in the form of intrusion-free logs and marks deviations from this normal behavior as anomalies [6]. Unlike the former, these systems are clearly more robust to unknown attacks and to malicious users who may keep on evolving their attack strategy.

One of the most common way of implementing an anomaly-based IDS is by exploiting data mining algorithms as the detection mechanism. In the past decade, a number of data mining techniques have been proposed for the purpose of detecting intrusions in databases. Among these are the use of data dependency and association rules [7][8][9]. Such methods, however, require the manual assignment of attribute weights; they also cannot be scaled easily to typical database sizes [3]. Another technique was proposed by Barbara et al. [10], who made use of hidden Markov models (HMM) to capture the change in database normal behavior over time. This too, however, is impractical to implement in large databases with many tables and attributes. Consequently, Ramasubramanian et al. integrated artificial neural networks (ANN) into their proposed IDS framework [11], while Pinzon et al. made use of support vector machines (SVM) and multilayer perceptrons (MLP) to detect outsider attacks [12]. These papers have focused mainly on the structure development of the database IDS framework, and did not sufficiently evaluate the underlying core mechanism, which is the data mining technique. Furthermore, Kamra et al. proposed an IDS which exploits a naïve Bayes (NB) classifier to detect abnormal behavior [4]. The latter had based their approach on the RBAC model, a standardized access control mechanism, building a profile for each role, and checking the behavior of each role with respect to the profile [14]. The main idea is to assign one or more roles to each user, and assign privileges to roles. This effectively minimizes the number of profiles to maintain, which makes it scalable to a large database user population, and is a much more efficient method compared to managing a profile for each individual user. We adopt the same rationale and build normal profiles through roles and SQL query access.

We stress, however, that all mentioned works lack the necessary evaluation step of analyzing the features they have extracted and comparing their proposed data mining approach to other state-of-the-art techniques. We believe that merely applying an algorithm to the problem and showing its satisfactory results are not enough to prove the effectiveness and efficiency of the system—a clear comparison and parallel evaluation must be made to know how these algorithms perform in detecting intrusions under common grounds, most especially, in the data mining perspective.

III. Data Mining for Database Intrusion Detection

We exploit the existence of the RBAC mechanism and model normal access behavior profiles through roles. Normal access behavior is represented by intrusion-free SQL queries, and they are used to train a data mining algorithm to produce normal profile models. We define an anomaly as an access behavior that deviates from these normal profiles. Given these profiles, clearly, we have a standard classification problem.

A. Intrusion Detection System

Figure 1 shows the intrusion detection process. Every time a query is issued, the profile logs are updated. During the training phase, normal access behavior, in the form of SQL queries grouped into profiles, are fed to the feature extractor, feature selector, and finally, the data mining algorithm or classifier; the classifier then produces a trained model out of normal access behavior. During the detection phase, each new query goes through the feature extractor and selector, and is evaluated by the trained classifier. An alarm is raised if the query deviates from normal profiles. We emphasize that role profiles should be regularly updated and classifier training periodically done, so as to be able to update the normal profile models and minimize false alarms.

Given this setup, there are three main problems: (1) how to extract and represent features, (2) which of these features to use, and (3) which data mining technique to employ. We discuss the solutions to these three problems in the following section.

B. SQL Query Parsing and Feature Extraction

One SQL query corresponds to an entry in the database log file, which follows the SQL language syntax. For simplicity, we illustrate the SQL grammar with the SELECT command:

```
SELECT     <Projection attribute clause>
FROM       <Projection relation clause>
WHERE      <Selection attribute clause>
ORDER BY   <ORDER BY clause>
GROUP BY   <GROUP BY clause>
```

We parse queries in this manner, line-by-line, and extract features accordingly in order to transform SQL log entries into feature vectors that can be understood and processed by data mining classifiers.

We gather proposed features from database IDS literature and combine them to form a more complete feature set [4][12][15]. We represent a query as a feature vector Q with seven fields: Q`(SQL-CMD[], PROJ-REL-DEC[], PROJ-ATTR-DEC[], SEL-ATTR-DEC[], ORDBY-ATTR-DEC[], GRPBY-ATTR-DEC[], VALUE-CTR[])`, as seen in Table I. Query mode, c, represents the query commands SIUD: if the query command is `SELECT`, it is represented by integer 1; if `INSERT`, integer 2; if `UPDATE`, integer 3; and if `DELETE`, integer 4. Query length, Q_L, is denoted by the number of characters in the whole query, including spaces. The number of string values, S_V, and numeric values, N_V, indicate how many times these values appear in the selection clause. The same logic is applied with the number of JOINs, J, and ANDs/ORs, AO.

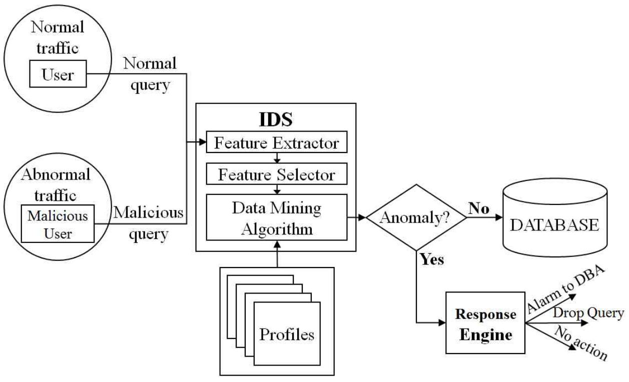


Figure 1. Flow of the IDS process.

TABLE I. LIST OF EXTRACTED FEATURES

Vector field	Description	Feature elements
`SQL-CMD[]`	Command features	query mode, c query length, Q_L
`PROJ-REL-DEC[]`	Projection relation features	Number of projected relations, P_R Position of projected relations, P_{RID}
`PROJ-ATTR-DEC[]`	Projection attribute features	$(P_A, P_A[], P_{AID}[])^a$
`SEL-ATTR-DEC[]`	Selection attribute features	$(S_A, S_A[], S_{AID}[])^a$
`ORDBY-ATTR-DEC[]`	ORDER BY clause features	$(O_A, O_A[], O_{AID}[])^a$
`GRPBY-ATTR-DEC[]`	GROUP BY clause features	$(G_A, G_A[], G_{AID}[])^a$
`VALUE-CTR[]`	Value counter features	Number of string values, S_V Length of string values, S_L Number of numeric values, N_V Number of JOINs, J Number of ANDs and ORs, AO

a. Convention (N_A, $N_A[]$, $N_{AID}[]$):
N_A – number of attributes in a particular clause
$N_A[]$ – number of attributes in a particular clause counted per table
$N_{AID}[]$ – position of the attributes present in a particular clause, represented in decimal

In addition, the number of relations, P_R, indicates how many tables are present in a specific clause. The position of relations, P_{RID}, is represented by a binary string, wherein each bit stands for a table in the database schema. If a table is present in the query, its bit representation is 1; if it is absent, it is represented by bit 0. Wu et al. stated that different input encoding schemes (binary or decimal) result to different algorithm performance results [16]. Decimal encoding was found to be more robust to noise and decreases computational complexity; thus, to get the final value of the ID feature, we convert the binary string into its decimal form. The same logic is applied to the mapping of the positions of attributes given a specific clause. Thus, all ID features are represented by a single decimal value.

Extending the parsing and feature extraction method to other commands such as `INSERT`, `UPDATE`, and `DELETE` is clearly straightforward. A total of 21 main features are extracted for every query in the SQL log, with some features (e.g., ID features) branching out to sub-features that depend on the number of tables and attributes in the database schema. For example, for a schema consisting of 2 relations with 4 attributes each, the resulting number of features will be 45.

C. Feature Selection Methods

Selecting good feature sets improves performance, eliminates noise, and enables faster and more accurate detection [17]. We use five feature selection methods to evaluate the extracted features, and they are categorized into two groups: ranking methods and filter methods.

Ranking methods output the complete feature set sorted from highest to lowest according to a certain evaluation measure. Since the top variables are considered to be the most discriminant features, a certain threshold should be determined to cut off features that are considered to have little or no contribution to the classification process. One of the most common evaluation measure when ranking features is information gain (IG). It is the expected reduction in entropy caused by partitioning a query data set according to a certain feature. Given a query data set S with K different roles/classes, entropy is given by:

$$I(S) = -\sum_{k=1}^{K} \frac{|s_k|}{|s|} \log \frac{|s_k|}{|s|}, \quad (1)$$

where s is the total number of queries in the data set and s_k is the number of queries in class k. We get the IG of feature Y which can partition S into M subsets by,

$$IG(Y) = I(S) - \sum_{m=1}^{M} \frac{|s_m|}{|s|} I(S_m), \quad (2)$$

where the second term is the conditional entropy, $I(S|Y)$, and s_m is the number of queries in subset m.

An improved variant of IG is gain ratio (GR), which overcomes the bias of the former towards features that can have a large number of possible values. GR applies a kind of normalization to IG by using the information value corresponding to M outcomes on feature Y, i.e.,

$$IV(S\,|\,Y) = -\sum_{m=1}^{M} \frac{|s_m|}{|s|} \log \frac{|s_m|}{|s|}. \quad (3)$$

Dividing (1) by (3) gives the GR of feature Y.

Principal components analysis (PCA) is an unsupervised ranking feature selection technique which can transform query data set S into a new coordinate system and produce a set of components $p \in P$ wherein the top components, called principal components (PCs), represent the greater part of the variance of S. With this, we can easily eliminate the tailing p's (those that does not contain much of the variance of the S). Scaling and standardizing is often applied before PCA to simplify the latter's calculation.

In contrast, filter methods automatically output a set of chosen features based on a certain evaluation measure. One of these methods, best first search (BFS), is a combination of forward selection and backward elimination which can greedily search through the query feature space. In the case when performance starts to drop, it can backtrack previous feature subsets (those with good enough performance) and start again from there. However, for a high dimensional query data set S (which depends on how big the database schema is), BFS can be computationally expensive.

Genetic algorithm (GA) is another filter method based on the principle of natural selection, which randomly creates a population N of possible feature subsets n (any combination of fields from Q) and evaluates each one by a certain measure (e.g. correlation). GA runs for several generations, each time creating a new N by performing crossover and mutation. This method has been proven to be very effective in practice [11].

D. Data Mining Algorithms

We consider the following state-of-the-art classifiers which have been successfully applied in the intrusion detection domain, namely: naïve Bayes, K-nearest neighbors, artificial neural networks and multilayer perceptrons, support vector machines, Bayesian networks, J48 decision trees, and random forest [4][11][12][13][19][20][21].

Naïve Bayes (NB) is a simple classifier with strong feature independence assumptions. Given a new query $q \in Q$ with a set of features $Y = \{y_1,\ldots,y_d\}$, and k roles/classes, we compute the posterior probability of class membership, i.e., the probability that Y belongs to role r_k, by,

$$p(r_k \mid Y) \propto p(r_k)\prod_{i=1}^{d} p(y_i \mid r_k). \tag{4}$$

Using (4), we can classify q into a role r_k that achieves the highest posterior probability.

Another method based on Bayes theorem is Bayesian network (BN), a probabilistic graphical model represented by a directed acyclic graph, wherein nodes signify the query features and edges represent the dependencies among them. A BN is learned by obtaining the log-likelihood, which is the probability of the data given the network, i.e.,

$$\log L(\Theta \mid Q) = \sum_{e}\sum_{d} \log p(q_{yi} \mid \pi_i, \theta_i), \tag{5}$$

where e is the number of queries in Q, q_{yi} is a feature instance of q_e, π_i is the set of parent nodes of node y_i, and $\theta_i \in \Theta$ is $p(y_i| \pi_i)$.

Artificial neural network (ANN) is a computational model based on the concept of human biological neurons. Weights between the so-called neurons, or nodes, are learned based on the query feature inputs; learning is done with the use of gradient descent and backpropagation algorithm. Multilayer perceptron (MLP) is a feedforward variant of ANN.

Support vector machines (SVM) are based on the concept of maximum margin hyperplanes that define a decision boundary between two classes/roles. They benefit from high dimensional feature spaces; high dimensionality means that there are more possible configurations that can be done in the feature space, which can produce more accurate results.

J48 decision trees are one of the most common techniques in data mining that have been successfully used in various fields. It makes use of tree-like graph decisions, selecting query features for every node based on (2). Although prone to overfitting and feature bias, it can achieve high performance with very little effort.

Accordingly, random forest (RF) is an ensemble model based on decision trees. It exploits bagging and random feature selection to create numerous simple trees to vote for the most popular class/role, and is considered to be better in performance and speed than plain decision trees.

Lastly, K-nearest neighbors (KNN) is an unsupervised classifier that groups new queries based on a distance function. Given a new query q, KNN will find the K nearest query data points with respect to q, the most popular class of the nearest neighbors being the inferred role of q.

IV. EXPERIMENTS

A. Benchmark Database

We have adopted the TPC-E benchmark database schema structure and its transactions for all our experiments. TPC-E is a database that simulates the online transaction processing (OLTP) workload of a brokerage firm [18]. Customers, brokers, and the market initiate read/write and read-only transactions against the database, which consists of 33 tables, an overall count of 191 attributes, and 11 standard transactions.

B. Synthetic Data Set Generation

We treat one TPC-E transaction as one role, and we set privileges of a role based on which tables and attributes the transactions are authorized to access, with the corresponding number of times they appear in the transaction. We emphasize, however, that depending on the context, one role may contain several transactions at once.

We employed the transaction database footprint and pseudo-code found in [18]. Each role has a set of specific tables T (and its corresponding attributes A) that it is allowed to access, as well as a set of commands C that it is allowed to execute. We specify the following probabilities for each role: (1) the probability of using a command $c \in C$ given a role r, $p(c|r)$, (2) the probability of projecting a table $t \in T$ given a command c and a role r, $p(P_t|c,r)$, (3) the probability of selecting a table t given a set of projected tables P_T, command c, and role r, $p(S_t|P_T,c,r)$, (4) the probability of projecting an attribute $a \in A$ given a projected table P_t, command c, and role r, $p(P_a|P_t,c,r)$, (5) the probability of selecting an attribute a given a selected table S_t, command c, and role r, $p(S_a|S_t,c,r)$, (6) the probability of including a random string or numeric value $v \in V$ in the selection clause given a command c and role r, $p(v_{sn}|c,r)$, (7) the probability of including a JOIN J given a command c and role r, $p(J|c,r)$, and (8) the probability of including an AND or OR given a command c and role r, $p(AO|c,r)$.

Note that probabilities 2 to 6 are uniformly distributed among a role's corresponding set of tables T, projected tables P_T, projected table P_t, selected table S_t, and list of random strings and numeric V, respectively. Probability 1 is based on a set of commands C (may compose of any combinations of SIUD) that a role/transaction is allowed to issue. For a certain role, probability 7 means that a query can contain a JOIN or not, while probability 8 may contain a combination of AND and OR, AND only, OR only, or none at all [18].

We generate 1,000 queries for each role, labeling each query with its corresponding class, for a total of 11,000 queries or training samples for our normal query log data set. Since we create the models with insider threats in mind, anomalous queries are generated using the same probability distribution as that with normal queries, only with role information negated. That is, if the role annotation for a certain normal query is class 1, we change it to any other role other than class 1, effectively making it anomalous [4].

C. Results

From this point on, we will refer to features describing the number of elements present in a query as counting features, and those that represent the position of elements as ID features.

The number of features generated largely depends on the number of tables and attributes in the schema. In the case of TPC-E, a total of 277 features were extracted.

Figure 2 shows the average merit based on IG and GR measures. The line indicates the threshold we adopted to get the feature subset for IG and GR. IG produced 12 features while GR produced 144 features. Observing the variables chosen by both measures, IG preferred counting features (those having more possible values), while GR produced a more spread-out merit graph, noticeably preferring pairs of counting by table and ID features while removing string features (S_V and S_L).

We determine the threshold for PCA by plotting the eigenvalues, as shown in Figure 3. Optimal coordinates method produced a subset of 13 features (PCA3), while parallel analysis yielded 63 features (PCA2) [22]. We obtain an additional subset by getting 99% of the variance of the data (113 features, PCA1) for comparison purposes.

For the filter methods, BFS yielded 19 features using correlation as the evaluation measure. Consequently, GA was run for 20 generations with a population size of 20 individuals, crossover rate of 0.6, mutation rate of 0.033 and correlation as evaluation measure. GA chose a total of 68 features, which are noticeably more diverse than the ones chosen by IG, GR, and BFS.

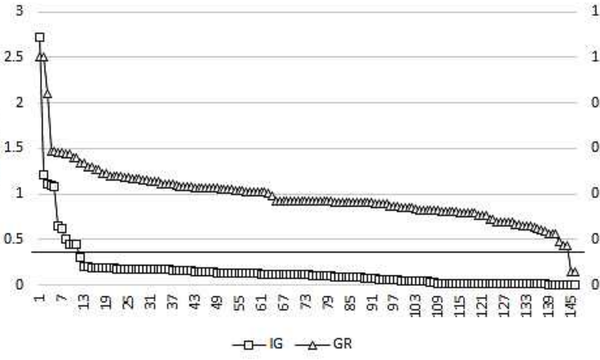


Figure 2. IG and GR values in terms of average merit (y-axis); features (x-axis).

The performances of classifiers in terms of false positive (FP) and false negative (FN) error rates are shown in Table II. False positives are those queries that should have been classified as normal but tagged as abnormal, while false negatives are those that should have been identified as anomalous but were categorized as normal. The Weka toolkit was used in all our experiments and all parameters were left to their default settings [23].

Based on the resulting feature subsets, it can be observed that counting features are vital to obtain a satisfactory classification performance (as seen in the performance of the IG subset). However, they are not enough on their own. PCA came out to be the best feature selection technique among the ones employed—from 277 features, it reduced the data set to 113 features (threshold of 99% variance, PCA1), yielding the overall best average performance. Halving PCA1 to form PCA2 does not have any significant effect on the FP and FN rates, and even when only one-third of PCA2 is retained (PCA3), it still yielded above average performance. This proves that PCA effectively eliminates most of the noise in

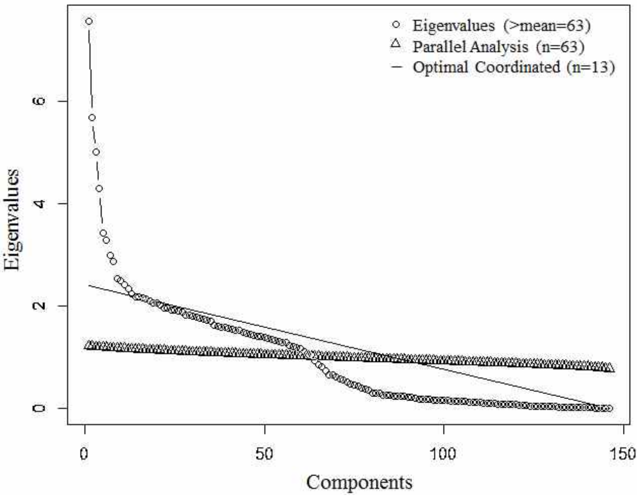


Figure 3. Eigenvalues, parallel analysis, and optimal coordinates plot.

TABLE II. PERFORMANCE OF CLASSIFIERS AND CORRESPONDING NUMBER OF FEATURES IN DECREASING ORDER

No. of features	**GR**		**PCA1**		**GA**		**PCA2**		**BFS**		**PCA3**		**IG**		**Avg.**	
	144		113		68		63		19		13		12			
Classifiers	***FP***	***FN***	***FP***	***FN***	***FP***	***FN***	***FP***	***FN***	***FP***	***FN***	***FP***	***FN***	***FP***	***FN***	***FP***	***FN***
NB	0.153	0.008	0.227	0.014	0.244	0.011	0.248	0.012	0.274	0.017	0.224	0.012	0.389	0.024	0.251	0.014
KNN	0.119	0.006	0.141	0.008	0.128	0.006	0.126	0.007	0.140	0.009	0.107	0.004	0.242	0.012	0.143	0.007
MLP	0.143	0.007	0.079	0.003	0.128	0.007	0.082	0.004	0.166	0.008	0.104	0.007	0.232	0.013	0.133	0.007
SVM	0.574	0.051	0.095	0.004	0.449	0.034	0.103	0.005	0.455	0.034	0.103	0.004	0.485	0.039	0.323	0.024
BN	0.064	**0.002**	0.131	0.007	0.083	0.004	0.168	0.010	0.089	0.005	0.160	0.009	0.097	0.004	0.113	0.006
J48	0.067	0.003	0.113	0.005	0.086	0.003	0.118	0.006	0.092	0.004	0.117	0.006	0.091	0.004	0.098	0.0044
RF	**0.055**	0.003	0.075	0.004	0.079	0.003	0.078	0.005	0.086	0.003	0.079	0.004	0.126	0.006	**0.083**	**0.0039**
Avg.	0.168	0.012	**0.123**	**0.006**	0.171	0.010	0.132	0.007	0.186	0.011	0.128	0.007	0.237	0.015		

the data, at the same time reducing its dimensions significantly. GA is second in line to PCA (in terms of performance and number of features used), followed closely by BFS.

Among the classifiers that we have evaluated, graphical models like J48, RF, and BN noticeably performed better than the other algorithms. This may be due to the fact that SQL language syntax has an inherent tree-like structure—a simple attribute that these classifiers are most likely to exploit. Conversely, SVM yielded the worst performance, producing a satisfactory result only with the PCA feature subsets. It is clear that the application of special kernel methods is necessary to obtain acceptable results with SVM. Moreover, NB is the second worst performer, having yielded the highest FP and FN rates for all PCA subsets. In terms of algorithm speed, SVM and MLP are significantly and impractically slower in build time compared to other classifiers, while J48 and RF yielded the fastest detect times.

V. Conclusion and Future Work

We have shown a clear, side-by-side comparison of data mining feature selection methods and classifiers as applied to the context of database IDS. PCA demonstrated exceptional performance in reducing noise and dimension in the data set, while graphical models, especially RF, came out to be the best suited classifiers for the intrusion detection task, exhibiting very reasonable FP and FN trade-offs and fast detection speed. We hope that these results will provide researchers with a more concrete direction towards designing a more efficient database IDS.

Although we have covered many algorithms in this work, there are still a lot of subjects to explore. Future works will include considering the sensitivity of the tables and attributes in the database. We are also considering on building an ensemble model to be able to develop a stronger classifier out of simple ones.

Acknowledgment

This work is supported by the National Strategic R&D Program for Industrial Technology (10044828), funded by the Ministry of Trade, Industry and Energy (MOTIE)..

References

[1] X. Jin and S. L. Osborn, “Architecture for data collection in database intrusion detection systems,” Secure Data Management, VLDB Workshop, vol. 4721, Sept. 2007, pp. 96-107.

[2] C. Mouza et al., “Towards an automatic detection of sensitive information in a database,” Int’l Conf. on Advances in Databases, Knowledge, and Data Applications (DBKDA), Jan. 2010, pp. 247-252.

[3] I. J. Rajput and D. Shrivastava, “Data mining based database intrusion detection system: A survery,” Int’l Journal of Engineering Research and Applications (IJERA), vol. 2, no. 4, July 2012, pp. 1752-1755.

[4] A. Kamra, E. Terzi, and E. Bertino, “Detecting anomalous access patterns in relational databases,” The VLDB Journal, vol. 17, no. 5, Aug. 2008, pp. 1063-1077.

[5] F. S. Rietta, “Application layer intrusion detection for SQL injection,” ACM Southeast Regional Conference, Mar. 2006, pp. 531-536.

[6] A. Adebowale, Idowu S.A, and O. Oluwabukola, “An overview of database centred intrusion detection systems,” Int’l Journal of Eng’g. and Advanced Technology, vol. 3, no. 2, Dec. 2013, pp. 273-275.

[7] Y. Hu and B. Panda, “A data mining approach for database intrusion detection,” ACM Symposium on Applied Computing, 2004, pp. 711-716.

[8] A. Srivastava, S. Sural, and A. K. Majumdar, “Database intrusion detection using weighted sequence mining,” Journal of Computers, vol. 1, no. 4, July 2006, pp. 8-17.

[9] S. Hashemi, Y. Yang, D. Zabihzadeh, and M. Kangavari, “Detecting intrusion transactions in databases using data item dependencies and anomaly analysis,” Expert Systems, vol. 25, no. 5, Oct. 2008, pp. 460-473.

[10] D. Barbara, R. Goel, and S. Jajodia, “Mining malicious corruption of data with hidden Markov models,” Research Directions in Data and Applications Security, Int’l Federation for Information Processing (IFIP), vol. 128, 2003, pp. 175-189.

[11] P. Ramasubramanian and A. Kannan, “A genetic algorithm based neural network short-term forecasting framework for database intrusion prediction system,” Soft Computing, vol. 10, no. 8, June 2006, pp. 699-714.

[12] C. Pinzon, A. Herrero, J. F. De Paz, E. Corchado, and J. Bajo, “CBRid4SQL: A CBR intrusion detector for SQL injection attacks,” Hybrid Artificial Intelligence Systems, vol. 6077, 2010, pp. 510-519.

[13] R. M. Elbasiony, E. A. Sallam, T. E. Eltobely, and M. M. Fahmy, “A hybrid network intrusion detection framework based on random forests and weighted k-means,” Ain Shams Eng’g. Journal, vol. 4, no. 4, Dec. 2013, pp. 753-762.

[14] E. Bertino, E. Terzi, A. Kamra, and A. Vakali, “Intrusion detection in RBAC-administered databases,” Computer Security Applications Conf. (ACSAC) , Dec. 2005, pp. 170-182.

[15] F. Valeur, D. Mutz, and G. Vigna, “A learning-based approach to the detection of SQL attacks,” Detection of Intrusions and Malware, and Vulnerability Assessment, vol. 3548, 2005, pp. 123-140.

[16] S. X. Wu and W. Banzhaf, “The use of computational intelligence in intrusion detection systems: A review,” Technical Report, Dept. of Computer Science, Memorial University of Newfoundland, Nov. 2008.

[17] A. Zainal, M. A. Maarof, and S. M. Shamsuddin, “Feature selection using rough set in intrusion detection,” IEEE Region 10 Conf. (TENCON) , Nov. 2006, pp. 1-4.

[18] Transaction Processing Performance Council (TPC), TPC benchmark E, Standard specification, version 1.13.0, 2014.

[19] H. A. Nguyen and D. Choi, “Application of data mining to network intrusion detection: Classifier selection model,” Challenges for Next Generation Network Operations and Service Management, vol. 5297, 2008, pp. 399-408.

[20] L. M. Lima de Campos, R. C. Limao de Oliveira, and M. Roisenberg, “Network intrusion detection system using data mining,” Eng’g. Applications of Neural Networks, Communications in Computer and Information Science, vol. 311, 2012, pp. 104-113.

[21] J. Zhang and M. Zulkernine, “Anomaly based network intrusion detection with unsupervised outlier detection,” IEEE Int’l Conf. on Communications, 2006, pp. 2388-2393.

[22] S. B. Franklin, D. J. Gibson, P. A. Robertson, J. T. Pohlmann, and J. S. Fralish, “Parallel analysis: A method for determining significant principal components,” Journal of Vegetation Science, vol. 6, no. 1, 1995, pp. 99-106.

[23] M. Hall et al., “The WEKA data mining software: An update,” SIGKDD Explorations, vol. 11, no. 1, 2009, pp. 10-18.

Improving ASR Recognized Speech Output for Effective Natural Language Processing

C. Anantaram, Sunil Kumar Kopparapu, Nikhil Kini, Chiragkumar Patel
Innovation Labs
Tata Consultancy Services Ltd.
Delhi & Mumbai, India
{c.anantaram, sunilkumar.kopparapu, nikhil.kini, patel.chiragkumar}@tcs.com

***Abstract*—The process of converting human spoken speech into text is performed by an Automatic Speech Recognition (ASR) system. While functional examples of speech recognition can be seen in day-to-day use, most of these work under constraints of a limited domain, and/or use of additional cues to enhance the speech-to-text conversion process. However, for natural language spoken speech, the typical recognition accuracy achievable even for state-of-the-art speech recognition systems have been observed to be about 50 to 60% in real-world environments. The recognition is worse if we consider factors such as environmental noise, variations in accent, poor ability to express on the part of the user, or inadequate resources to build recognition systems. Natural language processing of such erroneously and partially recognized text becomes rather problematic. It is thus important to improve the accuracy of the recognized text. We present a mechanism based on evolutionary development to help improve the overall content accuracy of an ASR text for a domain. Our approach considers an erroneous sentence as a zygote and grows it through an artificial development approach, with evolution and development of the partial gene present in the input sentence with respect to the genotypes in the domain. Once the genotypes are identified, we grow them into phenotypes that fill the missing gaps and replace erroneous words with appropriate domain words in the sentence. In this paper, we describe our novel evolutionary development approach to repair an erroneous ASR text to make it accurate for further deeper natural language processing.**

Keywords-evolutionary development; artificial development; speech recognition; natural language processing.

I. INTRODUCTION

Speech and natural language interfaces are becoming rather important means of communication with enterprise systems. As more and more end-users of enterprise applications are targeted (e.g., online shopping, banking), the demand for human-speech and natural language interfaces to such online application systems seems to be growing. Automated recognition of the user’s speech into natural language text and then processing that text is very important. It is imperative that this process becomes rather accurate. Similarly, some of the most used channels for customers to interact with human service-agents in an enterprise are still the telephony channel [1]. In several cases the customer actually speaks to a human agent to get an answer to the problem that he/she might face. With an increasing customer base and with a corresponding increase in transactional volumes, support personnel are rather stretched, and this results in a delay in responding to the customer. ASR systems with deep natural language processing have been found to help reduce this load by automatically routing calls to automated helpdesks, provided such recognition and processing is of good accuracy.

With self help solutions becoming popular, there has been a spurt of growth in Voice User Interfaces (VUI). A typical VUI-based solution would take as input a spoken utterance, recognize (speech to text conversion) the utterance, interpret it (natural language understating), fetch an answer in response from a structured or unstructured database, and communicate the response (text to speech) to the user. Clearly, the process of interpretation of the spoken query is subject to the accuracy of the speech recognition engine that converts the spoken speech into text. Most of the functional examples of speech recognition in day-to-day use work under the constraints of limited domain, and/or use of additional cues to enhance the speech-to-text conversion process. Typical recognition accuracies for state-of-the-art speech recognition systems have been observed to be about 50 to 60% for natural language spoken speech. Environmental noise, variations in accent, poor ability to express on the part of the user, or inadequate resources to build recognition system also affect the accuracy adversely. Natural language processing of such erroneously and partially recognized text becomes rather problematic. It is thus important to improve the accuracy of the recognized text. While it is desirable to have better speech recognition mechanisms through better training sets covering more sample scenarios for the speech recognition engine, the question of interest here is, how can we improve the accuracy of the recognized text that is output from an ASR engine for a particular domain?

We examine this problem and present a mechanism based on evolutionary development (evo-devo) processes [2][3] to help improve the overall content accuracy of a recognized text for a domain. Our approach considers an erroneous input sentence as a zygote and grows it through an artificial development approach, with evolution and development of the partial gene present in the input sentence with respect to the genotypes in the domain. Once the genotypes are identified, we grow them into phenotypes that fill the missing gaps and replace erroneous words with appropriate domain words in the sentence. This process of

ISBN: 978-1-61208-381-0

artificial rejuvenation improves the accuracy of the sentence, which can then be processed by a natural language processing application such as question answering [1][4][5], and workflow management [6]. Thus, the main contribution of the paper is in terms of proposing a bio-inspired novel procedure to repair the erroneously recognized text output by a speech recognition engine, in order to make the output text suitable for deeper natural language processing.

The rest of this paper is arranged as follows: in Section II, we describe the current state of Speech and Natural Language Processing (NLP) self help systems. In Section III, we describe our proposed evolutionary development approach, and give a detailed example of how it works in Section IV. In Section V, we show the use of the ASR repair approach in a self help system scenario.

II. State of Art

Although a number of attempts have been made to build speech and natural language interfaces for different applications [1][4][6][7], the attempts to build accurate speech and natural language processing systems for a domain is far from satisfactory [8].

While there are several ASR engines, both commercial and otherwise, their performance is highly dependent on the language, accent, dialect, and environmental noise. Even for the best of the ASR engine the accuracy of the recognition is as little as 50-60% for spoken natural language sentences [8]. Interpreting erroneously recognized text will result in erroneous interpretation of the query and the failure of the self-help solution to assist in addressing the queries of the users.

A typical speech recognition process is shown in Fig. 1. The input is the speech signal and the output is the recognized text. However, to achieve this simple process of speech to text conversion, there is a need for training. The training aspect involves the use of a well-structured speech corpus of a language containing several hours of speech to create Acoustic Model (AM) and Language Model (LM) for that particular language. The acoustic and language model can be assumed as a statistical representation of the spoken content, the dialect and the accent of a language.

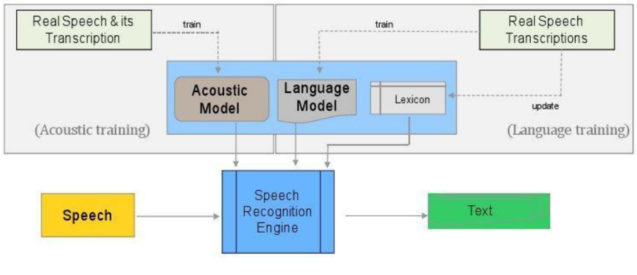


Figure 1. ASR Framework

There are several ways to improve the recognition performance of an ASR: (a) Fine tuning the ASR engine - This requires elaborate training, which in turn needs a rich amount of speech corpora. Very often, especially for not very popular languages, there is a dearth of speech corpora, and building a corpora is time consuming and expensive. There has been some work on building frugal speech corpora by exploiting the multimedia information on the Internet [9][10]. (b) Restricting what the user can say: This results in a restriction in the aspect of usability [11] and the VUI becomes user unfriendly.

In this paper, we examine the problem of improving the output of a speech recognition engine and present a mechanism based on evolutionary algorithms that help improve the overall content accuracy of a recognized text for a domain.

Much work has been done on automatic error detection in ASR output (a survey of this is presented in [12]), and also facilitating error correction for the user through easy-to-use interfaces [13][14], but to the best of our knowledge, there are no methods to automatically correct an ASR's output. Our work concerns not only detection, but automatic error correction after ASR. Previous work on this can be found in [15][16] [17].

As mentioned by Ringger and Allen [15], the reason this is an important problem is it allows for the ASR system to be a black box, whose output can be processed separately. This is particularly useful for improving proprietary systems where access to improve the system internally is not available. Also, such a post-correction system provides greater flexibility in terms of modeling domain variations and rescoring the output, in ways that are not possible in the ASR system [15].

Another example of error correction on ASR output can be found in [16]. In our research, we make use of evo-devo based repair methods that introduce an element of randomness in error correction, which is useful when training data for error correction is small or absent.

Our work is mainly concerned with ASR output that acts as input to another system, in this case, an NLP system that can retrieve answers to questions posed. IBM Watson Engagement Advisor [18] is an example of a commercial system that processes questions posed in text form and finds answers to them. Its basic working principle is to parse keywords in a clue while searching for related terms as responses [5]. Watson has deficiencies in understanding the contexts of the clues. Also, the setup cost and initial investment is too high, which makes it less suitable for being easily accessible and usable.

III. Our Approach

We consider the situation where a speech recognition engine takes a sentence spoken by a person as input and outputs a text sentence that is not an accurately recognized text. For example, when a person spoke the following sentence "Give me all the existing customers", the output sentence from our ASR engine was "The the contact existing customers". The question we tackle here is, how do we repair (or grow) the sentence back to the original sentence as intended by the speaker? It is in this context that our approach considers an erroneous input sentence as a zygote and grows it through an artificial development approach, with evolution and development of the partial gene present in the input sentence with respect to the genotypes in the domain. Once the genotypes are identified, we grow them into phenotypes that fill the missing gaps and replace erroneous words with appropriate domain words in the sentence. This process of artificial rejuvenation improves the accuracy of the sentence, which can then be passed onto a natural language processing application for further processing. The overall processing is described below.

A. Identifying the genotypes for the sentence

We consider each sentence as an individual in the population, i.e., as a zygote, and identify the genes that can apply on the sentence through partial match of concepts of the sentence with the ontological rules. The set of genes in the sentence form the genotypes.

1) Seeding the gene set for the domain

The ontology of a domain describes the domain terms and their relationships. A seed ontology details the meta-relations that are defined in the domain [4], for example “project has status”, “project has start_date” etc.

The application data (i.e., the database of the business application) is taken as input to instantiate the terms and relationships defined in the seed ontology in order to form the actual ontology of the domain [4]. This ontology (stored as a Resource Description Framework graph) forms the basic genes of the domain and their relationships with a <subject-predicate-object> structure for each of the genes.

2) Identifying the genotypes in the input sentence

We match sub-parts (or sub-strings) of the input sentence with the genes of the domain. The match will be partial due to the error present in the input sentence. The genes that match the closest are picked up, provided they satisfy fitness criteria.

B. Simulate the evolution and development process of the genotypes

Once the basic genes are identified, we develop the genes to better fit the situation on hand with evolution and development of the genes, and then score against a fitness function and select the “fittest” gene that survives. This gives us the set of genotypes that will form the sentence.

C. Developing the genotype to produce / extract its phenotype

The overall genotypes are collated together to form the input sentence. In this context some of the genotypes may need further development to form the final sentence that the user actually intended.

D. Evaluate the developed sentence

The developed sentence is then presented to a human Oracle who ranks the sentence if he/she deems it as a better fit (i.e., more accurate) for the domain.

IV. A DETAILED EXAMPLE

Let us consider an input speech where the user says the following sentence: “What is the status of the project in which Vinay is a team member?” A general purpose speech recognition system recognizes this spoken sentence as follows: “What is a hat us of the project in itch Vinay is a tea ember?” Thus the recognized sentence has errors and is not accurate.

We run the artificial development approach on this input sentence in order to repair the input and make it more accurate.

A. Identifying the genotypes for the sentence

We assume that the domain has the following ontology that is formed from the seed ontology and application data:

ds:project	ds:has	ds:status
ds:project	ds:has	ds:start_date
ds:project	ds:has	ds:role
ds:role	ds:is	ds:Team_member
ds:Vinay	ds:allocated	ds:ArtDevPrj
ds:Vinay	ds:role	ds:Team_member
ds:ArtDevPrj	ds:status	ds:Active

(The “ds:” prefix above is the namespace for this schema)

Firstly, the ASR output is parsed for identifying the parts of speech in the sentence. This process identifies the nouns, verbs, adjectives and adverbs in the sentence. Since the sentence itself is inaccurate, the parts of speech may not be accurate. For our example, parts of speech tagging gives the following output: ‘what/WP/what is/VBZ/be a/DT/a hat/NN/hat us/PRP/US of/IN/of the/DT/the project/NN/project in/IN/in itch/NN/itch vinay/NN/vinay is/VBZ/be a/DT/a tea/NN/tea ember/NN/ember’.

Using these identified parts of speech (especially nouns), the relevant subject-predicate-object of the domain that are referred in the sentence are marked (called partially matching genes). This is done through a partial-match algorithm wherein the ASR output sentence is matched with the ontology. The sub-string ‘project’ matches with an entry in the ontology, and thus a partial gene is triggered. The parts of speech identified before ‘project’ help narrow down to two possible genes: ‘project has status’ and ‘project has start_date’. Similarly, the sub-string ‘Vinay’ matches with two entries in the ontology, namely ds:Vinay ds:allocated ds:ArtDevPrj and ds:Vinay ds:role ds:Team_member and thus, a partial gene is triggered. The parts of speech following Vinay, especially “is/VBZ/be a/DT/a” help identify that a Verb and a Determiner follows and some relationship with ‘Vinay’ is expected. Thus, both these ontology entries are considered as genes of the input sentence.

The set of all possible genes that are identified in the ASR output sentence are considered as the genotypes in the sentence that need to be evolved and developed.

B. Evolution and Development of the genotypes

Using phonetic match, i.e., match of phonemes in words, between the ASR output sentence and the identified genes, we develop the partial gene present in the sentence. Phonetic match algorithms, such as Soundex [19], can be used for such a match. Thus ‘tea ember’ and ‘team member’ phonetically close, and the gene ‘Vinay role Team_member’ is selected. Similarly ‘hat us’ has a close match with ‘status’ rather than ‘start_date’. Hence, the fitness of the gene 'project has status' is better than the fitness of the gene 'project has start_date' in this context.

We use such fitness functions to select the genes that need further development for the ASR output. Currently in our approach, the development process is simulated by a replacement of the partially identified genes with the genes that are most likely. One can later introduce a more elaborate development process. Hence the ASR output sentence is modified to become "What is a status of the project in itch Vinay is a team_member?"

The genotypes in the sentence are 'status of the project' and 'Vinay is a team member'. The rest of the sentence needs to be further developed to make it more accurate.

C. Developing the genotypes to produce its phenotype

We now have a sentence that has been repaired through evolutionary development method that needs further development to make it accurate.

We parse the re-written sentence again to identify its new parts of speech. Thus, for the modified sentence we get: "what/WP/what is/VBZ/be a/DT/a status/NN/status of/IN/of the/DT/the project/NN/project in/IN/in itch/NN/itch vinay/NN/vinay is/VBZ/be a/DT/a team_member/NN/team_member".

We notice that there is a WP tag that refers to a Wh-Pronoun. However a WDT tag is missing that refers to a Wh-Determiner in the sentence. Using this clue we look for a phonetically matching word that could possibly match with a Wh-Determiner. Our match-function identifies "itch" as more phonetically close to "which" (that is a Wh-Determiner). This is a second-level fitness function and thus we can rewrite the modified sentence as follows "What is a status of the project in which Vinay is a team_member?" This sentence is now ready for accuracy evaluation.

D. Evaluate the developed sentence

In this step, we evaluate the accuracy of the artificially developed sentence to determine if it is a better fit for the domain than the ASR output. At present, we assume the presence of an oracle of the domain to evaluate the accuracy of the developed sentence. Later on, such a process can also be automated by formally defining accuracy and developing precise mechanisms to measure it.

The output of the artificial development approach is presented to an oracle who evaluates the accuracy of the sentence. The parts of speech for this newly developed sentence: "what/WP/what is/VBZ/be a/DT/a status/NN/status of/IN/of the/DT/the project/NN/project in/IN/in which/WDT/which vinay/VBP/vinay is/VBZ/be a/DT/a team_member/NN/team_member". The sentence has more ontology terms and relationships of the domain and the parts of speech are also complete. Thus, the oracle marks the newly developed sentence as accurate.

The artificially developed sentence, which is now marked as accurate, can now be processed by deeper natural language processing applications such as question-answering/workflow management/self help tools [1][4] [18].

V. Self help Case Study

In the scenario of a retail outlet that has a large number of products, lots of promotion offers, and catering to many customer queries, self help becomes a very important aspect of customer experience. Consider the following query asked by a customer via an interactive audio self help system:

User: Which camcorders have more than 20% discount?

The ASR system processes the speech and converts it to text. However, as described above, the output text may be erroneous. In this example:

ASR output: Itch came orders have more the 20% this count?

We will need to repair the output since there are recognition errors. Following our artificial development method described above we get the repair steps as:

Genes identified: Camcorder, discount

Genotype repair**:** "came orders" repaired to "camcorders" and "this count" repaired to "discount"

Phenotype repair**:** "Itch" repaired to "which"

Repaired sentence**:** "Which camcorders have more than 20% discount?"

This sentence is passed onto the self help question answering system. The output of the system is:

System**:** The Camcorders are
DXG 3MP Digital Camcorder - DXG-301V
Panasonic Mini DV Camcorder
Aiptek IS-DV2 Digital Camcorder
Panasonic 2.8" LCD Digital Camcorder with 3CCD Technology - Silver (SDR-S150).

Thus we can see how the speech and natural language system with repair of ASR output has answered the customer's question in a self help situation and improved the overall customer experience.

VI. Conclusion

We have described a mechanism to artificially develop and improve an ASR output sentence to make it more accurate for a domain by following the evo-devo based artificial development approach. The idea is to work with the inaccuracies in the recognition and repair/develop/grow-out the error and replace it with a more accurate sentence that can be processed further by a natural language processing system. This helps in better speech-and-natural language interface systems for enterprises and aids in self help systems.

VII. References

[1] http://www-03.ibm.com/innovation/us/watson/science-behind_watson.shtml [Retrieved: January, 2015]

[2] S. Harding and W. Banzhaf, “Artificial Development,” http://www.cs.mun.ca/~simonh/publications/evodevbookchapter.pdf [Retrieved: January, 2015]

[3] G. Tufte, “From Evo to EvoDevo: Mapping and Adaptation in Artificial Development,” http://www.intechopen.com/books/evolutionary-computation/from-evo-to-evodevo-mapping-and-adaptation-in-artificial-development [Retrieved: January, 2015]

[4] S. Bhat, C. Anantaram, and H. Jain, “Framework for text-based conversational user-interface for business applications.” In Proceedings of the 2nd international conference on Knowledge science, engineering and management, pp. 301-312. Springer-Verlag, 2007.

[5] IBM Watson - http://en.wikipedia.org/wiki/Watson_(computer) [Retrieved: January, 2015]

[6] S. Bhat, C. Anantaram and H. Jain, “An architecture for intelligent email-based workflow interface to business applications,” International Conference on Artificial Intelligence (ICAI-2008), WORLDCOMP’08, Las Vegas, USA, pp. 344-350. July 14-17, 2008.

[7] A. Imran, S. K. Kopparapu, “Building a Natural Language Hindi Speech Interface to Access Market Information,” The Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, NCVPRIPG, Hubli, 2011.

[8] C. Lee, S. Jung, K. Kim, D. Lee, and G. G. Lee, “Recent Approaches to Dialog Management for Spoken Dialog Systems,” Journal of Computing Science and Engineering, vol. 4, no. 1, March 2010.

[9] I. Ahmed and S. K. Kopparapu, "Speech recognition for resource deficient languages using frugal speech corpus," in Signal Processing, Communication and Computing (ICSPCC), 2012 IEEE International Conference on, 2012, pp. 750–755.

[10] S. K. Kopparapu and I. Ahmed, “A Frugal Method and System for Creating Speech Corpus,” Indian Patent 2148/MUM/2011; Jul 28, 2011.

[11] S. K. Kopparapu, “Voice Based Self Help System: User Experience Vs Accuracy,” Book Chapter, Innovations and Advances in Computer Sciences and Engineering edited by Tarek Sobh, Springer, 1st Edition, ISBN-13: 978-9048136575, March 2010.

[12] Y. Shi, "An investigation of linguistic information for speech recognition error detection," PhD diss., University of Maryland, Baltimore County, 2008.

[13] J. Ogata and M. Goto. "Speech repair: quick error correction just by using selection operation for speech input interfaces," In INTERSPEECH, pp. 133-136. 2005.

[14] D. Harwath, A. Gruenstein, and I. McGraw. "Choosing Useful Word Alternates for Automatic Speech Recognition Correction Interfaces." In Fifteenth Annual Conference of the International Speech Communication Association. 2014.

[15] E. K. Ringger and J. F. Allen. "Error correction via a post-processor for continuous speech recognition," Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings, 1996 IEEE International Conference on, vol. 1, pp. 427-430. IEEE, 1996.

[16] M. Jeong, B. Kim, and G. Lee. "Using higher-level linguistic knowledge for speech recognition error correction in a spoken Q/A dialog," Proceedings of the HLT-NAACL special workshop on Higher-Level Linguistic Information for Speech Processing, pp. 48-55. 2004.

[17] R. López-Cózar and David Griol. "New technique to enhance the performance of spoken dialogue systems based on dialogue states-dependent language models and grammatical rules," In INTERSPEECH, pp. 2998-3001. 2010.

[18] IBM WATSON Engagement Advisor - http://www-03.ibm.com/innovation/us/watson/watson_for_engagement.shtml [Retrieved: January, 2015]

[19] http://en.wikipedia.org/wiki/Soundex [Retrieved: January, 2015]

Digital Inclusion - The Vision and the Reality

Leela Damodaran, Wendy Olphert, Teresa Gilbertson,
Loughborough University
Loughborough, UK
L.Damodaran@lboro.ac.uk, C.W.Olphert@lboro.ac.uk,
T.Gilbertson@lboro.ac.uk

Mary Craig
University of Edinburgh
Edinburgh, UK
m.craig@ed.ac.uk

Jatinder Sandhu
Nottingham Trent University
Nottingham, UK
jatinder.sandhu@ntu.ac.uk

Abstract— **The benefits of a digitally inclusive society are vast and the need for such inclusion is now a requirement for full participation in our society. While the basic concept of universal digital inclusion is simple, the reality is a long way from the vision. Despite efforts to reduce it, inequality of access still exists. The beneficiaries of a digital society are not just the individual, but all stakeholders in the wider society. While the challenges to achieve a fully inclusive digital society are considerable, the knowledge of how to create such a society already exists. The creation of local venues for inclusively designed ICT (Information and Communications Technology), support and learning in familiar places along with the harnessing of political will could make such a society a reality rather than a vision. With the cooperation of all stakeholders, actualisation of the vision of a digitally inclusive society, while challenging, will yield opportunities that eclipse the cost of implementation.**

Keywords-Digital society; digital inclusion; accessibility; participation.

I. INTRODUCTION

The fundamental concept of digital inclusion is the deceptively simple premise that everyone in the world deserves to have equal access to whatever knowledge and information they require to enable them live their lives to their full potential, which crucially, now depends on fast and reliable internet access. Despite attention in society from academics, politicians, social activists and many others for almost two decades, the reality still seems to be a long way from the vision. This paper revisits the vision and aspiration of digital inclusion, and then, seeks to look beyond the rhetoric to provide an analysis of the status quo, a consideration of some facilitators and inhibitors to progress and some suggestions for moving forward with renewed energy and commitment.

A. Background: The digital divide

The term "digital divide" was adopted by the Clinton/Gore administration in the US in the late 1990s and used in 1998 in a commencement speech at the Massachusetts Institute of Technology:

"...the digital divide has begun to narrow, but it will not disappear of its own accord. History teaches us that even as new technologies create growth and new opportunity, they can heighten economic inequalities and sharpen social divisions" [1].

The initial focus of the digital divide was one of access to technology and the acquisition of the basic skills to use it. As access to computers has increased across all members of industrialised societies, the digital divide has become not just about access and the acquisition of basic skills and knowledge, but also about the ability to exploit technologies for personal benefit, empowerment and even transformation. For such a divide to be narrowed, it is not just equipment and basic skills that are required but, confidence, good support and appropriately designed equipment and services.

The benefits of an inclusive digital society are well-documented and emphasise extensive and transformational impact. There are numerous published articles by governments, academics, practitioners and others which show that being part of the digital world can improve life in numerous ways, means and forms [2]-[4]. Often, the emphasis of such publications is on the financial savings for the state and for individuals. While monetary gains are important, the potential transformation for individuals, society and the economy are vastly more far-reaching. Appropriate and competent use of digital technologies also helps to improve the well-being of individuals and maintain their independence, autonomy and social connectedness. Other benefits include civic participation and the opportunity to improve skills to gain employment or simply to ‘keep up with the times’.

Research identifies many factors that lead to inequality of access and opportunity. For example, Norris [5] recognises three kinds of digital divide:

- Global: the divide between the industrialised and the developing world
- Social: the divide between the have and the have-nots in society
- Democratic: the divide between those who use technology to participate in the public realm and those who do not [5].

In order for the divide to be significantly reduced or eliminated, each of these aspects need to be recognised and appropriate solutions and coping strategies put in place. This paper provides a vision for the development of a fully inclusive digital society by identifying the benefits, beneficiaries as well as a roadmap to achieve this vision.

The inspirational and aspirational ‘Declaration of Principles’ presented at the World Summit on the Information Society, Geneva, 10-12 December 2003 [6] states the following:

> **"We**, **the representatives of the peoples of the world** …declare our common desire and commitment to build a people-centred, inclusive and development-oriented Information Society, where everyone can create, access, utilize and share information and knowledge, enabling individuals, community and peoples to achieve their full potential in promoting their sustainable development and improving their quality of life ….”

Since that time, the overarching goal and some of the components embedded in the declaration have become familiar and well-established – if only by repetition. Some of the frequently voiced assertions, aspirations and principles of digital inclusion include the following:

- Global access to ICT (Information and Communications Technology) will make a significant contribution to improving health and wellbeing and quality of life for all
- Digital technologies make possible transformations which enhance quality of life of individuals, increase life chances, prolong independence and autonomy and improve social connectedness
- The transforming capabilities of digital technologies improve society and boost the economy
- Everyone should be able to access, create, utilise and share knowledge and information.

The remainder of the paper will examine and explore the subject of digital inclusion in greater depth. First, the definition of digital inclusion will be introduced, followed by a presentation of some examples of evidence of the value of a digitally inclusive society. The benefits and beneficiaries of such a society will then be introduced. Finally, the challenges to achieving a digitally inclusive society and the steps needed to make such a vision a reality will be discussed.

II. Defining Digital Inclusion

An inclusive digital society has been defined as one in which all members of a community are **able** to **access**, **use**, and **understand** digital technologies [7]. For this goal to be achieved the following preconditions are required:

- Connectivity – infrastructure and individual access to appropriate hardware, software, services
- Capability – education, tailoring for ability/disability, digital literacy and skills
- Content – availability of accessible, meaningful, relevant material

Meeting these preconditions will require the key stakeholders in society to collaborate effectively to make the vision a reality. The relevant stakeholders include local and national Government, all business sectors and professional groups, retailers, designers, developers and manufacturers of ICT products and services, third sector organisations, groups and communities of people and individuals. Crucially, they include diverse groups of the population, including the young and the old, the employed and the unemployed, the retired, the disabled and those unable to work. Only by recognising the wide range of needs and perspectives of all people and groups that make up our global society and making provision for them will digital inclusion come about. However, this will require that relevant influential stakeholders actively strive to achieve the digital participation of all the stakeholders in society.

III. Why is a digitally inclusive society Important?

Digital literacy is fast becoming a requirement for full participation is society. In our emerging e-society, ICTs are an important aspect of daily life. They offer particular advantages to older adults, such as helping them to stay connected to family and friends, to pursue their interests and hobbies, to organise the mechanics of daily living, to benefit from the financial savings of internet shopping and to access health and social care. Yet, large numbers of older adults are reported to be non-users of technology [8]. Recent statistics show that while 99% of people aged between 16 and 34, this drops to 88% for people aged between 55-64, 71% of people aged between 65-74 and a very low 37% for people aged over 74 years [9].

The real benefits lie in the transformational capability of digital technologies. One strong example of the power and significance of the benefits of digital connection is seen in the experience of an individual reported on the website of Digital Unite. The individual was a woman who had lost her sight 6 years previously and had to rely on her husband to read the post, write letters and other tasks that required reading such as checking recipes. Her life was ***transformed*** when she discovered that she could download magnification

ISBN: 978-1-61208-381-0

and text to speech software. By having access to such software and hardware, she regained her independence [10].

While this is only one example, there are numerous examples (e.g., from Leicestershire CareOnLine [11]) showing similar transformations, and the importance of such transformations cannot be understated. For those who are, or who become, digitally disengaged, there are real social and economic consequences in terms of accessing government services, accessing health information and social support as well as researching and procuring goods and services, or accessing price comparison and review websites. Indeed, it can be argued that the cumulative ripple effect of individual digital inclusion is far-reaching enough to have the potential to affect wider society, for example, by reducing the costs of maintaining people in the community who might otherwise require residential care or significant at-home support. The heart of this vision is to ensure that everyone is able to access and experience these transformational opportunities and impacts. The following section considers the benefits and beneficiaries.

IV. The benefits and beneficiaries

To promote a digitally inclusive society, it is important not only to raise awareness of the wide-ranging benefits but also to acknowledge the wide range of beneficiaries of such a society.

A. Benefits

The spectrum of benefits of being part of the digital world is vast and increasing all the time. This bigger picture can be overlooked when the specific aspects of digital inclusion, such as responsibility for delivering on-line services, are viewed in isolation. The benefits of a digitally inclusive society include the major advantages to the economy and society which derive from enabling individuals to become or remain economically active through learning digital skills, resulting in enhanced health, wellbeing and quality of life as well as increased opportunities for social interaction, cohesiveness and civic participation.

The keys to enabling economic development at the individual level are: enabling education; increasing opportunities; increasing self-efficacy and improving and expanding skills.

1) Health, wellbeing and quality of life

Digital participation also helps to reduce loneliness and improve independence and wellbeing. These benefits could in turn lessen demands on other (formal and informal) support systems for older people; improving the quality of life of older people while also reducing costs of care. By enabling rehabilitation, remote assessment and diagnosis and treatment delivery, the need for residential care could be delayed, or for some, avoided by supporting policies of early intervention.

2) Social interaction & cohesiveness

The social environment, as well as the social opportunities afforded by connectivity, promotes inclusion and helps to reduce social isolation. This increased social contact helps maintain good mental health by reducing the depression, stress and anxiety associated with social isolation. Increasing digital participation allows people to stay connected with friends and family, their local community and the wider world in a variety of ways. Civic engagement and participation, e.g., voting and knowing what is going on in your community, is maintained even for those who are housebound. Enabling social networks thus creates a digital haven of having fun and increasing and maintaining social contacts.

B. Beneficiaries

The beneficiaries of a digitally inclusive society are numerous and include key stakeholders in the following categories: government – national and local, service providers, retailers of on-line services and products, designers and developers of ICT products and services, AT (Assistive Technology) providers, individuals and society.

1) Government

a) National Government

Digital inclusion would benefit national government by enabling citizens to possess the skills that are required of a modern workforce. Additionally, if more people feel confident to access centralised services online, the burden on traditional services will be reduced. The provision of care would also benefit from a digitally-engaged populace with access to the health, care and wellbeing benefits enabled by technology.

b) Local Government

The ability for all members of society, especially older people who use large numbers of services, to utilise online services, would allow for cost-savings arising from greater individual health, care delivery and societal participation. Local government could expect see a lessening on the burden on some services as a result of the benefits of digital inclusion that enable independent living, greater well-being and the reduction of social isolation.

2) Service Providers

In the case of the public sector, the increased levels of confidence and capability of older people in digital participation will begin to increase the uptake of on-line services (e.g., Universal credit) and this will be of great value to overburdened local government bodies attempting to ‘achieve more with less’ (such an effect would be especially beneficial in achieving savings as the heaviest use of government services is by older people).

3) Retailers of On-line Products and Services

Well-publicised demographic change means that a major market exists in the 50+ age group which is not well-served and has yet to be ‘discovered’ by many retailers and other commercial companies. Understanding and being able to 'segment' appropriately the older market to achieve better tailoring of products and services offers competitive advantage to business. For businesses that provide online services, e.g., banks and retailers the existence of community venues will provide a venue for retailers and service providers to demonstrate their online services (without sales pressure!), provide information and support, build capacity in the older population, encourage customer loyalty and develop customer engagement.

4) Designers and Developers of ICT Products and Services

The design of technology can pose many problems – particularly, the speed of change and unnecessary complexity arising from function creep (i.e., the including of features outside that of the original specification). When combined with a raft of ageing issues and disabilities which are associated with ageing, these factors adversely affect older adults' experiences of using and sustaining use of ICTs. Good design can mitigate these effects and enhance the user experience.

The principles of designing for inclusion are well established and relate to the importance of eliciting detailed user requirements and then tailoring products, systems and services to these requirements. Despite this knowledge having been available for decades, it is still not the norm to design for inclusion. And yet, one of the major concerns identified by older people is that constant technological change poses a major hindrance for them in maintaining digital connection. It is entirely possible to have new functionality hidden, in order to keep the interface unchanged. If people want to do more, then they can; but when it suits them and when help is available. In other words, stability of the interface is maintained for them until/unless they want enhanced functionality.

The demands of digital engagement, especially cognitive load, can be reduced by designs which are:

- Tailored to diverse user requirements
- User friendly
- Accessible
- Intuitive
- Seamless
- Embedded where appropriate ("hidden functionality")
- Adaptive
- Making upgrading and administration transparent and easy

Meeting such design needs represents a commercial opportunity for ICT designers as well as offering an exciting intellectual challenge.

5) Assistive Technology (AT) Providers

Many ATs for people with recognized disabilities can help everyone in demanding or extreme usage situations. Accessibility = Profitability. Successful examples include:

- Text to speech
- Shorthand for text messaging
- Image stabilization
- Closed captions in video games

Such accessibility technologies open up new market opportunities in every sphere of life, e.g., healthcare, homecare, commerce, education and recreation.

6) Individuals

Individuals will benefit from all the aspects detailed above. For example, creating adaptable interfaces could be particularly important for older users who want to use what everyone else is using rather an AT, either because of the steep learning curve of some ATs, or because of a personal preference to use what everyone else is using [12]. This experience of built-in personal customisation is of benefit to all. Similarly adaptivity features that aid in automatic customisations can reduce or eliminate the learning curve of such changes - removing a number of access barriers. By ensuring stable, intuitive, usable and adaptive design, individuals will benefit in terms of not only accessing goods and services but also engaging in personal pursuits and living independently.

7) Society

The breadth of individual economic, health and wellbeing and social benefits combined with the reach of benefits across the private and public sectors benefits society as a whole. Moreover, the interaction between all of these advantages creates a synergy such that the total benefit to society is potentially far greater than the sum of the individual benefits.

A genuinely digitally inclusive society offers transformations which range, for example, from the empowerment that results from an individual who learns a specific skill on YouTube being able to solve a particular problem/meet a need, becoming economically active through online courses and social support, to being enabled to participate fully in society. There far-reaching benefits, will not only benefit the individual experiencing them, but also wider society and the economy.

V. THE CHALLENGES TO DIGITAL INCLUSION

The challenges to achieving digital inclusion are extensive. Not only are there significant challenges in getting everyone online – especially some older and disabled people, but in addition, there are special challenges for them in staying 'connected'. It is a common myth that 'once people are online, they stay online'. Various studies show that some people who have used the internet at some point, and for some period of time, have subsequently stopped doing so [13]-[16]. This phenomenon is a potential but largely unrecognised 'fourth digital divide' [17], i.e., it cannot be explained by a simple interpretation of lack of access, lack of skills or lack of interest or motivation, because the people in this category have formerly been users. One in ten people are reported to have given up on using computers and it has been found that the older generation are more likely to be the ones that 'give up' [18].

There are a number of factors which may lead to older people giving up use of computers and other digital technologies. It is well established that older adults are vulnerable to social, cognitive and physical changes in later life. These changes have important implications for older adults' experiences of learning to use and sustain use of ICTs. Changes in psychological and cognitive aspects will impact on remembering sequential processes and confidence levels in using ICTs. Changes in physical aspects, such as changes in vision have implications for seeing what is on the screen and dexterity issues will create problems for controlling the mouse. Additionally, social changes, such as family members moving away, also impact on the support available to older ICT users. Further, older adults encounter some or a range of the following barriers in learning to use and sustain use of ICTs: confidence and fear of using ICTs; problems with understanding technical jargon and dealing with pop-ups and spam; problems with updates, drivers, and software; dealing with the rate of change of technology and coping with poorly designed software and hardware [19].

The absence of adequate learning and support also impacts on older adults' abilities to continue use of ICTs. A survey of older ICT users showed that 56% of older people said they regarded support as the most important factor in sustaining their digital participation. Respondents reported using a wide range of learning mechanisms, including self-directed learning (54%), inter-generational learning, peer-to-peer learning (40%) and taught classes (47%). Respondents reported heavy reliance on support from family members or friends both to learn and to solve problems. Around a quarter of respondents said that human support and encouragement was the most important thing to help them use technology successfully [20]. Yet a further study on the Sus-IT project showed that the opportunities for learning and on-going support are extremely inadequate [19].

These learning mechanisms alone are not adequate and the UK Digital Inclusion Panel Report [3] reported that "there is a real risk that in the medium to long term, significantly more citizens will migrate from being digitally engaged to being unengaged than the other way round, as their capabilities change".

One of the major determinants of sustained digital engagement – and therefore of digital inclusion in the long term – and for older people in particular, is the quality of support available to them. For ICT users in the workforce, the majority of workplaces will have a dedicated member of staff tasked with ICT support i.e. setting up and maintaining the infrastructure, selecting which technology platform is used, installing updates etc. The ICTs are looked after for those in the workplace although they are the relatively able bodied, reasonably technologically au-fait members of society. What exists beyond the workplace for the many without the institutional support is very different. In contrast to the situation in many workplaces, many ICT users who are at home, unemployed, retired, living with disabilities, living on a reduced income, living with a reduced social circle and possibly living with reduced health find themselves having to cope with all these demands themselves. There are organisations, such as public libraries that can offer limited help, but more typically it is piecemeal and variable, and as such, is inadequate and unsatisfactory. In spite of this lack of support it is amazing that so many do succeed with so little support. Just think what these individuals could achieve with adequate support.

The current situation is very unsatisfactory from the point of view of older and disabled users and from the perspective of many stakeholders tasked with delivering services designed to be used on-line. For digital inclusion to become a reality for these groups, vastly improved access to ICT support in the community is required.

From the brief analysis of the status quo presented above, it will be evident that the challenges to digital inclusion are immense.

VI. How to meet the challenges?

Encapsulating the essence of the 2003 Declaration of Principles allows the vision of an inclusive society and economy to be articulated as "the enhancement of the quality of life for all, extending autonomy and independence through the use of digital technologies. This vision would be characterized by:

- Empowered people experiencing the benefits of digital inclusion
- Widespread participation in society and the economy
- Readily available support in the community for engaging with and managing all aspects of the digital world".

To create the digitally inclusive society encapsulated in the vision described above requires structural, political, and social change on a vast scale - which perhaps helps to explain the slow progress to date.

The process of achieving such change needs itself to be inclusive of all stakeholders across society. This means that to succeed, the co-creation of an inclusive society is required – and this will require collaboration on a grand scale to address the challenges through innovating, creating and evolving a digitally inclusive society that harnesses the power of ICT for the benefit of all. The knowledge to meet the challenges already exists, but before a fully inclusive digital society can be achieved, the prevailing myths concerning the idea that once people are online they stay online; that there is no clear cut way to get older people online; that the one-size fits all workplace model of ICT training is sufficient for widespread digital inclusion; and that older people are not interested in shaping the design of technologies to their benefit should be dispelled.

The key enablers for a digitally inclusive society are the provision of Community ICT learning and support venues and the will of influential stakeholders.

A. Community ICT Learning and Support Venues

Learning to use new technology is only the beginning of the journey to competent and confident use of ICTs. People

ISBN: 978-1-61208-381-0

need appropriate venues in which to learn to use ICTs. The majority of ICT courses and campaigns to promote internet use are centred on acquisition of basic skills to go online rather than on promoting confidence and sustained usage. Many older people express their discomfort with formal classes and the negative associations with their school days. Once initial training is over, older people can feel alone, anxious and frustrated when experiencing problems with on-going ICT use. Fifty-six percent of older people said they regarded support as the most important factor in sustaining their digital participation [20]. Evidence also indicated that older people would like ICT learning and support opportunities that are user friendly, accessible, affordable, local, adaptive and embedded in purposeful and enjoyable activities. Similarly, findings from the research suggest that once older people are doing what matters to them, facilitated by ICT use, they are more likely to be motivated to continue their participation in the digital world and to progress to being ‘digital by choice’ in other areas of their lives.

To enable individuals that are not familiar with ICT but wish to develop digital literacy skills, to have their ICT needs met, the requirements articulated by older people in the UK are for access at home/in the community to inclusively designed ICT, support and learning which utilises wherever possible, existing venues e.g., libraries and village halls with which people are familiar and comfortable and which are sustainable.

B. *Harnessing Stakeholder Will*

The vision and benefits of a digitally inclusive society are known and the structural changes needed to achieve the vision have also been identified. These are essential preconditions but are not sufficient to turn the vision into reality. Change comes about when (i) dissatisfaction with the ‘As Is’ situation (ii) a shared vision of the desirable future and (iii) a clear ‘roadmap’ to achieve the vision are cumulatively greater than the costs (both financial and psychological) of implementing the vision. Where this situation arises, it is often the case that a strong will develops among stakeholders to take action and move towards the shared goal. While there will inevitably be different perspectives and priorities among individuals, communities, businesses and government regarding what a digitally inclusive society looks like, the transformational outcomes envisioned by the Declaration of Principles help to inspire a shared vision and perspectives and encourages commitment to follow the path to achieving this. The sharing of perspectives between all stakeholders, especially those in positions of authority such as national and local government and those in business developing and selling ICTs are crucial to success in creating and sustaining a digitally inclusive society. Inter- and intra- stakeholder co-operation, ongoing negotiation for the mutual benefit of varied stakeholders and respecting the voices of less influential/authoritative stakeholders is also key to this journey.

VII. NEXT STEPS

To provide a path towards the vision of an inclusive digital society, the following steps are essential:

- Promote widespread awareness of the benefits – individual, societal and economic – of digital inclusion
- Engage and gain the ‘buy-in’ of key stakeholders to the Vision of digital inclusion
- Encourage and reward adoption of inclusive design principles and promote them as the industry ‘norm’ for designers, developers and manufacturers of ICT systems, services and products
- Create expectations of and demand for inclusive design/digital inclusion amongst buyers and users of ICT
- Promote awareness that current ICT learning and support provision in the community is variable in quality and availability and not ‘purpose’ tailored to the requirements of users
- Recognise that sustaining people online is an even greater challenge than getting them online in the short term – and invest in community provision indicated above.
- Create a framework for ubiquitous provision of ICT support in the community e.g., in public libraries
- Document and co-ordinate the various local initiatives that exist, e.g., some GPs are now “social prescribing” (e.g., recommending patients make an appointment at a local library to obtain digital skills training).
- Utilise local resources to meet local needs
- Harness the political will to push this vision
- Recognise and celebrate what we can all achieve given the right access to ICTs.

These steps offer a roadmap to a digitally inclusive society.

VIII. CONCLUSION

In summary, digital inclusion is fundamental to a flourishing democracy and to the full participation of people in society and the vision for universal inclusion and connectivity has been stated. At the individual level, digital inclusion is crucial to sustaining and enhancing independence and autonomy. However, the transformative potential of such inclusion transcends the individual and has wide-reaching benefits for the whole of society and the economy. To achieve the vision will require coordinated policies, strategies and practices which led or endorsed by national governments, coordinated and implemented by local government, service providers, businesses and third sector organisations. The barriers to achievement of a digitally inclusive society are well understood and, while commitment and effort to overcome them will be required, solutions are available and the return on investment in implementing these will be extensive. With leadership and commitment, sustained digital connectivity for everyone is within our grasp now. In particular, leadership is needed to promulgate the vision and to encourage the development of strategic alliances and partnerships within a framework of appropriate policies and strategies, involving all relevant stakeholders – especially older people and disadvantaged groups in society. Engagement at grass-roots is urgently needed to complement the 'top-down' digital inclusion campaigns currently in operation in some countries.

The path is understood and waiting to be travelled. The rewards are significant and achieving the vision of digital inclusion presents opportunities for innovation and change in business and society that are even greater than the challenges.

ACKNOWLEDGMENT

Sus-IT ('Sustaining IT use by older people to promote autonomy and independence') is a New Dynamics of Ageing project, funded by the 5 UK Research Councils – AHRC, BBSRC, EPSRC, ESRC and MRC (grant No. RES-353-25-0008).

REFERENCES

[1] Clinton B. Remarks by the President at Massachusetts Institute of Technology 1998 Commencement. 1995; transcript.

[2] Wangberg SC, Andreassen HK, Prokosch H-U, Vagos Santana SM, Sørensen T and Chronaki CE. Relations between Internet use, socio-economic status (SES), social support and subjective health. Health Promot Int 2007; 23: 70–77.

[3] UK Cabinet Office. Enabling a digitally United Kingdom. 2004; Available at: http://webarchive.nationalarchives.gov.uk/+/http://www.cabinetoffice.gov.uk/media/cabinetoffice/corp/assets/publications/reports/digital/digitalframe.pdf. Accessed 01/05, 2014.

[4] Gatto SL and Tak SH. Computer, Internet and Email use among older adults: benefits and barriers. Educ Gerontol 2008; 34: 800–811.

[5] Norris P. Digital Divide? Civic Engagement, Information Poverty and the Internet Worldwide. Cambridge: Cambridge University Press; 2001.

[6] World Summit on the Information Society. Declaration of Principles: Building the Information Society: a global challenge in the new Millennium. 2003.

[7] Information Policy and Access Centre. Public Libraries and Digital Inclusion. Digital Inclusion Survey 2013.

[8] Gorard S. and Selwyn N. Towards a le@rning society? the impact of technology on patterns of participation in lifelong learning. British Journal of Sociology of Education 2005 01/01; 2014/12;26(1):71-89.

[9] Office for National Statistics. Internet Access Quarterly Update, Q1 2014. 2014; Available at: http://www.ons.gov.uk/ons/rel/rdit2/internet-access-quarterly-update/q1-2014/stb-ia-q1-2014.html. Accessed 12/18, 2014.

[10] Digital Unite. Reclaim your independence by getting online!. 2013; Available at: http://digitalunite.com/blog/reclaim-your-independence-getting-online. Accessed 12/22, 2014.

[11] Leicestershire CareOnLine. Your Story. Available at: http://www.leicscareonline.org.uk/careonline_people. Accessed 01/21/15.

[12] Sayago S. and Blat J. About the relevance of accessibility barriers in the everyday interactions of older people with the web. Proceedings of the 2009 International Cross-Disciplinary Conference on Web Accessibililty (W4A) New York, NY, USA: ACM; 2009.

[13] Katz J.E. and Aspden P. Internet dropouts in the USA, the invisible group. Telecomm Policy 1998; 22:327-339.

[14] Emmanouilides C. and Hammond K. Internet usage, predictors of active users and frequency of use. J Interact Marketing 2000(24):17-18-32.

[15] Dutton W.H. and Blank G. Next generation users, the Internet in Britain. Oxford Internet Survey 2011 2011.

[16] Young W., Clarke J., Klima G., Gadag V, Gien L., and Hardill I. Sustaining information and communication technology use among Canadians with at least one activity limitation. Int J Tech Know Soc 2012;7:1-2-10.

[17] Olphert W. and Damodaran L. Older people and digital disengagement: a fourth digital divide? Gerontology 2013;59(6):564-570.

[18] Dutton W.H., Blank G., Groselj D. Oxford Internet Survey 2013 Report: Cultures of the Internet. 2013.

[19] Ramondt L., Sandhu J. and Damodaran L. Staying digitally connected – a study of learning and support provision for older people in seven cities in the UK and the implications for policy and practice. Int J Educ Ageing 2013;3:95-96-114.

[20] Damodaran L., Olphert C.W. and Sandhu J. Falling Off the Bandwagon? Exploring the Challenges to Sustained Digital Engagement by Older People. Gerontology 2014;60:163-164-173.

Challenges in E-government: Conceptual Approaches and Views

Rasim Alguliyev, Farhad Yusifov
Department of Information Society Problems
Institute of Information Technology of ANAS
Baku, Azerbaijan
emails: {rasim@science.az, farhadyusifov@gmail.com}

***Abstract*—The governments invest big amount of funds to realization of e-government projects for further upgrade of services supplied by the government to citizens and reduction of costs in whole world. From this point of view, development of scientific-theoretical principles of forming the e-government is of great importance. Current research reviews several relevant issues regarding e-government such as the definition of e-government, advantages of e-government implementation, the monitoring of forming processes and management, the intellectual analysis of web-resources, information security and electronic democracy problems and it focuses on the challenges of e-government implementation. Results of research may specify researchers in their continued investigation of e-government implementation, especially in the context of developing countries.**

Keywords—Electronic government; public administration, information security, web-analytics; social networks; data mining.

I. INTRODUCTION

Nowadays, the wide implication of information technologies in developed countries is affecting their social-economic development. The number of citizens, centers, organizations, institutes having access to and using internet for satisfying their needs is being rapidly increased. In this situation, there is an increasing need for more mobility and interactivity in transparency principles of public services and neutrality principles from political point of view. Note that the opportunities of political and social technologies in administration are being widened.

In some sources, “electronic government” term is used during translation and different definition, especially in the developing countries. As “electronic government” is currently under construction especially in the developing countries, it has not been fully formed as a definition. As shown in relevant documents, it does not only include the central executive authority, it also includes the three branches of government – the executive, the legislative and the judicial [1]-[3].

The conducted research shows that the definition of implication of the electronic government (e-government) is expanded not only as application of information technologies, but also as a tool of administration of public services in the world. Sometimes, as a key of success of e-government, Customer Relationship Management (CRM) systems are indicated [4][5].

The forming of national e-governments in post-industrial countries is carried out based on reform of all public administration system. The main objective here is the compliance of public administration with Information society. Modern public administration contains its substantial clarity, transparency, competition environment, and responsibility for the outcomes of its actions, increase of the role of ethical requirements, and active mutual relationship with civil society. It is essential that, during the use of Information - Communication Technologies (ICT) in public administration, also other factors affecting the character of socio-political, economic, cultural, mental, and government-society relationships are considered.

Wahid [6] provides a literature review on e-government in the context of developing countries published between 2005 and 2010. Results of research may guide researchers in their continued investigation of e-government implementation, especially in the context of developing countries. Some research direction were provided: paying more attention to research paradigm and methodology, preserving multiculturalism in e-government research, improving the research quality, developing conceptual basis of e-government, etc. [6].

Last research works provide an analytic review of the literature on the diffusion of e-government. In research works analyzed the related literature in the relevant journals and from international conferences in the field of ICT and public administration. Analytical results reveal the main conceptual and architectural principles, research methods, and research topics found in the relevant literature. Main research topics are included: the factors that influence the diffusion of e-government, e-government systems and applications, the impacts of e-government to public authorities and citizens, the relationships between ICT infrastructures and the influence of e-government, etc. [7].

Nowadays, the governments invest big amount of funds to realisation of e-government projects for further upgrade of services supplied by the government to citizens and reduction of costs in whole world. The governments can increase the efficiency of actions and carry out administrative operations more easily by using ICT. By considering this important fact, the specification of researches in direction of e-government establishment and also the most successfulle applied models and their research are remarkably necessary. From this point of view, development of scientific-theoretical principles of forming the e-government is of great importance. By considering the international practice in research, some up-to-date scientific-theoretical problems of forming the e-government has been researched.

II. THE CONCEPTUAL AND ARCHITECTUAL PRINCIPLES OF E-GOVERNMENT

It is known that e-government has started forming in the cross of two centuries. It is known from history that, each transition to new quality has been accompanied by several complications, sometimes by serious crisis. Following this experience, the government can prevent the possible social-economic crisis by modernizing the public administrative mechanisms. In this regard, the government was required to conduct some reforms for the transformation to a new phase in public administration. The transition to a new phase necessitates the conduction of important scientific-research works.

Literature review on e-government shows that potential factors of e-government implementation and classified these factors into four categories: institutional, resource-related, access-related and legal aspects [8]. Chen et al. [9] propose a set of elements for successful implementation of e-government. As for the benefits, such as efficiency and effectiveness in public administration, more countries are working towards adopting e-government [10]. Researchers suggest that e-government adoption is not merely a technological issue; also it is influenced by other factors such as human, social, cultural and economic aspects [11]. The implementation of e-government in developing nations faces many challenges [12].

Alshehri et al. [13] reviewed the updated the available literature about e-government implementation stages, its challenges and benefits. It reviews several relevant issues regarding e-government, such as the definition of e-government, implementation stages, the advantages of e-government implementation and it focuses on the challenges of e-government implementation [13].

According to “Electron government” law of USA dated 2002 [14], this term was accepted as an expansion of access to government information of agencies and government structures by means of information technologies and Internet, also as an implication of information technologies and the use for the increase of efficiency [14]. In official documents, e-government is comprehended as a mutual relationship system with information character of local government authorities and the society by using ICT. “Gartner Group” company reckons that e-government – is the concept of administration by incessant optimisation of services process, participation of citizens in political processes, also by changing of internal and external relations with the help of technical tools, Internet and modern mass media [15].

In some research works, “electronic government” term is defined as Internet-Technologies providing the informative mutual relationship of government authorities with population and civil society institutions. E-government is specified as an integral, socially responsible enterprise having regular counter-relation and open to information.

In general, e-government is specified as a mutual relationship between specialised complex system of public authorities and citizens, civil society and business structures by means of Internet. The following steps of a mutual relationship, Customer-to-Business (C2B) – between citizen and business; Business-to-Business (B2B) – between private companies; Government-to-Citizen (G2C) – between government services (on goverment, departments and regions level) and citizens; Government-to-Business (G2B) – between government and business sectors; Government-to-Government (G2G) – between public authorities can be shown.

Note that the conceptual model of e-government is based on government structure existing in countries of democratic society and market economy (Figure 1b). Conceptually, the approach of this problem, the reforms conducted in public administration in the beginning of 90-s showed the larger share of government in forming of e-government (Figure 1a).

Alongside, it must be considered that the society not only obtains the access to information, but also gets the opportunity to affect the decision-making process of government and participate interactively in the process of preparation of decisions; as a result, the transparency of public sector performance increases.

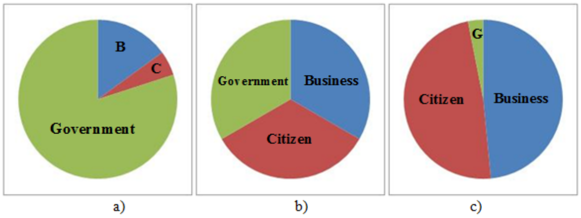


Figure 1. Evaluation model of E-government.

In general, e-government creates new opportunities for development of democracy. It provides the mutual information relationship between citizens and civil society institution and public authorities by means of ICT. In other words, e-government comprises the mutual relations system of citizens, civil society and business-structures, and executive government structures by means of Internet. Implication of ICT in government performance, transparency and accessibility of government information, feedback principle between citizens and public authorities, government responsibility for the decisions made, etc.; issues in different countries are the main characteristics specifying e-government.

It is essential that the transformation to Information society, e-government strategy based on democratice values necessitates the gradual change of government model, the increase of share of civil and business structures, minimization of government share (Figure 1c).

According to definition of European Committee, e-government - is the use of ICT in public structures and improvement of performance of government employees and public authorities in the background of realization of organizational reforms and forming of skills directed to the increase of level of services provided by them [16].

According to the concept of e-government, the whole system of public authorities performs as an integral service organization for the provision of services to citizens. The performance of e-government must be clear, transparent and available in terms of information for citizens. The specific attention is drawn to establishment of feedback mechanism, efficiency of services provision and execution period by using the centralized systems. These all enable to increase either the quality of provision of services provided by the government to citizens, or the performance efficiency of government.

III. THE ISSUES OF ASSESSMENT OF FORMING PROCESSES AND MANAGEMENT OF E-GOVERNMENT

One of the up-to-date issues regarding e-government is the assessment and monitoring its forming processes. It can be justified that e-government is an online environment with the quite complicated structure. On the other hand, e-government is the sum of vertically and horizontally interrelated corporative information spaces.

The issue of establishment of complex indicators system for the monitoring of the efficiency of governance and the use of ICT in different areas, the methodology of practical implication has been started to forming at the end of 1990's [3][17][18]. Nowadays, the existing practical experience and methodical potential for the assessment of the electronic readiness, the monitoring, potential analysis and comparative analysis of governments are present. As such methodologies, one can mention some well-known ones; (i) e-government development index of United Nations regarding the forming and the use of e-government [19], (ii) networked readiness index of World Economic Forum [20], (iii) ICT development index (IDI) of International Telecommunication Union regarding the assessment of Information society [21], (iv) digital opportunity index, an indicators system for the assessment of development level of e-government of European Union countries (Capgemini company) is presented in [22].

The implication of international indexes for the development of methods of the assessment and monitoring of e-government forming processes can be considered as an important factor. Also, note that the position of the country in international ratings has a great importance in terms of the position of the country attained in the region. These indicators exhibit the carrying out of development strategy of Information society of the country.

Note that alongside with the assessment and monitoring of digital differences either at national, corporative or at enterprise level; these are the important information for carrying out of expedient management of the forming process of aimed electronic environment (e-environment).

Taking into consideration the necessity of realization of 5 sequential phases (communication, computerization, networking, informatization, and virtualization) of e-environment forming, the balanced relation must be provided among the separate phases of it. Management centers are the intellectual systems enabling the efficient decision-making bases on the indicators characterizing the progress of the process.

The following can be shown as the indicators characterizing the virtualization, socialization phase in e-environment:

- The indicators characterizing the social networks created in considered e-environment;
- The indicators characterizing the classification and activeness (age, specialty, gender, space, time, etc.) of users;
- Classification and rating indicators of used contents;
- The indicators characterizing the transparency, accessibility and sequences of information in e-environment;
- The indicators characterizing the virtual relations established and contents turning over in e-environment, etc.

It is essential that the solution of several problems (technological, normative-legal base, cadres' education, scientific, etc.) is an important condition. By taking into consideration the leading practice, e-government establishment in country must be formed based on both horizontal and vertical management principles. From this point of view, each institution included in e-government must have an action plan, the indicators characterizing its plan must be specified the management of this process must be executed and the continuing (or on specific cycles) monitoring must be carried out.

Nowadays, the monitoring and assessment issues have a timely importance for the realization of e-government projects, programs specifically. From this point of view, it is complicated to assess on which phase country is in realization of e-government program. Different sources rely on information that is not always based on honesty, reliance by making specific results for themselves. From this point of view, there is a need for establishment of complex system of assessment and monitoring. The indicators accepted at international level and parameters meeting the local needs must be included in the system itself.

IV. INTELLECTUAL ANALYSIS ISSUES OF WEB-RESOURCES PERFORMING IN E-GOVERNMENT ENVIRONMENT

While considering the e-government programs carried out in different countries of the world, it becomes clear that e-government will be more accessible and efficient under the "single window" principle in the near future. This is mainly related to rapid development of content mining methods, web technologies and social networks [23]-[26]. From this point of view, the analysis of web-resources and development of management mechanisms is of great importance in carrying out the e-government projects.

The implication of web content mining, web-analytics and social networks are strong tools in improvement of e-government management effectiveness and establishmet of feedback mechanism. If ones consider that one of the main issues in realisation of e-government programs are the analysis of web-resources and establishment of the effective management policy, then the implication of innovations, new technologies widens the communication capacity significantly. This, in turn, enables to achieve new integration forms between business sector and citizens.

Web-resources creates online interactive social communication environment between public authorities and citizens. More information is gathered in this environment eventually. Thus, new opportunities are created for the intellectual analysis of web-infrastructure and more efficient management of the society.

The internal structure of the electronic community existing inside each online environment in disguise can be revealed by applying the social networks theory. The analysis of e-government web-infrastructures, web mining technology can be implied for obtaining the following information [23]-[25]:

- Which issues are mostly discussed by the citizens;
- The monitoring whether the discussions are related to government sector;
- The classification based on different criteria's (space, time, age, specialty, activeness, etc.) of citizens applying mostly to which institutions;
- The online monitoring of realisation status of requirements claimed against web-resources;
- The classification based on different criterias (countries, institutions, issues, time, etc) of inquiries to web-resources from foreign countries performing in e-government environment, etc.

An effective web analytics of sites, portals and also, sites providing online services to citizens – is revealing existing program, technical, content related errors and adjustment to requests of citizens, and users. By using web-analytics, the reasons for leaving the site by users, their actions, and behaviors at web-site regarding a site or particular service can be revealed. It is clear that web-analytics is not limited with particular statistics and enables to obtain more detailed information for analysis.

To analyze log-files gathered in servers, information gathered in e-mails play a prominent role in effective decision-making by e-government parties in the process of establishment of online relations between citizens and public authorities [23][24]. This, in turn, enables the development of feedback mechanisms for e-government management.

V. Information security provision issues of e-government

People became more dependent on information as society gets computerized. Non-provision the information security can cause major consequences for the society. The priorities of information security in a particular country are specified based on the balanced ratio of government, society and citizens interests. As one of the main components of the safety of society, the duties of information security are the confidentiality of information, information integrity, information accessability and the fight with harmful computers [27][28].

Information security is critical in e-government initiatives. Confidentiality of any information available on the network is crucial point. All data and whole the government document and other important material have to be protected from unauthorized persons in case of realizing e-government initiatives. Information security is critical for successful realization of such initiatives [29].

E-government forming has created new options of interactions between government organizations and citizens in the delivery of government services. Of course, these interactions have needs for maximum information security. In general there are five broad requirements of information security in e-government: confidentiality, integrity, availability, authenticity, and accountability [30]. Developing effective information security practices requires of both the technological perspective and the socio-organizational perspective [31][32].

Efforts have been made to develop frameworks for ensuring information security in organizations. For example, propose an information security culture framework to heighten information security awareness in organizations [33]. Martin [34] recommends a total quality management based framework to manage information security in organizations. These studies have shown the opportunities of individual frameworks for maintaining information security in organizations from different perspectives.

The several issues with technical and administrative-legal characters must be solved for carrying out the e-government program. The preparation of mutual relationship reglaments, the creation of government services classification, also integral technical architecture, realization of program platform and the provision the information security can be indicated among them.

For the provision of normal performance of e-government, it is necessary to provide the security of each level constituting e-government.

In general, the up-to-date issues in the framework of provision the information security of e-government we can classified as following:

- Development of conceptual-architectual models for provision and management of e-government information security and sustainable performance;
- Development of models for the analysis of information security risks and management;
- Development of cybercrime defences technologies;
- Revealing the disguised criminal social networks creating threats for e-government environment and development of analysis methods;
- Development of intellectual monitoring system of corporative network environment;
- Development of spam busters methods and algorithms by means of data mining technologies;
- Protection of individual information in e-government environment and development of user-oriented security mechanisms;

- The creation of Computer Emergency Responce Team (CERT) in e-government environment;
- Investigation of information war, information attack and information attack defences technologies and development of new methods and algorithms.

Note that a complex and systematic approach is required to information security provision issues of e-government. With the development of Information society, the necessity of establishment of integral and multilevel nation-wide information security system appears in the process of e-government building. In general, building the Information society perplexes the provision the information security of countries and the sole fight against threats of different nature and scale. Thus, building the global information security environment must be of interest of all countries, civil societies, companies and people.

VI. E-GOVERNMENT AND ELECTRONIC DEMOCRACY ISSUES

Different factors are considered as variables which impact electronic democracy (e-democracy) based on e-government literature. E-government is not only a term that refers to the transformation of public services, so-called e-governance, but also about the transformation of political systems, so called e-democracy. E-democracy is considered such organisation form of citizens' social-political activity that the wide use of ICT provides the establishmet of more effective relations at new level either among citizens, or between citizens and government bodies, civil society and business sector [35][36]. In other words for the strengthening of democratic institutions, the expansion of participation of citizens in political activity and the use of ICT constitues the essence of e-democracy. E-democracy term means the consideration of citizen's thoughts and the engagement of citizens and organisations to political relations and processes. In this phase, the issue of how close the citizens are engaged in social-political processes is characterized with electronic citizen problems.

Starting from initial phases of awakening of e-democracy, the provision of access opportunities to socially important information of government bodies by the citizens was constrained by creating of voting opportunities regarding partiular decisions of the government [35]-[37]. The further development has widened the opportunities of both sides, the citizen and the government and close participation of citizens in social-political processes was provided. This meant the establishment of the opportunity of expressing the thoughts by citizens in any level of decision-making and the notworthy increase of transparency.

The following are related to e-democracy mechanisms [35][36]:

- Electronic voting (voting with mobile phone, Internet-elections, etc);
- The collective discussion mechanisms of subjects with social-political content and socially important issues in online regime;
- The forming mechanisms of online communities, groups, social networks;
- The mechanisms of realisation of citizens' incentives;
- Citizens' control mechanisms on public authorities' performance, etc.

Figuratively speaking, the main currency of the democracy is information and communication. With these two, the citizens are self-organised, start to govern themselves and e-citizen is shaped. Social networks, blogs and others play a prominent role in forming of civil society. For the civil society, this means that the horizontal relations system is being shaped and self-governance opportunities (municipalities, non-governmental organizations, etc.) are created, i.e., it takes some functions of the government. Other functions are carried out by the business sector.

Alongside with what was mentioned above, the transition to Information process does not require only an automation of existing processes in government management, but also their re-building based on particular interests of citizens and a group of interests of the society. Considering those principles, nowadays, the direct e-democracy projects are not sufficiently supported by the business sector. E-democracy concept has several inconsistencies and is reasonably criticized. Hence, recently e-government concept is dominant in socio-political and scientific literature, which is the basis for carrying out the reforms in government management sphere by means of ICT.

VII. CONCLUSION

During the review of practice of leading countries, it is revealed that existing e-government projects have different objectives and different models, conceptual approaches are suggested by institutions, organizations for the development of e-government. By considering this fact, the inspection of research conducted in direction of e-government establishment in international practice is remarkably necessary. By considering the international practice in research, some up-to-date scientific-theoretical problems of forming the e-government has been researched. In research work, reviews several relevant issues regarding e-government such as the definition of e-government, advantages of e-government implementation, the monitoring of forming processes and management, the intellectual analysis of web-resources, information security and electronic democracy problems and it focuses on the challenges of e-government implementation. Some conceptual and architectural principles of forming the e-government are investigated and some recommendations are given.

Important research directions were specified by considering the main principles of e-government concept. Results of research may specify researchers in their continued investigation of e-government implementation, especially in the context of developing countries.

In future researches, the specific attention will be drawn to establishment of feedback mechanism, efficiency of services provision and execution period by using the centralized systems. These all enable to increase either the quality of provision of services provided by the government to citizens, or the performance efficiency of government.

ACKNOWLEDGMENT

This work was supported by the Science Development Foundation under the President of the Republic of Azerbaijan - Grant № EİF/GAM-2-2013-2(8)-25/03/1

REFERENCES

[1] S. M. Alhomod and M. M. Shafi, "Best Practices in E government: A review of Some Innovative Models Proposed in Different Countries," International Journal of Electrical & Computer Sciences, vol. 12, no. 01, 2012, pp. 1-6.

[2] Definition of E-Government, World Bank, 2002, www.worldbank.org [retrieved: November, 2014]

[3] M. Yildiz, "E-government research: Reviewing the literature, limitations, and ways forward," Government Information Quarterly, 24, 2007, pp. 646–665.

[4] M. Vulić, J. Dadić, K. Simić, D. Mazinjanin, and A. Milić, "CRM e-government services in the cloud," www.fos.unm.si [retrieved: November, 2014]

[5] L. M. Lowery, "Developing a Successful E-Government Strategy," http://unpan1.un.org [retrieved: November, 2014]

[6] F. Wahid, "The Current State of Research on eGovernment in Developing Countries: A Literature Review," In H. Scholl, M. Janssen, M. Wimmer, C. Moe & L. Flak (Eds.), Electronic Government, Springer, vol. 7443, 2012, pp. 1-12.

[7] H. Zhang, X. Xu, and J. Xiao, "Diffusion of e-government: A literature review and directions for future directions," Government Information Quarterly, vol. 31 (4), pp. 631–636.

[8] Sh. Rahman, N. Rashid, A. Yadlapalli, and L. Yiqun, "Determining factors of e-government implementation: a multi-criteria decision–making approach," Proceedings of PACIS 2014 Chengdu, China, 24 – 28 June, 2014.

[9] Y. C. Chen and R. Knepper, "Digital Government Development Strategies. Lessons for Policy Makers from a Comparative Perspective," In Electronic Government Strategies and Implementation, Idea Group publishing, 2005.

[10] S. Ozkan and E. I. Kanat, "e-Government adoption model based on theory of planned behavior: Empirical validation," Government Information Quarterly, vol. 28(4), 2011, pp. 503-513.

[11] M A. Shareef, V. Kumar, U. Kumar, and Y.K. Dwivedi, "e-Government Adoption Model (GAM): Differing service maturity levels," Government Information Quarterly, vol. 28(1), 2011, pp. 17-35.

[12] K. J., Bwalya, T. Du Plessis, and C. Rensleigh, "E-government implementation in Zambia – prospects," Transforming Government: People, Process and Policy, vol. 8 (1), 2014, pp. 101-130.

[13] M. Alshehri and S. Drew, "E-government principles: implementation, advantages and challenges," Intnational Journal Electronic Business, vol. 9, no. 3, 2011, pp. 255-270.

[14] E-Government Act of 2002, USA, www.gpo.gov [retrieved: October, 2014]

[15] Gartner company, www.gartner.com [retrieved: October, 2014]

[16] ICT for Government and Public Services, European Commission, http://ec.europa.eu [retrieved: November, 2014]

[17] D. D. Potnis, "Measuring e-Governance as an innovation in the public sector," Government Information Quarterly, 27, 2010, pp. 41–48.

[18] C. E. Koh, V. R. Prybutok, and X. Zhang, "Measuring e-government readiness, Information & Management," 45, 2008, pp. 540–546.

[19] The United Nations E-Government Survey 2014: E-Government for the Future We Want, www.unpan.org [retrieved: October, 2014]

[20] Global Information Technology Report 2014,www.weforum.org [retrieved: December, 2014]

[21] Measuring the Information Society 2012, www.itu.int [retrieved: November, 2014]

[22] eGovernment Benchmark Framework 2012-2015, http://ec.europa.eu [retrieved: November, 2014]

[23] A. Kaushik, "Web Analytics 2.0 - The Art of Online Accountability and Science of Customer Centricity," Wiley Publishing, Inc. 2010, 447 p.

[24] R. M. Alguliyev, R. M. Aliguliyev, and F. F. Yusifov, "Automatic Identification of the Interests of Web Users, Automatic Control and Computer Sciences," vol. 41, no. 6, 2007, pp. 320-331.

[25] H. Liu and V. Keselj, "Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests," In: Data and Knowledge Engineering, vol. 61, no. 2, 2007, p. 304-330.

[26] J. Vosecky, Dan Hong, and V. Y. Shen, "User identification across multiple social networks," Proccedings of First International Conference on Networked Digital Technologies, 2009, pp. 360–365.

[27] Creation of a global culture of cybersecurity, 2002, www.un.org [retrieved: October, 2014]

[28] Global Cybersecurity Agenda, 2008, www.itu.int [retrieved: November, 2014]

[29] Sh. Singh and S. Karaulia, "E-Governance: Information Security Issues," International Conference on Computer Science and Information Technology (ICCSIT'2011), Pattaya, 2011, pp. 120-124.

[30] D. Zissis and D. Lekkas, "Securing e-Government and e-Voting with an open cloud computing architecture," Government Information Quarterly, 2011, pp. 239-251.

[31] B. Bulgurcu, H. Cavusoglu, and I. Benbasat, "Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness," MIS quarterly, no. 3, 2010, pp. 523-548.

[32] G. Dhillon and J. Backhouse, "Current directions in IS security research: towards socio ‐ organizational perspectives," Information Systems Journal, 2001, pp. 127-153.

[33] A. Da Veiga and J. H. P. Eloff, "A framework and assessment instrument for information security culture," Computers & Security, 2010, pp. 196-207.

[34] C. Martin, A. Bulkan, and P. Klempt, "Security excellence from a total quality management approach," Total Quality Management & Business Excellence, 2011, pp. 345-371.

[35] A.-V. Anttiroiko, "Building Strong E-Democracy - The Role of Technology in Developing Democracy for the Information Age," Communications of the ACM September, vol. 46, no. 9, 2003, pp. 121-128.

[36] A. Meier, "eDemocracy & eGovernment," Springer-Verlag. Berlin, Heidelberg, 2012

[37] M. Hilbert, "The Maturing Concept of E-Democracy: From E-Voting and Online Consultations to Democratic Value Out of Jumbled Online Chatter," In: Journal of Information Technology & Politics, vol. 6, 2009, pp. 87–110.

Development of the Model of Dynamic Storage Distribution in Data Processing Centers

Rashid Alakbarov
Institute of Information Technology
of ANAS, Baku, Azerbaijan
rashid@iit.ab.az

Fahrad Pashayev
Institute of Control Systems after
Academician A. Huseynov of
ANAS, Baku, Azerbaijan
pasha.farhad@gmail.com

Mammad Hashimov
Institute of Information Technology
of ANAS, Baku, Azerbaijan
m.hashimov@iit.ab.az

***Abstract*—The paper reviews an optimal distribution of storage resources among the users in data processing centers. Stochastic model of the dynamic distribution of storage resources is proposed. The model ensures the use of storage resources without wasting.**

Keywords*—data processing center; cloud computing; storage capacity; Markov process; stochastic model.

I. Introduction

Nowadays, computing and storage resources of personal computers are not sufficient for the solution of complex problems requiring big computing and storage resources such as real time modeling of physical and chemical processes, nuclear reactions, global atmospheric processes, economic development in various fields of science, as well as Cryptography, Geology, development of new drugs. Supercomputers with high performance computing and big storage are widely used in the above-mentioned issues [1]. As a strategic product, the high price of supercomputers reduces its availability for many countries to be used in scientific and technical research. However, these countries have demand for big computing resources. On the other hand, computing and storage resources of the data processing centers connected to the computer networks are not used effectively. Researches show, that only 60-70% of computing and storage resources of computers manufactured by giant companies (Intel, IBM, Google, etc.) are used effectively [2]. In this case, remaining unused computing and storage resources of data processing centers can be used to solve complex problems. Applying remote access to the data processing centers in daily practice with the help of high speed communication channels open up new possibilities for the users. Now, the quantitative increase of opportunities of users to get information caused qualitative change in the organizational principles of distributed computing systems in the networks.

At present, research is conducted out for an effective use of computing and storage resources of data processing centers with the help of Cloud Computing. Such systems with big computing and storage resources are based on computer networks, provided with high-speed communication channels. Cloud Computing enables organizations to use computing and storage resources of data processing centers more efficiently. The concept of Cloud Computing provides the development and utilization of infrastructure and software of computer technology in the network. With the help of this technology, the user data is stored and mined on Cloud Computing servers, at the same time, the results are viewed through browsers [3]. Cloud Computing allows the users to access powerful computing and storage resources, and at the same time, the user is not interested where these resources are located and installed. The paper is dedicated to the optimal distribution of storage resources among the users. The proposed model allows the data processing center to attract more users providing optimal distribution of available storage and system resources among the users.

The content of the article is organized as follows:

- State of the memory use in data center was set to change as Markov process;
- The characteristics of Markov process of the memory usage change were identified;
- Recommendations were given for the use of obtained results.

II. Developing the Model of Dynamic Storage Distribution in Data Processing Centers

Let us analyze the process of dynamic storage distribution among the users of the systems where Hypervisors are applied. The process of dynamic storage distribution is modeled as Markov process [4]-[7]. Let us suppose that *M* is the number of users specified in advance and suggesting storage need. In this case, *m*-th user ($m \in [1, M]$) uses *Vm* amount of memory. Thus, an amount of memory required by users is as follows:

$$V_1, V_2, \dots, V_M. \tag{1}$$

If V_t denotes storage volume that is used instantly at any t time

$$V_{tmin} = 0, V_{tmax} = \sum_{m=1}^{M} V_m. \tag{2}$$

Nevertheless, in practice,

$V_{tmax} = \sum_{m=1}^{M} V_m$ is almost impossible. In the peak of storage use it is practically as follows:

$$V_{peak} \leq 0,7 * V_{tmax}. \tag{3}$$

This may lead to attracting additional users.

Each storage user can apply for storage at random moments of time, regardless of its physical identity and functionality. Therefore, the process of memory use is determined by a random *Vt* storage volume used instantly at any time *t*. In this case, state space of the process is defined by the storage of different capacity in use. The process of transition from one state to another does not depend on the transition path. This transition depends on the current state, and it is one of the signs of Markov process. It is sometimes called the process without memory.

The second key feature of Markov processes the finite number of states. In our case, the required number of different volumes (states) is finite. Thus, M number of users can apply for storage in a short time interval (time instant) 0,1,2,... . Storage capacity defining the process state depends on these combinations. Obviously, *n* number of *M* users can be choosen from

$C_M^n = \frac{M!}{n!\,(M-n)\,!}$ methods. That is, different combinations C_M^n with *n* number of userscan be set up from *M* number of users. If n ∈[0; M] the number of all possible combinations can be as follows:

$$K = C_M^0 + C_M^1 + ... + C_M^n + ... + C_M^M \qquad (4)$$

According to Newton binominal [8]-[9]:

$$\text{As}(1+1)^M = C_M^0 + C_M^1 + ... + C_M^M, \qquad (5)$$

$$K = C_M^0 + C_M^1 + ... + C_M^M = (1+1)^M = 2^M. \qquad (6)$$

In other words, the number of possible different states are $K = 2^M$, which is finite. Thus, Markov process is covered in this state.

So the model of the need for storage in the Data Center by M number of users is developed as in Markov process. The process parameters are defined as follows:

1. E_1 , E_2 , ... , E_K - state set.

As it is mentioned above, this set is finite and can be defined as $K=2^M$. Each state corresponds to regular order of storage volumes which can be required in different combinations.

2. Stochastic transition matrix**.** The matrix for continuous-time processes is an intensity matrix of the transition from one state to another. Intensity of transition from E_i state to E_j can be denoted by g_{ij}.

$$G = \begin{pmatrix} g_{11} & \cdots & g_{1K} \\ \vdots & \ddots & \vdots \\ g_{K1} & \cdots & g_{KK} \end{pmatrix} \qquad (7)$$

3. Probability vector is the key characteristics of the modeled process. Let $P_t = \{p_1^{(t)}, p_2^{(t)}, \dots, p_K^{(t)}\}$. This vector gives complete information about the process at any time t: where $0 \le p_i^{(t)} \le 1$, $\sum_{i=1}^K p_i^{(t)}=1$ states are covered.

In continuous-time Markov processes, the following system is used to calculate these probabilities:

$$\begin{cases} \frac{dp_j^{(t)}}{dt} = \sum_{i=1}^K p_i^{(t)} * g_{ij} \\ p_j^{(0)} = p_j^{(0)} \\ \quad ... \\ \sum_{i=1}^K p_i = 1 \end{cases} \qquad (8)$$

This system can be solved by approximate methods. If the process has ergodic features, it is a system of algebraic equations, which has only one solution.

$$\begin{cases} \sum_{i=1}^K p_i g_{i1} = 0 \\ \sum_{i=1}^K p_i g_{i2} = 0 \\ \quad ... \\ \sum_{i=1}^K p_i g_{iK} = 0 \\ \sum_{i=1}^K p_i = 1 \end{cases} \qquad (9)$$

As the ergodic process shifts to stationary mode, it does not depend on time and it shifts to

$$P(t) = (p_1^{(t)}, p_2^{(t)}, ... , p_K^{(t)}) = (p_1, p_2, \dots, p_K) = P \quad (10)$$

The equation (10) is the vector of transition probabilities of the process from the i-th state into another. This vector provides the probable distribution of the condition of the process in which it will be in the future.

If $j \in [1, K-N]$ the probability of the calls to the computing resources of the users of $i \in [j+1, j+N]$ number is calculated as follows: $\sum_{i \in [j+1, j+N]} p_i$. This sets the condition of attracting new users as well.

III. CONCLUSION

The paper reviews an optimal distribution of storage resources among the users in data processing centers. The model of the dynamic distribution of storage resources is proposed. The model provides that the storage resources allocated for any purpose hold the space in the system as much as they are used. Thus, resources are allocated as much as they are used without wasting. This is beneficial for both cloud provider and the user. Accordingly, the user does not pay for reserved resource, but only for the actual resource use, and the provider reduces unnecessary purchase and installation of additional equipment's. Furthermore, it can be able to offer the same service for lower price, which leads to greater user involvement in this type of services.

REFERENCES

[1] V. V. Voevodin, Vl. V. Voevodin. Parallel computing. St. Petersburg. "BHV – Petersburg", 2002.

[2] I. Yu. IBM invests "Cloud Computing", www.pcnews.ru/news/ibm-300.

[3] R. M. Alguliyev, R. K. Alekperov. Cloud Computing: Modern State, Problems and Prospects. Telecommunications and Radio Engineering, 2013, vol.72, no.3, pp.255-266.

[4] A. A. Markov. "Extension of the limit theorems of probability theory to a sum of variables connected in a chain". Reprinted in Appendix B of: R. Howard. Dynamic Probabilistic Systems, volume 1: Markov Chains. John Wiley and Sons, 1971.

[5] S. Meyn. Control Techniques for Complex Networks. Cambridge University Press. 2007, p.615.

[6] S. P. Meyn and R. L. Twedie. Marcov Chains and StochosticStability. Springer-Verlag 1993, p 552.

[7] H. A. Taha. Operations Research: An Introduction. Seventh Edition M:, “Williams”, 2007., p 901.

[8] G. Ronald, K. Donald, P. Oren. “(5) Binomial Coefficients” Concrete Mathematics (2nd ed.) Addison Wesley. 1994. PP. 153-256. ISBN 0-201-55802-5. OCLC 17649857.

[9] G. E. Shilov, (1977). Linear algebra. Dover Publications. ISBN 978-0-486-63518-7.

Reliable Document-centric Processing in a Loosely Coupled Email-based System

Magdalena Godlewska
University of Gdansk
Faculty of Mathematics, Physics and Informatics
Gdansk, Poland
Email: maggod@inf.ug.edu.pl

***Abstract*—Email is a simple way to exchange digital documents of any kind. The Mobile INteractive Document architecture (MIND) enables self-coordination and self-steering of document agent systems based on commonly available email services. In this paper, a mechanism for providing integrity and reliability of such an email based agent system is proposed to cope with message soft or hard bounces, user interrupts, and other unexpected events. This mechanism consists of a system acting as a ”ground control” for migrating documents and a set of protocols that improve the implementation of document coordination patterns. It allows for an estimation of the global state of a distributed loosely coupled agent system and making top-down decisions in unforeseen situations.**

***Keywords*–multi-agent systems; collaborative work; electronic documents; email-based systems**

I. Introduction

A knowledge-based organization is a management idea, describing an organization in which cooperating people use knowledge resources to achieve organizational goals. People are the key intellectual resource but only collaboration with other workers in accordance with the organizational procedures enables converting knowledge of individuals to knowledge of organization [1].

Knowledge workers communicate through the exchange of documents constituting units of information. Nowadays, email has a dominant position in the computer mediated communication and document exchange in the workplace [2]. Email messaging provides an easy to use simple textual form and allows to disseminate attachments in any format to one or multiple recipients.

The MIND architecture [3] is a proposition of a document-centric uniform interface to provide both effective communication of content and coordination of activities performed on documents. MIND is a solution that augments email messaging with proactive documents, capable of initiating process activities, interacting with individual workers on their personal devices and migrating on their own between collaborators. Thus, each MIND document is a mobile agent. Document-agents have built-in migration policy to control their own workflow and services to proper processing contained information. Section II contains a more detailed overview of the MIND architecture.

The migration path of the document-agent contains all information and status of the workflow process to perform it locally on users' devices. An email client installed on each worker's device participating in the process needs to be extended with functionality to activate the document-agents and switch documents between the activity and transition phases of the workflow. This special email client with workflow enactment capability has been implemented as a Local Workflow Engine (LWE) [4]. All LWEs participating in the process and performing independently form together both a loosely-coupled agent system and a distributed workflow enactment service. Section III outlines generic functionality of LWE and the idea of distributed workflow enactment service.

In the LWE-based MIND system, individual knowledge workers perform activities on documents independently, using their personal devices, and yet collaborate on achieving a common goal. This is possible owing to the migration policy embedded in each document. This policy defines for each document a workflow process composed of specific document-flow patterns that provide process wide coordination. The document-flow patterns [4] are the result of analysis of the coordination patterns proposed by van der Aaalst [5] under the assumption that email is the transport layer for document migration. The work of van der Aaalst shows that a relatively small and well defined set of collaboration patterns contains building blocks of arbitrary complex workflow processes in real organizations. Thus, the document-flow patterns that directly follow the collaborations patterns for the proposed MIND architecture enable modeling and coordination of any workflow process.

The crucial services for MIND implementation are executability and mobility. The former involves activating document-agents to enable their autonomous execution, while the latter involves transporting them between users' devices in accordance with the migration policy. These services have already been implemented and described before [3][4]. In the prototype system, mobility has been implemented based on email as the transport layer.

In most cases, these mentioned services are sufficient to properly perform the MIND agent system. However, in human organizations, some situations unforeseen by the designer of the process may occur and the reliable system should be able to cope with them. For a distributed loosely coupled and interactive system it is impossible to determine a global state of all document-agents. Consequently, it is not possible to determine where all documents are located at a specific time. A document may be lost and it often remains beyond the knowledge of the authors, who have edited it earlier. There are various reasons why documents may be lost: a problem with a transport layer, an error of a local environment or an unexpected user behavior. The initial workflow process definition makes it possible to search for documents in the specified places. However, the process may be modified during its execution. Therefore, a document may take a path that was not originally designed and a document originator has

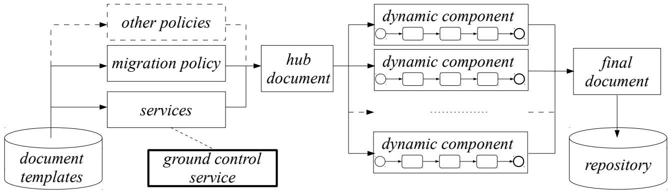


Figure 1. The MIND document lifecycle

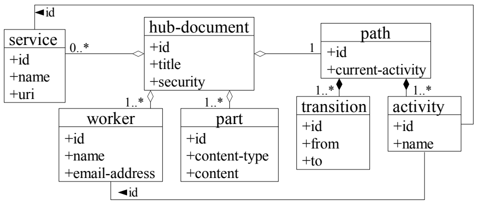


Figure 2. Dynamic form of the MIND architecture [4]

not information about its definition.

Thus, the *reliability* of the MIND agent system is a service that makes the system more useful and trustworthy than the typical email-based communication. It allows for an estimation of the global state of a distributed loosely coupled system, taking into account transport layer errors, unforeseen actions of users and process modifications.

Thus, Section IV identifies problems associated with distributed workflow execution. Section IV-A focuses on the problems associated with email as the transport layer for the MIND documents, while Section IV-B presents problems that may occur in specific document-flow patterns. In particular, it is interesting the canceling pattern due to the loosely-coupling principle at operating of the MIND agent system.

Section V presents a concept of a ”ground control” service which introduces the ability to track document-agents globally and solve some of the problems associated with the documents flow. The service is designed to receive signals containing the status of the document from the LWE clients and send control signals to LWEs that resolve situations incompatible with the designed workflow. The proposed syntax of a notification sending to the ”ground control” service is adapted to document-flow patterns. Further, this section outlines two pilot implementations of the ”ground control” service – one using the Handle System [6], and another based on an email-based notification system.

Section VI surveys previous work related to a document-centric processing and a reliability of workflow enactment in distributed loosely-coupled systems.

II. The MIND architecture

The MIND architecture enables the new agent-based distributed processing model. Traditionally, electronic documents have been static objects downloaded from a server or sent by email. MIND allows static documents to be converted into a set of dynamic components that can migrate between collaborative workers according to their migration policy.

The concept of the MIND document life cycle is illustrated in Figure 1. At the beginning of the knowledge process, some originator forms a *hub document* based on document templates that includes migration policy, which specifies the steps of the process and services that will be performed on different parts of the document during the process. The hub document is changing to mobile components that meet their mission in the distributed agent system. Each component performs its migration policy and interacts with workers of the organization.

The MIND architecture makes possible a radical shift from *data-centric* distributed systems, with hard-coded functionality, to flexible *document-centric* ones, where specialized functionality is embedded in migrating document components and some generic or supporting services are provided by local devices or external servers. The essence of the MIND architecture is that the documents have capability of self-organization and self-steering during the process execution.

Figure 2 outlines the dynamic form of the MIND architecture. It includes five components: *hub-document* is the main component and it contains basic information about the document, *worker* component contains data about workers who participate in the process, *part* component defines parts of the document, *service* component contains information about services that can be performed on different parts of the document during the process, and *path* component defines migration policy of each part of the document. It specifies the steps of the process and activities that should be performed at each step of the process.

The service objects provide document functionality that makes it proactive. Three types of services are possible: *embedded* that are transferred together with the document, *local*, which may be acquired by the document components from local worker’s device, and *external*, called on the remote hosts by the worker’s system at the request of arriving document.

III. Distributed workflow enactment

A key feature of the MIND architecture is physical distribution of business process activities, performed dynamically on a system of independent personal devices. MIND documents have built-in process definition and functionality (the respective path and embedding service components mentioned in the previous section). This makes them agents, which are autonomous and mobile. Especially, they are independent of any particular platform supporting workflow enactment and they are capable of launching individual activities onto various workers’ devices which maintain process coordination across the organization.

Workflow enactment service interprets the process description and control sequencing of activities through one or more cooperating workflow engines [7]. Even if the workflow engines are distributed, workflow enactment is centralized in most of the implementations, because the control data must be available for all engines. In the MIND architecture, all data needed for workflow enactment are embedded in documents [4]. This allows for implementation of distributed workflow enactment service consisting of LWEs.

The idea of the distributed workflow enactment service built on top of LWE clients and email transport layer is illustrated in Figure 3. In the prototype system, LWE was implemented as lightweight email client installed on personal devices of each worker. Each LWE is independent of other LWEs, so it can be implemented in any technology and adapted

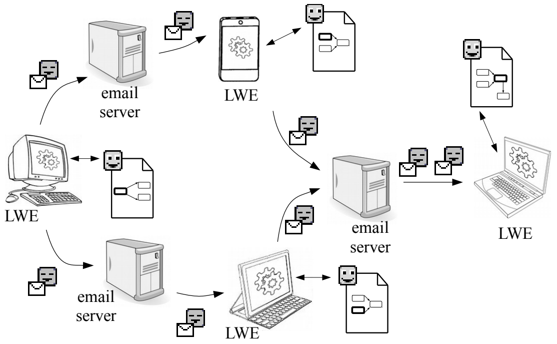


Figure 3. Distributed workflow enactment service based on LWE clients and email transport layer (LWEs are symbolized as gear wheels).

to requirements of particular devices, especially mobile devices such as tablets and smartphones. Also, it may be implemented as a plug-in to existing email clients.

States of the LWE correspond to the phases of a document lifetime and the initial state is when a message with a document is received, i.e., noticed by LWE in the worker's mailbox. The LWE downloads the document on the local device and activates it, which means launching its embedded functionality. The activated document may interact with the knowledge worker, his/her local system and some external services. The interaction begins with obtaining the document path component and determining the current activity that should be performed in this particular step of the process. If the next activity is intended for another worker, the document is serialized, packed and sent as an attachment to the next worker's email address.

LWE is capable of recognizing and executing all document-flow patterns contained in the path component. More details about the document-flow patterns are in Section IV-B, which presents a discussion about their execution in a loosely-coupled distributed system consisting of the LWE clients.

IV. Problems of reliable workflow execution

MIND and LWE clients form a distributed workflow enactment system in which the coordination of activities is based on control data contained in the documents. In most cases, it ensures that the documents arrive at a specific location at a specific time. Nevertheless, some situations unforeseen by the designer of the process may occur in loosely-coupled system.

First of all, the document may be lost: during the transfer by email, due to failure of the local system, accidentally deleted by the user. The transfer of the document may also be delayed to miss the designed deadline. The knowledge worker may also make a decision unforeseen by the workflow, e.g., cancel some document flow or modify the workflow path, which is just not possible in typical message passing via email.

Figure 4 shows the path of the document from a sender to a recipient and indicates points where some problems may occur. Points ②–④ are associated with several well known email transport layer problems briefly described in Section IV-A, while points ① and ⑤ indicate problems with document-flow patterns execution by LWE clients, detailed in Section IV-B.

A. Email transport layer problems

Email message is a simple textual form combined with attachments in any format. It can be sent to one or multiple recipients and supports asynchronous work. Email mechanisms have a reputation of being robust and trustworthy since its invention a few decades ago, as email messages reach their recipients in most cases without problems. Nevertheless, there is a list of problems associated with the delivery of messages.

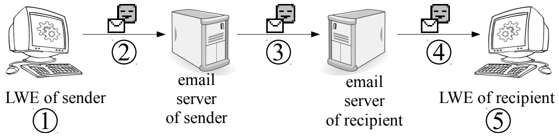


Figure 4. LWE to LWE connection. The numbers indicate points, where some problems with the transport of the document may occur.

The first step in the email processing model is to submit email message by an email client (Mail User Agent – MUA) to a sender Simple Mail Transfer Protocol (SMTP) server (Mail Transfer Agent – MTA) [8]. Figure 4 indicates it as point ②. This step may fail due to the lack of network connection, incorrect SMTP server configuration or SMTP server failure. The message usually remains in the sender outbox and the email client tries to send it again. Configuration of SMTP server for LWE client does not differ from the configuration of other email clients and does not require any special functionality. Temporary lack of network connection is a typical situation for mobile devices. SMTP server failure is a rather transient situation that can be solved by resending the email message.

In the next step, sender MTA transfers messages to the receiver MTA mostly by SMTP protocol (point ③ in Figure 4). SMTP server should deliver the message or notify about any problem [8]. The SMTP reply consists of a three digit number often followed by some text for the human user. The message may be rejected, however, in a transient or permanent way. In transient situations, the SMTP client should try to send the message again. In the case of permanent errors, the SMTP client should not repeat the exact request. After a failed attempt to send a message, the sender SMTP agent sends a notification message to the mail user agent. This notification message is known as a Delivery Status Notification (DSN) or email bounce [9].

Nevertheless, receiving of email bounces does not necessarily mean that the message has not been delivered and, conversely, the lack of notice does not necessarily mean that a message has arrived to the recipient. For instance, the receiver SMTP server may silently drop message to protect themselves from attacks [8]. Many SMTP servers are configured to block messages categorized as spam based on DNS blacklists or anti-spam filters [10].

Receiving a message by the SMTP server and placing it in a user's mailbox does not imply that the user will read it. Point ④ in Figure 4 indicates the problem of the recipient's email server – email client communication. Firstly, some messages may be marked as spam and placed in the spam folder in user's mailbox. In this case, the frameworks to build mail applications (like Java Mail [11] and IMAP – Internet Message Access Protocol [12] used in the LWE implementation) often enable access to the spam folder. In fact, also the email client may have its own spam filters and other solutions to manage received messages automatically, like the automatic responses software (e.g., ”out of office” message) [13].

Next to email bounces and automatic responses there is one more type of notifications, the Message Disposition Notifications (MDNs) [14]. These notifications are intended to report of the disposition of a message after it successfully reaches a recipient's mailbox. The MDN can be used to notify the sender of any of several conditions that may occur after successful delivery such as display, printing or deletion of the message. Allow mail user agents to keep track of the message (only) in its subsequent step of the flow. The sending of the response depends on the functionality of recipient email client and often on the decision of the recipient.

Message tracking is also possible through email tracking services like ReadNotify [15] or WhoReadMe [16]. These services add to the message some hidden information: picture, or pieces of HyperText Markup Language (HTML) code (like IFRAMEs). Tracking is hidden from the recipient and not too elegant.

There is yet another reason for which the message may not reach the mail user agent - the human action. The recipient may accidentally or intentionally delete the message from his/her mailbox, move it to a different folder or mark it as a spam. Also, his/her email client or a local system may fail.

B. Document-flow patterns execution

In mailing systems, notification mechanisms can provide the status of messages in their the next step of flow, but never any further. It can be said that the email message can store history of its own flow, inform about its next step, but does not ”know” its future flow. The MIND document has an embedded workflow path, thus it has information about whole its flow and about flow of other documents in the process. However, a worker which finished his activity has no control on further flow of document – this knowledge is built in document, which has left his device.

In some cases, the location of the document may be required for the proper execution of the workflow process, especially in unexpected situations, like a lost document. LWE temporarily stores copy of documents in the worker's mailbox, in case the process has to be recreated from a certain place. Searching for a document in all places indicated by a workflow is possible but often time-consuming and costly, and may not take into account the modification of the path during the process execution.

This paper proposes a ”ground control” external service for receiving and storing notifications from LWEs about status of documents. Each notification from LWEs contains information about: process id, document id, current activity id, and sender of the notification. This section presents what other information about the document should be included in the notifications for reliable coordination of all document-flow patterns.

Based on the work of van der Aaalst [5] and the result of previous research [4], three categories of document-flow patters have been identified: distributed state patterns, coupled state patterns, and embedded state patterns.

1) Distributed state patterns: These patterns describe situations in which the next activity or activities can be determined solely on the state of the current activity. Four patterns of this type have been distinguished: sequencer, splitter, merger and iterator.

a) Document sequencer: This pattern involves a knowledge worker sending a document to another worker. The document may be sent in its entirety in one message or it may be partitioned into several messages. In this basic situation, the following problems may occur: a sender may receive bounce notifications from each sent message and recipient may not receive all messages. However, a bounce notification does not always mean that the recipient has not received the message in a timely manner. In this pattern, the notification should contain one of the three route-status: sent (sends from sender's LWE after sending the document), received (sends from recipient's LWE - after receiving the document) or bounced (sends from sender's LWE - after receiving the bounce notification). It is possible, that some notification does not reach to the ”ground control” service or arrives in the wrong order. Thus, in all patterns, the *received* status and the subsequent *sent* status are considered to more important then previous *bounced* and *received*.

b) Document splitter: This pattern creates identical copies of the document or partitions it into separate fragments. The resulting documents are next sent to the respective knowledge workers specified in the migration policy. These documents, either copies or fragments, get new document IDs. The parent document is considered to be delivered if all its child documents have been delivered. Thus, the *sent* route-status is given to each parent and child documents. The parent document has also assigned a *splitted* document-status and references to the child documents are indicated. Each arrived child document gets the *received* status individually. Once all the child documents have the *received* status (or the subsequent *sent* status), the ”ground control” service gives automatically the *received* status to the parent document. The child documents are determined by the references. The *bounced* status is also assigned to each child document separately.

c) Document merger: This pattern complementing the document splitter pattern merges all received documents in one. Of course, this pattern may involve various document functionality, depending on whether the preceding splitter has been cloning or decomposing. But before merging, all the expected documents must be delivered. The LWE client on the basis of path component of the first received document determines the number of expected documents that have to be merged. Each of the arrived document gets the *received* status. When all documents are collected, they are merged and a new document gets the *received* route-status and constitutes documents get the *merged* document-status and reference to this new merged document. The document merger fails when at least one child document has been not received. In exceptional situations, decision about completing merger before receiving all child document may be made.

d) Document iterator: This pattern enables repeated execution of some sequence of activities controlled by a condition specified in the respective document migration policy. The route-status is assigned as in document sequencer, but the activities can be performed several times and notification may be received by ”ground control” service in incorrect order. Thus, the activity id and route-status is not enough to determine where the document resides. To solve this problem, some basic partial ordering mechanism, like Lamport's timestamps [17], has been used. The path component of the document has a timestamp attribute that is incremented by LWE. When the documents are merged, the new one gets a maximum value of all merged documents' timestamps plus 1. Thus, each notification contains also a timestamp value.

2) Coupled state patterns: Sometimes completion of an activity performed by one worker may require a notification on a state of some activity performed by another worker somewhere in the organization. That involves the notion of asynchronous signals, sent between different parts of the workflow process. Three document-flow patterns of this kind have been distinguished: deferred choice, milestone and cancel activity.

a) Deferred choice and milestone: These patterns are used to deal with situations when the current activity of one worker has to be blocked until a signal notifying on some external event has been received from another worker. Both patterns require a proactive document to provide a worker's device with a semaphore and embedded functionality to handle it. Initial value of the semaphore is closed, so if the signal from another worker has not been received, the current activity is blocked. Upon receiving a signal, the waiting activity is resumed. Deferred choice is used when sending a given document has to be postponed until the worker gets information to whom it should be sent. Milestone just blocks some activity of one worker by another. The problem appears, if the signal does not arrive within the specified time and the received document activity can not be proceed. In this case, the route-status of the document is *received* but the signal-status is *waiting*.

b) Cancelling pattern: Implementation of this pattern depends on what exactly should be cancelled. If a particular activity should be cancelled, a cancellation signal is sent only to the LWE client responsible for its performance. The decision on canceling the activity is immediate for the receiving device or does not make sense any more if the document has been sent to another worker.

More problematic situation is to cancel the document, because it requires the designation of its location. It is possible to search for a document in all places indicated by the workflow, but the ”ground control” service can significantly reduce this set of places. If the route-status of document is *received*, the cancellation signal is sent only to the sender of that notification. After a successful cancellation, LWE sends the *cancelled* route-status.

If instead the *cancelled* route-status, the ”ground control” service receives the *sent* status, it can start chasing the document. This situation is shown in Figure 5. To increase the chance of success, a cancellation signal is sent to, say, three subsequent activities for each possible path of the document flow. The three cases are possible for each activity: an activity was finished, an activity is currently being preformed or waiting for a document. Figure 5a) shows successful cancellation, i.e., the ”ground control” service received a *cancelled* notification from all possible paths of the document. Figure 5b) shows cancellation potentially successful but not yet completed. While Figure 5c) shows the failed cancellation - the cancelling process should be continued for the subsequent activities on this particular path.

It is worth mentioning that the rate of the document flow is measured in minutes or hours, even days, rather than seconds. For example, the Intel's Email Service Level Agreement defines the acceptable time frame for replying to emails to 24 hours [18]. Thus, chasing the document will not be so much demanding as it might appear to be.

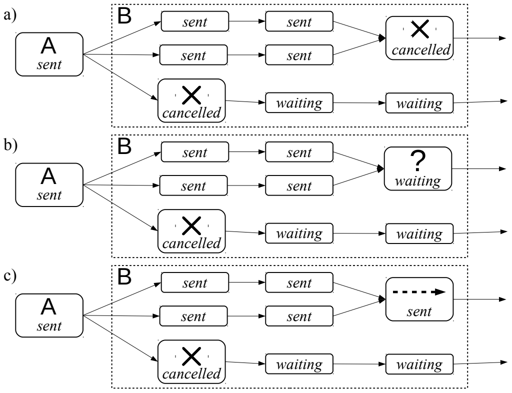


Figure 5. Cancellation of the document

3) Embedded state patterns: Performing an activity by some worker may require a subprocess delegated to someone else, with activities not specified originally in the migration policy of the arriving document. States of such a subprocess are embedded in the state of the current activity enabling that.

a) Internal subprocess: If the current worker is authorized to extend the original migration policy of a document with new activities, they constitute an internal subprocess. Neither the structure of the internal subflow nor identity of added workers have to be known earlier to the workflow designer. The notification from the subflow activities are the same like from other activities, but the ”ground control” service has only the structure of the designed workflow. Thus, for reliable coordination of subflow, its structure and identity of added workers must be sent to the ”ground control” service. If the ”ground control” does not have the current data of the subprocesses, tracing a document, and in particular, the cancellation may not be possible.

b) External subprocess: The performed activity may call some external subprocess, which structure are unknown for both, workflow designer and the performer of the current activity. The external subprocess is often performed outside of the organization, thus, it is not traced by the ”ground control” service. Only the lack of received notification at the end of the subprocess within the specified time may indicate troubles.

The document-flow patterns analysis allowed for formulating the syntax of notifications. The route-status type should be one of the: sent, received, bounced, cancelled and finished, the optional document-status can be one of the: splitted or merged. The signal-status can be waiting or just indefinite. Each notification contains also timestamp value. LWE performed a first activity adds to the notification a migration path and information about workers. If it notices a modification of the path component by adding a subprocess, it also sends definition of subprocess and information about new workers to provide the most recent data of the process. Table I summarizes the problems associated with the reliable execution of presented patterns.

V. Ground control service

This is an external service intended for a central document tracing to ensure the reliability of distributed workflow

TABLE I. RELIABILITY OF PATTERNS EXECUTION

PATTERN	PROBLEMS TO SOLVE
Sequencer	Check whether the document has reached the recipient. Route-status: sent, received, bounced.
Splitter	Check whether all constituents of the splitted document have reached the recipients. Route-status: sent, received, bounced. Document-status: splitted (for splitted document + references to constituents).
Merger	Check whether all documents that should be merged into one have reached the recipient. Route-status: sent, received, bounced. Document-status: merged (for merged documents + references to new document).
Iterator	Check whether the document has reached the recipient as many times as it has been established in the loop. Route-status: sent, received, bounced. Timestamp to determine the order of the activities.
Deffered choice Milestone	Check whether both the document and the signal have reached the recipient. Route-status: sent, received, bounced. Signal-status: waiting (or indefinite).
Cancelling	Check whether the document has been cancelled. Route-status: cancelled (when it succeeded).
Internal subprocess	Track a subprocess added during the workflow process execution. Attach subprocess sources to the notification.
External subprocess	This pattern is not tracked by the ”ground control” service.

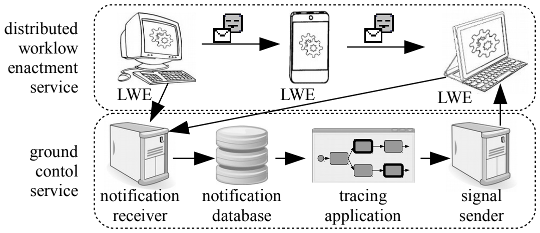


Figure 6. The *Ground control* service architecture

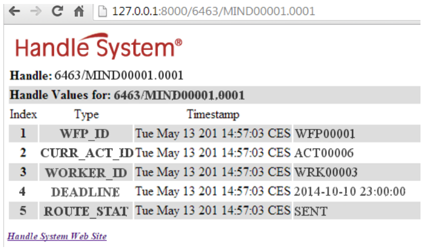
execution. The workflow enactment remains distributed and may be still performed without it, however. The intention of the ”ground control” service is to collect notifications from LWEs in order to determine the approximate global state of the distributed document flow and to make top-down decisions in some unforeseen cases. The document policy component decides whether the notification has to be sent or not.

Figure 6 presents the concept of this service. It constitutes a notification receiver, i.e., a service receiving notifications from the LWEs via the particular transport layer. Then, the notifications are parsed and placed in the database. The notification database has some functionality, e.g., trigger that gives automatically the *received* status to the splitted document, after all its child documents have also got this status.

A tracing application visualizes the workflow process and marks the currently executing activities, designated on the basis of the notifications. It is also an interface for some users allows for monitoring the process and/or makes some top-down decisions. Some decisions may require sending the notification signal to the particular LWE. Signals are generated by the tracing application and transfered by the signal sender service.

A. Implementation

Two possible implementations of transferring and storing notifications were taken into account. The former uses the Handle System, while the latter uses an email-based notification system.



127.0.0.1:8000/6463/MIND00001.0001

Handle System®

Handle: 6463/MIND00001.0001

Handle Values for: 6463/MIND00001.0001

Index	Type	Timestamp	
1	WFP_ID	Tue May 13 201 14:57:03 CES	WFP00001
2	CURR_ACT_ID	Tue May 13 201 14:57:03 CES	ACT00006
3	WORKER_ID	Tue May 13 201 14:57:03 CES	WRK00003
4	DEADLINE	Tue May 13 201 14:57:03 CES	2014-10-10 23:00:00
5	ROUTE_STAT	Tue May 13 201 14:57:03 CES	SENT

Handle System Web Site

Figure 7. A handle for the MIND document

1) Handle System [6]: is a solution for assigning, managing, storing and resolving persistent identifiers for digital objects on the Internet. It includes a set of protocols enabling a distributed computer system to store identifiers of digital resources and resolve those identifiers into the information necessary to locate and access the resources. This information can be changed to reflect the current state of the identified resource without changing the identifier. The most popular system based on Handle System is DOI (Digital Object Identifier) [19] used for persistent citations in scholarly materials, research datasets or European Union official publications.

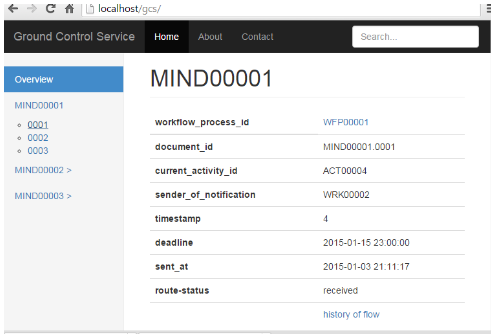
The Handle System defines a hierarchical service model. The top level consists of a single handle service, known as the Global Handle Registry. The lower level consists of all other handle services, known as Local Handle Services. The Handle System provides the Java-based Handle Server and a set of tools needed for the Local Handle Service installation. The Global Handle Registry is used to manage any handle namespace and provides the service used to manage naming authorities. The Local Handle Service and its responsible set of local handle namespaces must be registered with the Global Handle Registry and gets a unique prefix.

The Handle System provides unique persistent identifiers called handles for digital objects, such as the MIND document. The handle is a character string that consists of two parts: its naming authority and a local name separated by the ASCII (American Standard Code for Information Interchange) character ”/”. Each handle may have a set of values assigned to it. A handle value may be thought as a record that consists of a group of data fields. Every handle value must have a data type. The Handle System predefines a set of data types and allows for defining another.

Thus, the ”ground control” service can use the Handle System to create an unique handle for each migrating document (see Figure 7). Each handle has a set of values corresponding to the syntax of the LWE notifications. The LWE modifies it at each change of document status.

Nevertheless, this solution has some disadvantages. Modifications of handles occur frequently, and each time they require a connection with Global Handle Registry. Besides, the Local Handle Service administration requires additional skills and needs control of other than email transport layer. The Handle System indicates the current location of the document. However, extraction of the list of all the documents in given process requires additional functionality. The Handle System tracks each document separately.

2) Email-based notification system: The ”ground control” service has been also implemented as the email-based notifica-



MIND00001

workflow_process_id	WFP00001
document_id	MIND00001.0001
current_activity_id	ACT00004
sender_of_notification	WRK00002
timestamp	4
deadline	2015-01-15 23:00:00
sent_at	2015-01-03 21:11:17
route-status	received
	history of flow

Figure 8. GUI of the email based ”ground control” service

tion system on basis of email transport layer. Email is intended for frequent passing of messages so that it can easily receive multiple notifications and does not require any additional users skills in the installation, configuration and operation. It does not require unlocking new ports for the transport layer, which affects the security of the organization.

The LWE notifications are sent to one or more email addresses. The notification receiver services run on some organizational server check dedicated mailboxes frequently, parse attached LWE notifications and insert new records to a database.

There are three main tables in the database: Notifications, Documents, and WorkflowProcesses. Each new notification is inserted into the Notifications table. The new notification is distinguished by the address of the mailbox and email's Unique Identifier (UID - a unique number referencing an email in a mailbox). Only those notifications are inserted to the Documents table, which have higher value of logical timestamp than the already registered. The last record for each document ID refers to the current state of this document.

When a notification for a new process appears or process was modified, a new record is inserted to the WorkflowProcesses table. This table stores workflow process IDs, the migration path files and workers data files.

Thus, the WorkflowProcesses table stores structure of the process, while the Documents table stores the states of documents flow. The tracing application selects only the most recent records from Documents and WorkflowProcesses tables and constructs a current workflow process structure with its approximate global state.

In contrast to the LWE, which has been implemented in Java, the ”ground control” service has been implemented based on PHP (Personal Home Page) and Postgresql database [20]. PHP technology has been chosen in order to test it for email messaging and XML (Extensible Markup Language) manipulating. PHP provides classes to access mailbox, e.g., by IMAP protocol and functions to XML manipulation. PiBX (XML-Data-Binding framework for PHP) is similar to JAXB (Java Architecture for XML Binding), but it is in the alpha-state at the moment. PHP technology has been good enough for the rapid implementation of the ”ground control” service, but many other technologies could be used for this purpose. In fact, the syntax of notifications is essential for the ”ground control” service, since implementation does not require any new or advanced technology.

Figure 8 shows information about one MIND document selected from the database. An interface allows the user to view all documents related to the process and to view a history of document flow.

During the experiments, emails were received from the dedicated mailbox every minute (for this purpose, the ”Cron” software was used). So emails often appear in mailbox in the different order than they were sent. First sent notification provides a workflow process resources (process definitions, data about workers) to the ”ground control” service. However, sometimes this notification is received later then subsequent notifications. In such a situation, only worflow process ID is inserted to the WorkflowProcesses table and the table is updated at a later time.

Emails with notifications generally were delivered to the inbox without any problems. However, the service does not require to deliver all notifications. However, there was a problem during testing that emails have been received from mailbox and deleted, but the service crashed while writing data to the database. To prevent such situations, emails are stored in the inbox for a month.

VI. Related work

The presented proposal combines existing technologies and new idea to extract some new functionality in the topics of the distributed electronic document and collaborative environments.

The first significant step in the document-based processing was the Multivalent Document architecture MVD [21] that introduced active functionality to manipulate a document content with dynamically loaded objects called *behaviors*. The concept of behaviors is similar to the MIND embedded services, howerer MIND expands this concept with local and external services, which can also affect a document behavior, but are not components of the document. This gives documents more flexibility on opening, suiting them better to exploit local resources of visiting devices and to easily add a new functionality.

The Placeless Documents [22] implements document functionality with active properties that cannot only manipulate a document content but also manage of a document structure and workflow. The Placeless Documents are reactive, i.e., they respond to external events, while MIND documents are proactive – they initiate their own behavior.

The concept of a proactive document, capable of traveling between computers under its own control has been introduces with a document-agent platform MobiDoc [23]. This platform was, however, closely related to the particular technology, and thus lacked forward compatibility. On the other hand, solutions proposed by MIND found document-agent mobility on stable email messaging standards. Owing to proactive MIND attachments, any email system could be almost like an agent platform with all the benefits of multi-agent systems, but without a need to implement a full-size agent platform that would have to be updated regularly and require additional skills from administrators to run it.

Workflows have been also implemented by WADE (Workflow and Agents Development Environment) [24] agent platform based on JADE (Java Agent DEvelopment framework) [25]. WADE agents embed a micro-workflow engine, capable of executing workflows and compiled before launching the workflow. Performing of activities may be delegated by one agent to another and in principle is not related to agent mobility. This solution follows the classic central workflow enactment philosophy, and differs from it only in decentralization of a global process state into local process states controlled by micro-workflow engines running inside agents. In the MIND architecture, workflow as a XML Process Definition Language (XPDL) file is a part of the whole document and it contains its internal state. LWEs run outside of agents as local workflow engines. Workflow, in the form of plain XPDL, may be also modified during the process execution. Moreover, a document-agent is the only communication interface, making MIND based platforms technologically independent and truly loosely coupled distributed systems.

The reliability of distributed workflows processing is associated with the assurance that the object would not be lost and would arrive to the designated location. It requires some tracking service that in distributed loosely-coupled systems may only estimate the real states of migrating objects. The JADE platform provides some control remote agents (Agent Management System – AMS, Remote Monitoring Agent – RMA) that receives messages from JADE agents, while the "ground control" service has a similar task - it receives messages from the MIND documents to tracking their states. Contrary to JADE control agents, the "ground control" is an external, technologically independent service that communicates with the MIND documents through notifications. Document determines whether the notification has to be sent or not. A syntax of the notifications includes also all document-flow patterns.

VII. Conclusion

Reliable workflow execution of distributed mobile document must be able to handle unforeseen situation when migrating documents fail to reach their destination or get stuck in some worker's device. The "ground control" service, proposed in this paper, is a track and trace service that enables observation of the current document location and stores the history of migration. It allows for checking if the document has reached the recipient LWE or is processed too long on the current device, i.e., if passed the appointed deadline and *sent* notification has not received, it may indicate that the document got stuck in some place. The service does not "tighten" the idea of loosely-coupled distributed system, because it can still execute without this service and the MIND document may decide in which steps the notifications should be sent and in which should not.

Next to reliability, there is also a security issue, which answers the question: what to do if lost document will get to an unauthorized person? The LWE may require authentication of the worker before unpacking and activating document components. The LWE also verifies if the performer assigned to the current activity is the same person as the recipient of the document. The interesting idea has been proposed in [26] by the MENAID (Methods and Tools for Next Generation Document Engineering) project [27]. It introduces a security by the face recognition algorithm built in the MIND documents.

Acknowledgment

This work was supported in part by the National Science Center, Poland, under grant DEC1-2011/01/B/ST6/06500.

References

[1] G. D. Bhatt, "Organizing knowledge in the knowledge development cycle," Journal of Knowledge Management, vol. 4, 2000, pp. 15–26.

[2] L. A. Dabbish and R. E. Kraut, "Email overload at work: an analysis of factors associated with email strain," in Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work, ser. CSCW'06. New York, USA: ACM, 2006, pp. 431–440.

[3] M. Godlewska, "Agent system for managing distributed mobile interactive documents in knowledge-based organizations," in Transactions on Computational Collective Intelligence VI, ser. LNCS 7190, N. T. Nguyen, Ed. Berlin: Springer-Verlag, 2012, pp. 121–145.

[4] M. Godlewska and B. Wiszniewski, "Smart email - almost an agent platform," in Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering, ser. LNEE, S. Tarek and E. Khaled, Eds. Berlin: Springer-Verlag, 2015, pp. 581–589.

[5] N. Russell, A. Hofstede, W. Aalst, and N. Mulyar, "Workflow control-flow patterns: A revised view," 2006, BPM Center Report BPM-06-22.

[6] Corporation for National Research Initiatives, "Handle.net (version 7.0): Technical manual," 2010.

[7] WfMC. Workflow Management Coalition, "Terminology and glossary," WfMC, Winchester, UK, Tech. Rep. WFMC-TC-1011, Issue 3.0, 1999.

[8] J. Klensin, "Simple Mail Transfer Protocol," RFC 5321, IETF, 2008.

[9] K. Moore, "Simple Mail Transfer Protocol (SMTP) Service Extension for Delivery Status Notifications (DSNs)," RFC 3461, 2003.

[10] C. Lewis, "Overview of Best Email DNS-Based List (DNSBL) Operational Practices," RFC 6471, 2012.

[11] "Java Mail," URL: http://www.oracle.com/ [retrieved: Dec., 2014].

[12] M. Crispin, "Internet message access protocol - version 4rev1," RFC 3501, 2003.

[13] K. Moore, "Recommendations for Automatic Responses to Electronic Mail," RFC 3834, IETF, 2004.

[14] T. Hansen and G. Vaudreuil, "Message Disposition Notification," RFC 3798, IETF, 2004.

[15] "Readnotify," URL: http://www.readnotify.com [retrieved: Dec., 2014].

[16] "Whoreadme," URL: http://www.whoreadme.com [retrieved: Dec., 2014].

[17] G. Coulouris, J. Dollimore, T. Kindberg, and G. Blair, Distributed Systems: Concepts and Design, 5th ed. USA: Addison-Wesley Publishing Company, 2011.

[18] J. Spira and C. Burke, "Intel's war on information overload: A case study," 2009.

[19] International DOI Foundation, "DOI Handbook," 2013.

[20] The PostgreSQL Global Development Group, "PostgreSQL 9.4.0 Documentation," 2014.

[21] T. A. Phelps and R. Wilensky, "Multivalent documents: A new model for digital documents," EECS Department, University of California, Berkeley, Tech. Rep. UCB/CSD-98-999, 1998.

[22] P. Dourish, W. K. Edwards, A. LaMarca, J. Lamping, K. Petersen, M. Salisbury, D. B. Terry, and J. Thornton, "Extending document management systems with user-specific active properties," ACM Trans. Inf. Syst., vol. 18, no. 2, 2000, pp. 140–170.

[23] I. Satoh, "Mobile agent-based compound documents," in Proceedings of the 2001 ACM Symposium on Document engineering, ser. DocEng '01. New York, USA: ACM, 2001, pp. 76–84.

[24] Telecom Italia, "Workflows and Agents Development Environment," 2014, URL: http://jade.tilab.com/wade [retrieved: Dec., 2014].

[25] Telecom Italia, "Java Agent Development Framework," 2014, URL: http://jade.tilab.com [retrieved: Dec., 2014].

[26] J. Siciarek, M. Smiatacz, and B. Wiszniewski, "For your eyes only – biometric protection of pdf documents," in EEE'13 - The 2013 International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government, Las Vegas, USA, 2013, pp. 212–217.

[27] MeNaID, "http://menaid.org.pl/," 2012-2014, National Science Center, Poland, grant DEC1-2011/01/B/ST6/06500.

Spectrum Allocation Policies in Fragmentation Aware and Balanced Load Routing for Elastic Optical Networks

André C. S. Donza, Carlos R. L. Francês
High Performance Networks Processing Lab - LPRAD
Universidade Federal do Pará (UFPA)
Belém, Brazil
{acdonza, renato.frances2010}@gmail.com

João C. W. A. Costa
Applied Electromagnetics Lab - LEA
Universidade Federal do Pará (UFPA)
Belém, Brazil
jweyl@ufpa.br

***Abstract*— The rigid nature of wavelength division multiplexing (WDM) routed networks leads to inefficient capacity utilization. Thus, flexible networks are a possible breakthrough for Internet technology, as long as they provide higher spectrum efficiency use. Several discrete-time simulations were carried out in Matlab in order to analyze different spectrum allocation policies (First-Fit, Exact-Fit and Random-Fit) in some routing algorithms: The Fragmentation Aware Assignment (FA), the Shortest Path with Maximum Spectrum Reuse (SPSR) and the Balanced Load Score Spectrum Assignment (BLSA). Two network topologies were used: a small 6-node subset of Cost239 and a 7-node random topology. As physical layer effects were not included as constraints, Fragmentation Aware and Balanced Load Spectrum Assignment strongly outperformed Shortest Path with Maximum Spectrum reuse, with much better results for BLSA. The separation between First-Fit and Exact-Fit curves was smaller in SPSR than in FA and BLSA. In general, Exact-Fit spectrum allocation policy presented slightly better performance than First-Fit.**

Keywords-Routing; spectrum allocation; elastic optical networks.

I. INTRODUCTION

The rigid nature of wavelength division multiplexing (WDM) routed networks leads to inefficient spectrum utilization, a problem that is expected to become much more critical with the deployment of higher capacity WDM networks. Thus, flexible networks are required to provide high spectrum efficiency use in order to achieve scalability, reduce network power consumption and decrease per unit bandwidth cost.

Optical Orthogonal Frequency-Division Multiplexing (O-OFDM) is one of the promising modulation techniques for optical networks. Optical OFDM distributes the data on several low data rate subcarriers. The spectrum of adjacent subcarriers can overlap, since they are orthogonally modulated [1], providing good spectral efficiency, flexibility and tolerance to impairments.

OFDM-based elastic optical networks achieve multiple data rate sub-wavelength or super-wavelength paths through flexible granular grooming and switching in the spectrum domain, using data-rate/bandwidth-variable transponders and bandwidth-variable optical cross-connects [2].

The data-rate/bandwidth-variable transponder provides no more than the enough subcarriers to treat sub-wavelength traffic. It is also possible to create super-wavelength paths to transport multiple-rate data traffic, by merging several OFDM channels.

Transmitted signals are routed over the optical path through bandwidth variable optical cross-connects, designed to allocate a cross-connection with the suitable spectrum to create an appropriate-sized end-to-end optical path.

The approach of the following sections is directed to routing and spectrum allocation algorithms, as long as traditional routing and wavelength assignment algorithms can no longer be directly applied to establish an elastic optical path that uses flexible spectrum width to accommodate multi-data rate services.

Section II defines the routing and spectrum allocation problem in flexible optical networks. In Section III, First-Fit, Exact-Fit and Random-Fit spectrum allocation policies are described in details. Section IV presents the results of the deployed simulations. Finally, Section V presents the conclusion and the perspectives of this work.

II. ROUTING AND SPECTRUM ALLOCATION

There is an increasing number of works searching for solutions to the elastic optical network Routing and Spectrum Allocation (RSA) problem. The RSA treats routing and spectrum allocation in order to save spectral resources, in an optimized optical network operation.

The capacity requirement of each connection request is characterized by a number of subcarrier slots, based on the capacity of the subcarriers.

Integer Linear Programming formulations are not scalable to large networks and many heuristic algorithms have been developed to treat connection requests sequentially.

The RSA problem may have an one-step approach or a two step-approach. In one-step approach, as in Modified Dijkstra's Shortest Path (MSP) and in Spectrum-Constraint Path Vector Searching (SCPVS), the algorithms define the route and the available contiguous spectrum simultaneously. On the other hand, in a two-step approach, routing and spectrum assignment are sub-problems that are solved sequentially. In the following subtopics, the Fragmentation Aware Assignment, the Shortest Path with Maximum Spectrum Reuse and the Balanced Load Score Assignment algorithms are described. After routing, spectrum allocation is performed using one of the possible existing policies.

A. Constraints

There are several restrictions to be observed in RSA problems: The traffic demands for a node-pair should be exactly added in the source node and dropped at the destination point; one sub-carrier in a fiber can only be used to serve one spectrum path; each optical path should use the same subcarriers along its entire way; overlapping spectrum paths must be separated by a number of subcarriers called guard-carriers and, finally, the employed subcarriers in a spectrum path must be consecutive in the frequency domain.

B. Shortest Path with Maximum Spectrum Reuse

For a given set of spectrum path request pairs, the sub-carrier reuse can be increased by reducing the maximum sub-carrier number. Shortest path with maximum spectrum reuse (SPSR) algorithm, proposed in [3], combines the shortest path routing with the maximum reuse spectrum allocation (MRSA) algorithm, where simultaneous spectrum path requests are first sorted according to the size of the traffic demand and larger traffic demands have higher allocation priority.

C. Balanced Load Score Spectrum Allocation (BLSA)

Proposed in [3], BLSA determines the routing through a load balancing, in order to decrease the maximum subcarrier number on a fiber. In the beginning, a k-shortest path algorithm is used to generate k paths for each source-destination pair. Then, a path is selected by estimating the load of the fibers. The maximum fiber load (MFL) of each path is taken and the better path is the one that presents the smaller MFL. Finally, after the path selection, the spectrum allocation is made through one of the possible allocation strategies (First-Fit, Exact-Fit or Random-Fit, for example). In the situation of simultaneous requests, larger traffic demands have higher allocation priority over smaller ones.

The load of a fiber (FL) can be estimated using the sum of the size of all requests, i.e., the total amount of occupied slots (SUM) in the considered link, the quantity of guard carriers (GC) and the number of optical paths (I) at that fiber, according to (1).

$$\text{FL} = \text{SUM} + \text{GC} \times (\text{I}-1) \qquad (1)$$

D. Fragmentation Aware Routing (FA)

Fragmentation is not directly related to the spectrum utilization but it can be used as a decisive parameter for routing. The process of adding and ending connections in a non-uniform bandwidth assignment generates an interleaved spectrum. Obviously, spectrum fragmentation becomes a problem when free resources are broken into portions that are smaller than incoming bandwidth requests. Hence, these small non-contiguous free frequency bands reduce the spectrum efficiency gained by flexibility in the bandwidth allocation.

In the developed FA algorithm, the external fragmentation formulation, presented in [4], is used to select one of the k paths generated by the k-shortest path algorithm for each incoming source-destination connection request. Thus, a path is selected by the fragmentation estimation of each fiber along the way. The maximum fiber fragmentation (MFF) of each path must be taken and the chosen path is the one that presents the smaller MFF. After the path selection, the spectrum allocation is made through one of the possible allocation strategies. In the situation of simultaneous requests, larger traffic demands have higher allocation priority over smaller ones.

The external fragmentation of a fiber can be calculated by (2), where the largest free block (LFB) is the number of slots of the largest contiguous free space, and total free (TF) is the total amount of free slots.

$$F_{ext} = 1 - (\text{LFB}/\text{TF}) \qquad (2)$$

III. Spectrum Assignment Policies

In elastic optical networks, the assignment of spectrum slots is in different granularities to the arriving connection requests.

A first-fit assignment policy serves the request in the first available frequency band fitting the spectrum demand, a random allocation policy places incoming requests in any available block large enough to satisfy the necessary bandwidth and an exact-fit assignment, proposed in [4], searches for the exact available block in terms of the number of slots requested for the connection. If there is no block that matches perfectly, the spectrum is allocated in the first largest free block available.

IV. Results

Some simulations were carried out in Matlab to compare different existing routing and spectrum assignment algorithms for elastic optical networks. The dynamic events were simulated in a 6-node subset of Cost239 and in a 7-node fiber based structure, as shown in Figures 1 and 2, respectively.

As long as the main purpose of the simulations is to compare different spectrum assignment policies in Fragmentation Aware Routing and in Balanced Load Spectrum Assignment, a distance-adaptive modulation level selection has not been configured for use. Actually, for each source-destination request, the algorithms simply select a route and try to allocate the incoming connection in one of the available sets of contiguous frequency slots, according to the chosen technique and observing all the constraints. For each source-destination pair, routes are chosen from a list created by Yen's k-shortest path algorithm.

Contiguous spectral paths were set to be separated by one guard carrier. Each type of connection is represented by one of the values in the set C = {4, 5, 6, 8, 10, 12, 14, 16}. The elements of C are the number of contiguous slots needed to satisfy bandwidth requirements. For each one of these incoming request types, it has been considered an arrival rate following a Poisson distribution.

At each time unit of the simulations, all the incoming request types were randomly associated to source-destination pairs. 400 spectral slots were defined for each fiber link. The service time of each connection follows exponential probabilities of Poisson.

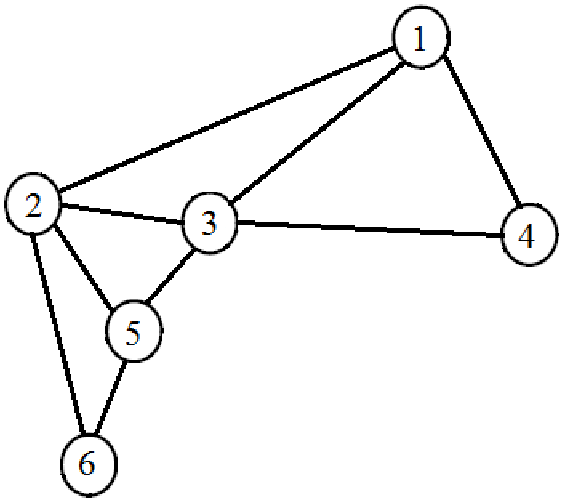


Figure 1. The 6-node subset of Cost239 network used for numerical evaluations

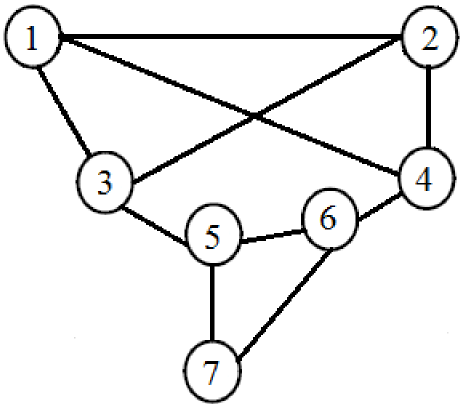


Figure 2. The 7-node random topology used for numerical evaluations

Each number associated with incoming requests load (IRL) in Tables I and II represents the blocking frequency average of five deployed simulations for the routing techniques specified in the columns. The IRL was defined as the product between the average service time, the average arrival rate and the number of connection requests. In the lines of the presented tables, this load is displayed normalized in relation to the IRL of one of the lines. Table I refers to the 6-node subset of Cost239 and Table II refers to the 7-node random topology, with First-Fit policy as spectrum allocation technique used. The last line presents the general averages of the 60 implemented simulations. Figures 3 and 4 depict the data presented in Tables I and II, respectively, in order to compare the performances of BLSA, FA and SPSR.

TABLE I. BLOCKING FREQUENCY AVERAGES FOR SPSR, BLSA AND FA IN THE 6-NODE TOPOLOGY.

Normalized IRL	*SPSR*	*BLSA*	*FA*
1	0.0751	0	0.0121
1.5	0.1214	0.0110	0.0563
2.5	0.1302	0.0155	0.0673
4	0.1501	0.0254	0.0905
General Average	*0.1192*	*0.0130*	*0.0565*

TABLE II. BLOCKING FREQUENCY AVERAGES FOR SPSR, BLSA AND FA IN THE 7-NODE TOPOLOGY.

Normalized IRL	*SPSR*	*BLSA*	*FA*
0.2	0.0413	0	0.0052
0.4	0.1731	0.0116	0.0775
1	0.1774	0.0426	0.1214
2	0.1899	0.0504	0.1331
General Average	*0.1454*	*0.0261*	*0.0843*

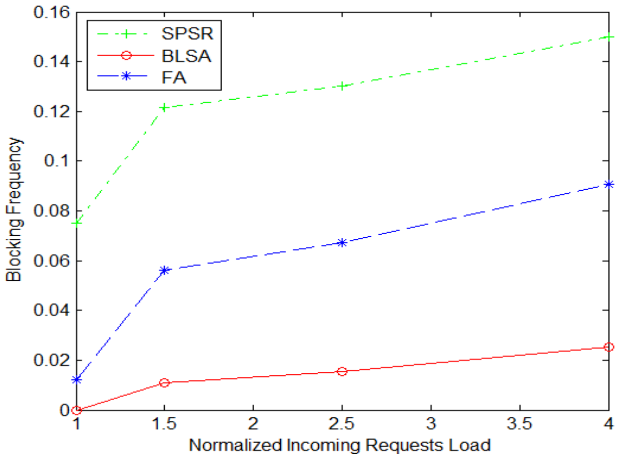


Figure 3. Blocking Frequency Comparison between SPSR, BLSA and FA for the 6-node network.

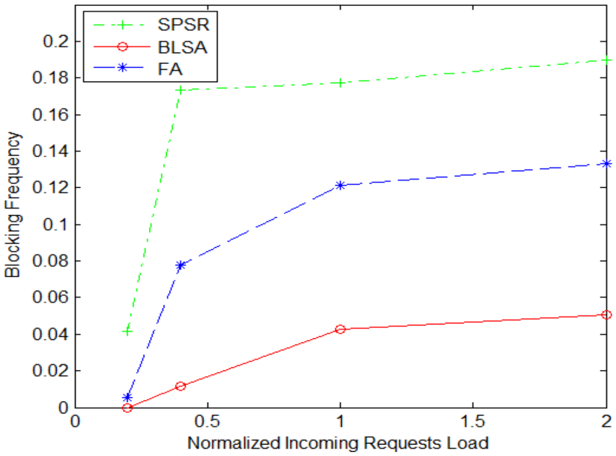


Figure 4. Blocking Frequency Comparison between SPSR, BLSA and FA for the 7-node network.

Tables III, IV, V, VI, VII and VIII present blocking frequency averages of some deployed simulations for SPSR, Balanced Load Score Spectrum Assignment and Fragmentation Aware Routing, respectively, for each topology used for numerical evaluations. Each number associated with an IRL in any of these tables is the average of 5 simulation results for the spectrum allocation policies specified in the columns, in order to compare the performances of First-Fit, Exact-Fit and Random-Fit strategies in the routing techniques proposed. The last lines are general averages.

TABLE III. BLOCKING FREQUENCY AVERAGES FOR DIFFERENT SPECTRUM ALLOCATION POLICIES IN SPSR FOR THE 6-NODE TOPOLOGY.

Normalized IRL	***SPSR***		
	First-Fit	***Exact-Fit***	***Random-Fit***
1	0.0623	0.0623	0.0609
1.6	0.1272	0.1272	0.1285
4	0.1497	0.1483	0.1510
12	0.1550	0.1550	0.1603
General Average	*0.1235*	*0.1232*	*0.1252*

TABLE IV. BLOCKING FREQUENCY AVERAGES FOR DIFFERENT SPECTRUM ALLOCATION POLICIES IN SPSR FOR THE 7-NODE TOPOLOGY.

Normalized IRL	***SPSR***		
	First-Fit	***Exact-Fit***	***Random-Fit***
0.3	0.1364	0.1395	0.1519
0.4	0.2124	0.2093	0.2078
1	0.2264	0.2264	0.2326
1.6	0.2589	0.2543	0.2543
General Average	*0.2085*	*0.2074*	*0.2116*

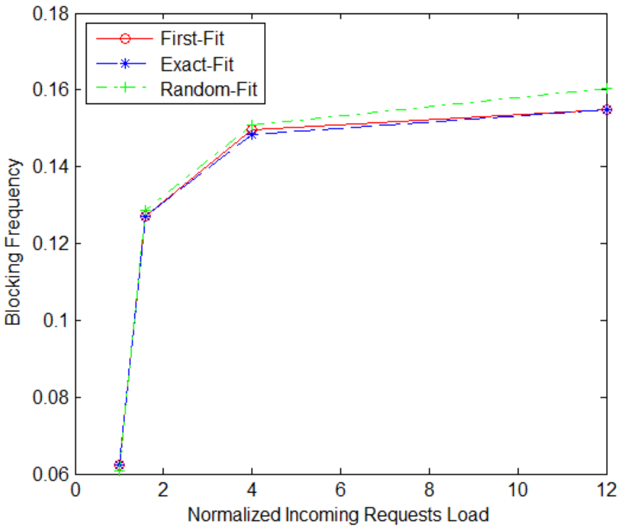


Figure 5. Blocking Frequency Performance Comparison between First-Fit, Exact-Fit and Random-Fit in SPSR for the 6-node topology.

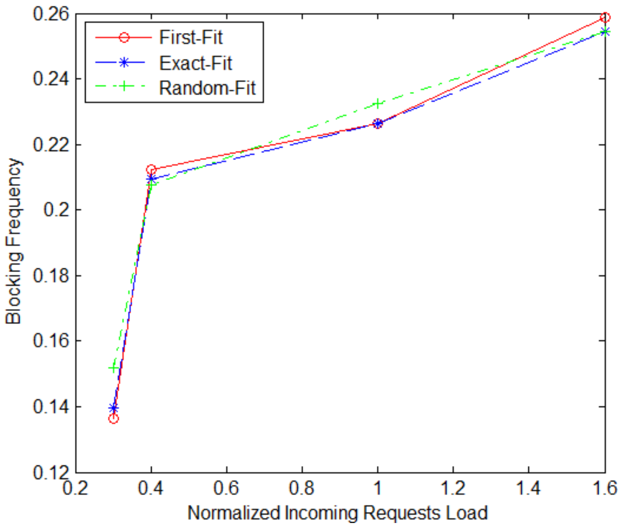


Figure 6. Blocking Frequency Performance Comparison between First-Fit, Exact-Fit and Random-Fit in SPSR for the 7-node random topology.

TABLE V. BLOCKING FREQUENCY AVERAGES FOR DIFFERENT SPECTRUM ALLOCATION POLICIES IN BLSA FOR THE 6-NODE TOPOLOGY.

Normalized IRL	***BLSA***		
	First-Fit	***Exact-Fit***	***Random-Fit***
1	*0*	*0*	*0.0053*
1.6	*0.0106*	*0.0093*	*0.0397*
4	*0.0344*	*0.0411*	*0.0689*
12	0.0517	0.0490	0.0570
General Average	*0.0242*	*0.0248*	*0.0427*

TABLE VI. BLOCKING FREQUENCY AVERAGES FOR DIFFERENT SPECTRUM ALLOCATION POLICIES IN BLSA FOR THE 7-NODE TOPOLOGY.

Normalized IRL	***BLSA***		
	First-Fit	***Exact-Fit***	***Random-Fit***
0.3	0.0109	0.0078	0.0155
0.4	0.0217	0.0171	0.0341
1	0.0372	0.0372	0.0450
1.6	0.0434	0.0419	0.0620
General Average	*0.0283*	*0.0260*	*0.0391*

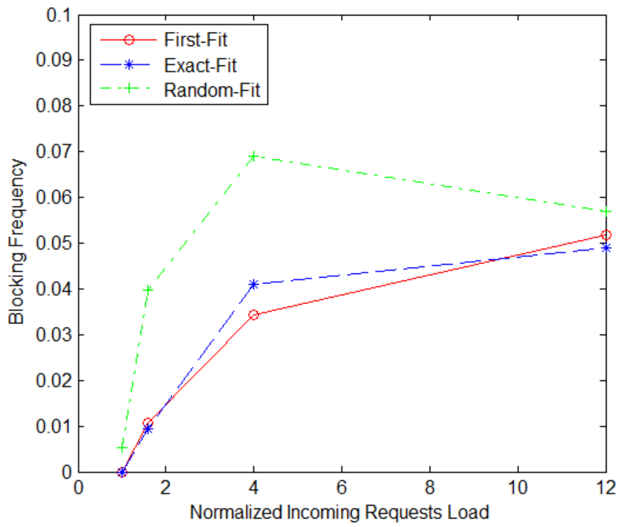


Figure 7. Blocking Frequency Performance Comparison between First-Fit, Exact-Fit and Random-Fit in BLSA for the 6-node subset of Cost239.

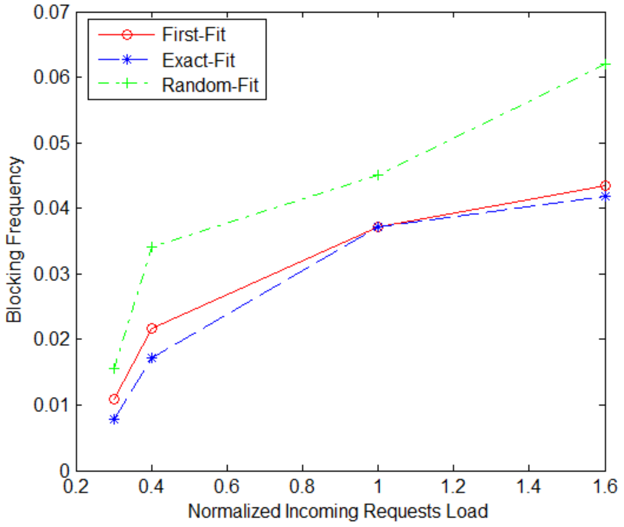


Figure 8. Blocking Frequency Performance Comparison between First-Fit, Exact-Fit and Random-Fit n BLSA for the 7-node random topology.

TABLE VII. BLOCKING FREQUENCY AVERAGES FOR DIFFERENT SPECTRUM ALLOCATION POLICIES IN FRAGMENTATION AWARE ROUTING FOR THE 6-NODE TOPOLOGY.

Normalized IRL	*FA*		
	First- Fit	*Exact-Fit*	*Random-Fit*
1	0.0159	0.0265	0.0556
1.6	0.0583	0.0623	0.1099
4	0.0954	0.0848	0.1285
12	0.1060	0.1113	0.1192
General Average	*0.0689*	*0.0712*	*0.1033*

TABLE VIII. BLOCKING FREQUENCY AVERAGES FOR DIFFERENT SPECTRUM ALLOCATION POLICIES IN FRAGMENTATION AWARE ROUTING FOR THE 7-NODE TOPOLOGY.

Normalized IRL	*FA*		
	First- Fit	*Exact-Fit*	*Random-Fit*
0.3	0	0.0465	0.0667
0.4	0.0078	0.0992	0.1256
1	0.0248	0.1442	0.1442
1.6	0.0667	0.1349	0.1643
General Average	*0.0248*	*0.1062*	*0.1252*

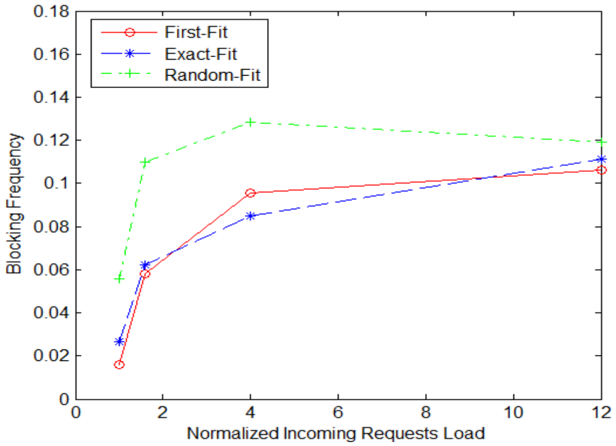


Figure 9. Blocking Frequency Performance Comparison between First-Fit, Exact-Fit and Random-Fit in FA for the 6-node subset of Cost239

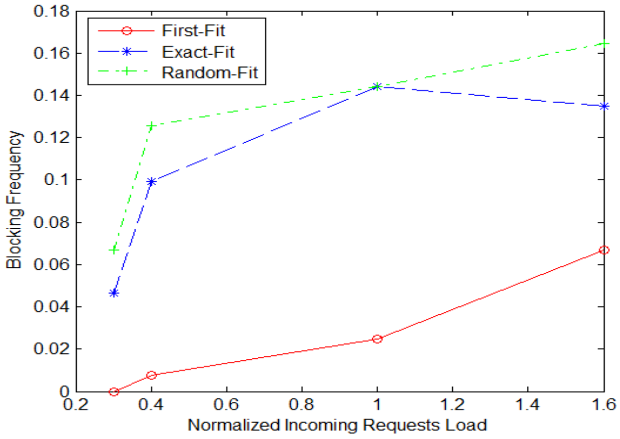


Figure 10. Blocking Frequency Performance Comparison between First-Fit, Exact-Fit and Random-Fit spectrum allocation policies in FA for the 7-node random topology.

Figures 5, 7 and 9 illustrate the data presented in Tables III, V and VII, respectively, in order to compare the performances of First-Fit, Exact-Fit and Random-Fit spectrum allocation policies in the 6-node subset of Cost239 deployments. Figures 6, 8 and 10 illustrate the data presented in Tables IV, VI and VIII, respectively, in order to compare the performances of First-Fit, Exact-Fit and Random-Fit spectrum allocation policies in the 7-node random topology proposed.

V. Conclusions

As non-linear physical layer effects were not included in the deployed network simulations, Balanced Load Score Spectrum Assignment and Fragmentation Aware Routing Techniques strongly outperformed Shortest Path Routing, with much better results for BLSA. However, it is essential the continuous searching for better formulations and strategies over the problem of fragmentation to optimize the performance of algorithms that work with it as a decisive parameter for routing. In the developed algorithms, only the concept of external fragmentation was explored.

The separation between First-Fit and Exact-Fit curves was smaller in SPSR than in BLSA and FA. In general, Exact-Fit spectrum allocation policy presented slightly better performance than First-Fit. However, as it did not happen in all the cases, a much larger number of simulations in other topologies is necessary to reach a strong conclusion on this matter.

Thus, future works might include studies on correlation between fragmentation and blocking probability, as long as other routing and spectrum allocation algorithms.

References

[1] K. Christodoulopoulos, I. Tomkos, and E.A. Varvarigos, “Elastic Bandwidth Allocation in Flexible OFDM-Based Optical Networks,” Journal of Lightwave Technology, vol. 29, May 1, 2011, pp. 1354-1366.

[2] G. Zhang, M.Leenheer, A.Morea, and B. Mukherjee, "A Survey on OFDM-Based Elastic Core Optical Networking," IEEE Communications Surveys & Tutorials, vol. 15, no.1, First Quarter 2013, pp. 65-87.

[3] Y. Wang, X. Cao and Y.Pan, “A Study of the Routing and Spectrum Allocation in Spectrum-Sliced Elastic Optical Path Networks,” IEEE Infocom, 2011, pp. 1503-1511.

[4] A. Rosa, C.Cavdar, S.Carvalho, J.Costa and L.Wosinska, “Spectrum Allocation Policy Modeling for Elastic Optical Networks”, in Proc. HONET, 2012, pp. 242-246.

Bilateral Multi-Issue Negotiation Between Active Documents and Execution Devices

Jerzy Kaczorek[1], Bogdan Wiszniewski[2]

Department of Intelligent Interactive Systems
Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology
Gdansk, Poland
Email: [1]jkaczorek@gmail.com, [2]bowisz@eti.pg.gda.pl

***Abstract*—Mobile document-agents are often in conflict with execution devices when attempting to perform activities of the business process they implement, since preferences of device owners may change depending on their current location and the actual class of the device in use. The paper proposes a bilateral negotiation mechanism based on a simple bargaining game that can effectively resolve such conflicts without any third party support.**

***Keywords*–eCollaboration; Mobile computing.**

I. Introduction

Individuals, who collaborate in a network organization, interact by exchanging electronic documents that constitute units of *information* and at the same time units of *interaction*. This dichotomy has become apparent with the advent of *active documents*, often implemented as software agents. In particular, a mobile interactive document (MIND) can migrate over the network and carry both, the *content* to be worked on and specification of its migration path with *activities* and *transitions* [1]. Each activity represents a piece of work to be performed by the user with the incoming document content, whereas transition indicates where the outgoing document (or documents), constituting a result of the activity, should migrate next. This idea is outlined in Figure 1; activities are represented by rectangles and transitions by arrows.

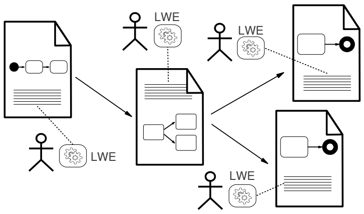


Figure 1. Mobile documents with embedded workflows

The process is started by the document originator, who sends the MIND document to the first collaborator indicated in its workflow. Each collaborator may interact with the content of the received MIND document using any currently available personal device that can receive and send email messages – from simple cellphones to smartphones or tablets to laptops or workstations. Each device has a lightweight workflow engine (LWE), implemented as an email client, which is capable of unpacking and packing the documents and sending them to other workers of the organization using email as the transport layer [1].

A. Document execution context

A single activity performed in the execution context provided by the device, on which the mobile document is currently located, depends on the policy of the document originator, operational characteristics of the currently used worker's device, and preferences of the knowledge worker responsible for the current processing step. Proactive MIND documents may handle that in several ways: activity may be performed automatically by the *embedded* document code, if allowed by the worker operating his/her device, may be performed manually by the worker using *local* services or tools installed on that device, also the device may call some *external* (third party) service requested by the document, if Internet connection is available at the time of executing the activity. This task is not trivial – as execution contexts may vary, because the same worker may use different devices when performing activities of the same business process, e.g., using a workstation when in office, a smartphone during the travel between office and home, and a laptop at home. Moreover, user preferences for the same device may depend on its current location, e.g., when out of office and accessing an untrusted network, and often conflicting with the document-agent policies. Finally, document-agents arriving to the particular device may have incomplete information on the specific execution context provided by the device.

Further in the paper, we propose *negotiation* to provide a solution to the problem of reaching an agreement between the document-agent and its execution device – even when the parties are in conflict and have incomplete information on each other preferences on how the current activity should be performed. This is the novel concept in the area of document engineering [2].

B. Negotiation model

Offers exchanged by negotiating parties are m-vectors of items $o = \langle item_1, item_2, \ldots, item_m \rangle$. Each $item_i$, where $i = 1, \ldots, m$, can be assigned a value of any attribute-specific type chosen from the set of values: $A_i = \{a_{i_1}, a_{i_2}, \ldots, a_{i_n}\}$, where $i_n = |A_i|$. Operator $|\ |$ denotes cardinality of its argument set, A_i represents the i-th attribute of the negotiated service. Set A_T of all m-vectors $A_T = \times_{i=1}^{m} A_i$ is called a *space of offers*. Based on that we define offer $o_j \in A_T$, where $j = 1, \ldots, |A_T|$, as the vector of attribute values $o_j = \langle v_1^j, v_2^j, \ldots, v_m^j \rangle$, where $v_i^j \in A_i$. Each single attribute value has an assigned numerical value, which reflects utility of the related attribute value. Utility of attribute value $a_{i_k} \in A_i$, for $i_k = 1, \ldots, |A_i|$, is calculated by function $u_i : A_i \rightarrow N$. Each party has its own set of functions $\{u_1, u_2, \ldots, u_m\}$ to calculate utility of any item in the offer. Given that, utility of each offer o_j may be calculated as $U(o_j) = \sum_{i=1}^{m} u_i(v_i^j)$, where $j = 1, \ldots, |A_T|$. Throughout the rest of the paper we will use values of the utility function normalized against $\max_j(U(o_j))$, so that $U : A_T \rightarrow (0, 1]$. Formally, the problem of the paper is to find the best offer $o_c \in A_T$ that is acceptable to the document-agent P_1 and the execution device P_2, given their utility functions, U_1 and U_2. In other words o_c, called the *contract* between P_1 and P_2, maximizes their utility, i.e.,

$$o_c = \arg\max_{o \in A_T} U_1(o) U_2(o). \quad (1)$$

Since neither party knows its opponent's utility function, nor is willing to reveal its own, exchanging of offers and counter-offers is needed to systematically search space A_T. We use for that an alternating-offer protocol, modeled in the paper as the *simple bargaining game (SBG)*.

The rest of the paper is structured as follows. Section II introduces the method of representing non-functional attributes that are used to build offers exchanged by negotiating parties. The game-theoretic model of that process is defined in Section III. Next, in Section IV results of the simulation experiments are provided, which show that despite of conflicting preferences and incomplete information on each other, the document and the device may reach a satisfying solution. Section V compares the proposed method to the related proposals in the recent literature, and Section VI concludes the paper.

II. Bargaining sets

We have implemented the generic bilateral negotiation model introduced above for $m = 5$ attributes, thus $A_T = \times_{i=1}^{5} A_i$, $i = 1, .., 5$. Each set A_i contains attribute values that specify respective options of the execution context.

1) Performer of the activity (A_1): Disjoint subsets $D, W, J \subset A_1$ contain values indicating respective contexts where the document is allowed to execute automatically on the device (D), where it is not able or is not allowed to execute on its own, so only the user (worker) can perform the activity on its content (W), or where the activity is supposed to be executed jointly by the document and the worker (J).

2) Availability of network resources (A_2): Disjoint subsets $S, E, I \subset A_2$ contain values indicating respective contexts where the execution device is actually separated from its host organization, i.e., no network connection exists or is not allowed by the device owner (S), is connected from outside of its host organization (E), or alternatively from inside of it (I).

3) Performance of network resources (A_3): Disjoint subsets $U, R, M, A, N \subset A_3$ contain values indicating respective contexts where the parameters of the network connection (if any) are unknown (U), or optionaly, wireless (R), telephone modem (M), asymmetric digital subscriber line (ADSL) modem (A), or twisted pair (wire) connection (N), is used by the execution device.

4) Security of the network connection (A_4): Disjoint subsets $P, K, T, C \subset A_4$ contain values indicating respective contexts where the connection (if any) is not secure at all (P), uses wireless (if any) protected by the access key (K), can connect to remote sites using the secure transfer protocol (T), or combines the latter two mechanisms to provide the most secure connection possible (C).

5) Reliability mechanisms supporting interaction (A_5): Disjoint subsets $L, B, F, H \subset A_5$ contain values indicating respective contexts where no support is provided by the document embedded functionality or the execution device system to protect the document content from user errors (L), some backup support is provided by the autosave option (B), failsafe option is provided by the acceptance button, i.e., no changes to the content are permanent until accepted by the worker (F), or high reliability can be provided by combining the later two with the "undo" button and the automatic check of the content performed by the document itself (H).

Note that based on the above model each single vector (offer) $o \in A_T$ specifies in fact a concrete, multi-aspect execution context that may be negotiated by the MIND document and its currently available execution device.

A. Multi-option offers

In our model, the space of offers A_T is discrete, as values of the respective attributes constituting each offer are selected from enumerable sets of available options. Tables I–V illustrate our approach to modeling of multi-option (multi-issue) offers. Specific options that contribute to each respective attribute value $v_i^j \in A_i$, are represented by binary flags. The label of each respective option considered for the given attribute value is listed in the header of each corresponding table, values '0' and '1' listed in each respective column below the option label indicate options 'is' or 'is not' present when calculating value of the given attribute; if the flag may assume two possible values in the context defined by the given table row we denote that by the regular expression $[0, 1]$. The respective attribute values are generated by combining flag values in each row, which we also specify with regular expressions, listed in the rightmost column of each corresponding table. The first letter denotes the respective subsets of A_i, explained in p.II-1–II-5, while digits indicate each meaningful combination of option flags, considered in our current implementation of MIND.

Option values for attribute A_1 are listed in Table I and specify potential performers required by the document to complete its activity: the *Worker* (*Wkr*) using it, the *Embedded Service* (*EmS*) brought by it to the device, some *External Service* (*ExS*) the device should allow it to call, any *Local Tool* (*LoT*) the document may want to use, and any *Local Service* (*LoS*) the document may want to access, when interacting with the local operating system of the device. Combinations of these options' flags distinguish in total nine different subsets of A_1, labeled with symbols listed in the rightmost column of Table I.

TABLE I. Execution context options for 'performer'

Wkr	*EmS*	*ExS*	*LoT*	*LoS*	*Option labels*
0	1	[0,1]	0	[0,1]	D[1-3]
1	0	[0,1]	1	[0,1]	W[1-2]
1	1	[0,1]	[0,1]	[0,1]	J[1-4]

For example, a proactive document that intends to perform its activity without interacting with the worker using the execution device ($Wkr = 0$), only by the means of its embedded functionality ($EmS = 1$) and some external service ($ExT = 1$), would not need any local tool ($LoT = 0$) or service ($LoS = 0$) installed on the device; this particular ’performer’ attribute value ’01100’ would be labeled with $D3$ in our model.

Option values for attribute A_2 are listed in Table II; they specify various resources that should be available to the document when performing its activity on the device. The device connected from inside of the worker’s organization has a *Local IP* (*LIP*), or otherwise an *External IP*. The *Specific Browser* (*SpB*) required by the document may be available on the device, or just *Any Browser* (*AnB*). Similarly, the *Specific Tool* (*SpT*) requested by the document may be provided by the local operating system, or just any *Substitute Tool* (*SuT*). Moreover, the device may be equipped with the *Full Keyboard* (*FKb*), or alternatively a smaller set of *Selection Buttons* (*SeB*) can be provided. Combinations of these options’ flags distinguish in total 22 different subsets of A_2, labeled with symbols listed in the rightmost column of Table II.

TABLE II. Execution context options for ’availability’

LIP	*EIP*	*SpB*	*AnB*	*SpT*	*SuT*	*FKb*	*SeB*	*Option labels*
0	0	0	0	[0,1]	[0,1]	[0,1]	[0,1]	S[1-8]
0	1	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	E[1-7]
1	0	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	[0,1]	I[1-7]

For example, a proactive document that is ready to perform its activity on the device connected from outside of its host organization ($LIP = 0$ and $EIP = 1$) may not care about the type of browser ($SpB = [0, 1]$ and $AnB = [0, 1]$), and assumes no other support from the local system ($SpT = SuT = 0$) nor the device ($FKb = SeB = 0$); four possible ’availability’ attribute values ’01**0000’ would be considered equal and labeled with $E1$ in our model.

Option values for attribute A_3 are listed in Table III and specify performance aspects of the execution device during the current activity. The connection may be *Wired* (*Wre*), or using a TV *Cable* or telephone *Line* (*C/L*), the plain *Telephone Modem* (*TMo*), or any *Wireless* (*Wrs*) network within the reach of the device. The device’s processor (*CPU*) may be above the average specified by the document, as well as the device may provide more memory (*RAM*) as the minimum required by the document. Combinations of these options’ flags distinguish in total 20 different subsets of A_3, labeled with symbols listed in the rightmost column of Table III.

TABLE III. Execution context options for ’performance’

Wre	*C/L*	*TMo*	*Wrs*	*CPU*	*RAM*	*Option labels*
0	0	0	0	[0,1]	[0,1]	U[1-4]
0	0	0	1	[0,1]	[0,1]	R[1-4]
0	0	1	0	[0,1]	[0,1]	M[1-4]
0	1	0	0	[0,1]	[0,1]	A[1-4]
1	0	0	0	[0,1]	[0,1]	N[1-4]

For example, a proactive document accessing only a wireless network ($Wre = C/L = TMo = 0$ and $Wrs = 1$) during its current activity, accepting less powerful CPU ($CPU = 0$) but consuming RAM above average ($RAM = 1$) would have its ’performance’ attribute value ’000101’ labeled with $R2$ in our model.

Option values for attribute A_4 are listed in Table IV and specify security aspects of the current activity. Data can be *Securely Transferred* (*SeT*) if the remote site provides HTTPS. Moreover, if the device uses wireless connection its security may be improved if the network is protected by the *Access Key* (*AcK*). The document may also be *Digitally Signed* (*DSg*), and the local system of the device may be protected by some *Anti-Virus* (*AnV*) tool. Combinations of these options’ flags distinguish in total 16 different subsets of A_4, labeled with symbols listed in the rightmost column of Table IV.

TABLE IV. Execution context options for ’security’

SeT	*AcK*	*DSg*	*AnV*	*Option labels*
0	0	[0,1]	[0,1]	P[1-4]
0	1	[0,1]	[0,1]	K[1-4]
1	0	[0,1]	[0,1]	T[1-4]
1	1	[0,1]	[0,1]	C[1-4]

For example, a proactive document not requiring a secure transfer for its data ($SeT = 0$) but expecting the wireless network protected by the access key ($AcK = 1$), with its content digitally signed ($DSg = 1$) and the local system protected with some antivirus software ($AnV = 1$) would have its ’security’ attribute value ’0111’ labeled with $K4$ in our model.

Option values for attribute A_5 are listed in Table V and specify reliability aspects of the operations performed on the document content during the activity. If the *Acceptance Button* (*AcB*) is provided the user can decide on permanence of the document content modifications. The *Autosave Mode* (*ASv*) provided by the related tool or service can prevent the user from loosing accidentally the content entered so far. Functionality providing any *Automatic Check* (*ACh*) of the content being entered by the worker may improve its correctness, whereas the *Undo Button* (*UdB*) would improve comfort of work of the worker and further reduce the rate of errors he/she can make when modifying the document content. Combinations of these options’ flags distinguish in total 16 different subsets of A_5, labeled with symbols listed in the rightmost column of Table V.

TABLE V. Execution context options for ’reliability’

AcB	*ASv*	*ACh*	*UdB*	*Option labels*
0	0	[0,1]	[0,1]	L[1-4]
0	1	[0,1]	[0,1]	B[1-4]
1	0	[0,1]	[0,1]	F[1-4]
1	1	[0,1]	[0,1]	H[1-4]

For example, if the proactive document performs the activity entirely on its own, it may reasonably expect the device to provide just the acceptance button ($AcB = 1$ and $ASv = ACh = UdB = 0$), to allow the worker to accept the concluded activity and send it to the next activity of its workflow. The related ’reliability’ attribute value ’1000’ would be labeled with $F1$ in our model.

B. Bargaining over option trees

Realistically, negotiating parties will continue choosing offers from a certain subset of offers $C_B \subset A_T$, which we call the *bargaining set*. The range of options, available to the particular document-agent in the given execution context, is determined by the type of a personal device currently in use by the worker. In other words, the device class defines a concrete bargaining set content. We distinguish five basic classes of devices in the current implementation of MIND.

1) Workstations: They are immobile and used mainly at the user's workplace (in office) or at home. If used in office they are usually wired from inside to the organization's network, have access to various secure services, offer reliable interfaces and have relatively high computational power. If used at home they offer the similar level of service, except for network connections that may be external to the organization – if no virtual private network (VPN) connection is possible – and may use various types of modems.

2) Laptops: Performance of their hardware makes them not less powerful than workstations, but owing to their ability to access networks in many ways, including wire, WiFi, and modems, they are more versatile. The only distinction that may be taken into account when characterizing execution contexts they provide is the software they use. If the laptop is a private property of the worker, it may lack some proprietary software tools provided by the organization to its workers, so sometimes substitute tools may have to be used to perform specific activities on the document content – especially when performed from outside of the organization.

3) Tablets: They usually have less computational power than laptops, and are less versatile, due to the limited range of networking solutions they support (embedded WiFi cards or/and ADSL modems). Some specialized software (local tools or services) required to handle properly the content of the document-agent may be unavailable, what can affect results of the activity to be performed.

4) Smartphones: Although their recent technological advances are impressive, they may lack (like tablets) specific local tools or services required to properly process the content brought to the device by the document-agent. These devices also slightly differ from tablets in the networking solutions they support – since the telephone modem may be used as the alternative to WiFi. Reliability of interaction is usually slightly reduced compared to tablets, due to their smaller screen sizes.

5) Cellphones: They are the weakest execution devices, although they can support most elementary execution scenarios performed by workers on the document content, e.g., simple form filling or modifying/accepting simple text prepared by someone else. These operations can be performed if the cellphone currently in use by the worker is able to read email messages, usually via the cellular network.

In Table VI, we specify bargaining sets for various classes of devices listed above. The range of option values of the five attributes in our current MIND implementation reflects the specificity of each device characterized above, which is connected to some network and willing to use it during the entire activity to be performed.

TABLE VI. BARGAINING SETS FOR EXECUTION DEVICES

Bargaining set:	*Attribute A_i value sets/subsets*				
device	1	2	3	4	5
C_1: workstation	A_1	$E \cup I$	$M \cup A \cup N$	$P \cup T$	A_5
C_2: laptop	A_1	$E \cup I$	$A_3 - U$	A_4	A_5
C_3: tablet	A_1	E[4-6]$\cup I$[4-6]	$R \cup A$	A_4	A_5
C_4: smartphone	A_1	$\{E3, I3\}$	$R \cup M$	A_4	L
C_5: cellphone	A_1	E[1-2]	M[1-2]	$P \cup T$	$\{L1\}$

Notice that the range of offers which may be exchanged during negotiation between the document-agent and the device of any particular class is limited, i.e., $C_1, .., C_5$ are rather small subsets of A_T. Either negotiating party, the document and the device, selects offers from the corresponding set C_i independently, with regard to its own valuation of each single offer. The alternating-offer protocol assumed by MIND requires parties to individually sort all offers from the relevant bargaining set in the order implied by that valuation. The ordering of offers preferred by each party is represented by the related *option tree*.

Consider two parties, a document-agent and a laptop, negotiating over C_2; let it consist of just five offers:

$$\begin{aligned}
o_1 &= < D3, E1, A4, T4, H3 >, \\
o_2 &= < D3, E1, M4, T4, H3 >, \\
o_3 &= < D3, E1, R2, K4, F1 >, \\
o_4 &= < D3, E1, R4, C4, H1 >, \\
o_5 &= < D3, I1, N4, T4, H4 > .
\end{aligned}$$

Detailed interpretation of the example offer o_3 attribute values has been presented before in connection to Tables I-V, while the valuation and preferences of all options negotiated by the document and the device are shown in Figures 2 and 3.

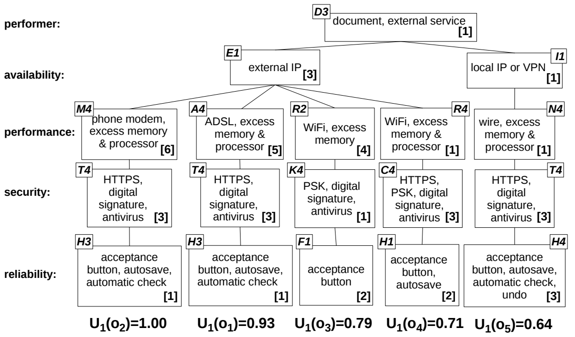

performer:
availability:
performance:
security:
reliability:
D3
document, external service
[1]
E1
external IP
[3]
I1
local IP or VPN
[1]
M4
phone modem, excess memory & processor
[6]
A4
ADSL, excess memory & processor
[5]
R2
WiFi, excess memory
[4]
R4
WiFi, excess memory & processor
[1]
N4
wire, excess memory & processor
[1]
T4
HTTPS, digital signature, antivirus
[3]
T4
HTTPS, digital signature, antivirus
[3]
K4
PSK, digital signature, antivirus
[1]
C4
HTTPS, PSK, digital signature, antivirus
[3]
T4
HTTPS, digital signature, antivirus
[3]
H3
acceptance button, autosave, automatic check
[1]
H3
acceptance button, autosave, automatic check
[1]
F1
acceptance button
[2]
H1
acceptance button, autosave
[2]
H4
acceptance button, autosave, automatic check, undo
[3]
U$_1$(o$_2$)=1.00
U$_1$(o$_1$)=0.93
U$_1$(o$_3$)=0.79
U$_1$(o$_4$)=0.71
U$_1$(o$_5$)=0.64

Figure 2. Example option tree of the proactive document

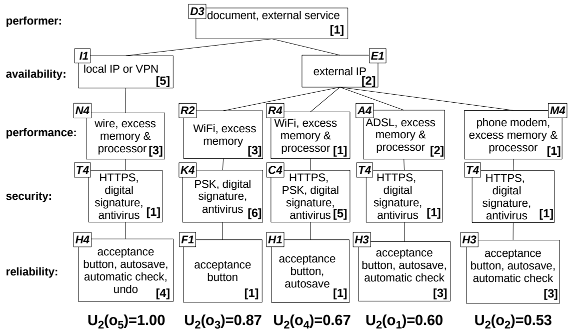

performer:
availability:
performance:
security:
reliability:
D3
document, external service
[1]
I1
local IP or VPN
[5]
E1
external IP
[2]
N4
wire, excess memory & processor
[3]
R2
WiFi, excess memory
[3]
R4
WiFi, excess memory & processor
[1]
A4
ADSL, excess memory & processor
[2]
M4
phone modem, excess memory & processor
[1]
T4
HTTPS, digital signature, antivirus
[1]
K4
PSK, digital signature, antivirus
[6]
C4
HTTPS, PSK, digital signature, antivirus
[5]
T4
HTTPS, digital signature, antivirus
[1]
T4
HTTPS, digital signature, antivirus
[1]
H4
acceptance button, autosave, automatic check, undo
[4]
F1
acceptance button
[1]
H1
acceptance button, autosave
[1]
H3
acceptance button, autosave, automatic check
[3]
H3
acceptance button, autosave, automatic check
[3]
U$_2$(o$_5$)=1.00
U$_2$(o$_3$)=0.87
U$_2$(o$_4$)=0.67
U$_2$(o$_1$)=0.60
U$_2$(o$_2$)=0.53

Figure 3. Example option tree of the laptop execution device

Nodes in each example option tree are labeled with option values specified in Tables I–V, while utility of each single attribute value calculated by the party is specified in square brackets. Utilities of the respective offers of C_2 are listed at the bottom of the tree in a normalized form, as explained before; see in Figure 2 that o_2 is the document's most preferred offer, since $\sum_{i=1}^{5} u_i(v_i^2) = 1 + 3 + 6 + 3 + 1 = 14$ and $U_1(o_2) = 14/14 = 1.0$, whereas o_5 is its least preferred offer, with $U_1(o_5) = 8/14 = 0.64$.

III. SIMPLE BARGAINING GAME

We model negotiation between the document-agent and the execution device as a *multi-stage game*. It consists of *stages*, denoted as $(\alpha_1^{2n}, \alpha_2^{2n+1})$, where $n \in N \cup \{0\}$ (or $n \in Z_{\geq 0}$ for brevity) denotes the stage number, α_i^k denotes a *move* of each respective player P_i, $i = 1, 2$, and k is the move number. After making move α_i^k player P_i chooses offer $o \in C_B$ and valuates it with its payoff function $\pi_i : A \times N \rightarrow [0, 1]$, calculated as $\pi_i(o, k+2) = \delta_i \pi_i(o, k)$, where $\delta_i \in (0, 1]$ is called a *discount factor* and $\pi_i(o, 0) = U_i(0)$. Each player has its own discount factor that remains constant during the game. Throughout the rest of the paper we will denote the opponent of player P_i by P_{-i}, their respective utility functions by U_i and U_{-i}, payoff functions by π_i and π_{-i}, and discount factors by δ_i and δ_{-i}.

Rules of our simple bargaining game are the following:

1) The game is started by player P_1.
2) Players P_i, $i = 1, 2$ keep in secret their private information, including U_i, δ_i and π_i, but share knowledge on the bargaining set C_B.
3) Utility values of players' offers are discounted at each transition to the next stage.
4) Players exchange offers until the game is concluded, i.e., one of the players accepts an offer or quits the game.
5) The game is concluded by player P_i when:
   a) P_i repeats its own offer o' what implies quitting the game by P_i.
   b) P_i repeats player's P_{-i} offer o'' what implies accepting it as the contract and exiting.

A. Negotiation algorithm

A generic form of the algorithm implemented by each SBG player (thread) is given below. Two threads operate on the same resource, which is the bargaining set and individual indexes for marking offers in it as sent, received or not yet sent nor received. Functions of the form $max\Pi_{[B,R,N]}$ calculate offers to be chosen from the bargaining set or its respective subsets, according to π_i and the move number k, as explained before.

```
Public: Bargaining set $C_B$; sets of offers: received $C_R$, sent
    $C_S$, remaining $C_N$ (all initially empty); received $o_R$ and
    sent $o_S$ offers; move number $k$ (initially $k = 0$).
Private: Discount factor $\delta_i$; payoff function $\pi_i$;
if $k = 0$ then
    {Opening move}
    $o_S \leftarrow max\Pi_B$;
    $send(o_S)$;
else
    $receive(o_R)$;
    if $o_R \in C_S$ then
        {Opponent has accepted the offer}
        $terminate$;
    end if
    $C_R \leftarrow C_R + o_R$;
    $C_N \leftarrow C_B - (C_R + C_S)$;
    if $C_N = \emptyset$ then
        {Last move}
        $o_S \leftarrow max\Pi_R$;
        $send(o_S)$; $terminate$;
    else
        {Intermediate/penultimate move}
        $o'_S = max\Pi_R(k)$; $o''_S = max\Pi_N(k)$;
        if $\pi_i(o'_S) \geq \delta_i \cdot \pi_i(o''_S)$ then
            $send(o'_S)$;
        else
            $send(o''_S)$;
        end if
    end if
    $k = k + 1$
end if
```

B. Collaboration agreement

Table VII specifies history of SBG played by document-agent P_1 and execution device P_2, which shared the example bargaining set $C_2 = \{o_i | i = 1, \ldots, 5\}$, with their respective option trees specified in Figures 2 and 3. The document started negotiation by offering to the device its most preferred option o_2 in move α_1^0. Then after the next four moves, option o_3 offered by the device was finally accepted by the document. The negotiated contract between the proactive document and the device provides the former with the commonly agreed *execution context* supplied by the latter.

TABLE VII. BARGAINING OVER EXAMPLE OPTION TREES

Stage	Move	Player	Offer o_i	$U_i(o_i)$	$U_{-i}(o_i)$
0	0	P_1	o_2	1.00	0.53
0	1	P_2	o_5	1.00	0.64
1	2	P_1	o_1	0.93	0.60
1	3	P_2	o_3	0.87	0.79
2	4	P_1	o_3	0.79	agreed

In the example we have used simplified option trees to keep them small. In real applications involving MIND documents the upper bound for the maximum size of a single option tree could be as high as the product of the numbers of attribute value labels used in Tables I-V, which is well over 5000, whereas the upper bound for the number of possible negotiation histories is a product of their respective permutations, in the order of magnitude of 10^{13}.

IV. SIMULATION EXPERIMENT

Our simple bargaining game can provide a solution of (1) when neither player P_i knows its opponent's U_{-i}, δ_{-i} nor π_{-i}, known in the literature as the *Nash equilibrium.*

In order to show the above let us formally define the simple bargaining game $SBG = \{P, D, U, C_B, S, \Pi, T\}$ as a multi-stage game. Sets $P = \{P_1, P_2\}$, $D = \{\delta_1, \delta_2\}$ and $U = \{U_1, U_2\}$ are self explanatory. In the bargaining set C_B we distinguish subsets $C_i, C_{-i} \subset C_B$ of offers, submitted respectively by P_i and P_{-i}. Therefore, the set including offers not yet submitted by any party would be $C' = C_B - (C_i \cup C_{-i})$, and $C' \subset C_B$. Set $S = S_i \cup S_{-i}$ consists of strategies used by the respective players; each single strategy $s_i \in S_i$ is a function $s_i : Z_{\geq 0} \rightarrow C_B$ that associates each possible move α_i^{k+2} of P_i that follow α_i^k with the relevant offer from C_B. There are at most $|C_B|$ possible strategies for each player P_i, while the maximum number of steps is $k_{max} = |C_B|$. Set $\Pi = \{\pi_1, \pi_2\}$ consists of the respective players' payoff functions; for any offer $o \in C_B$, submitted in the k-th move, each respective payoff function returns $\pi_i(o, k) = \delta_i^k U_i(o)$ if $o \in C_{-i}$, or $\pi_i(o, k) = 0$ if $o \in C_i$. Finally, $T(o)$ denotes a condition for concluding the game, i.e., for any $o \in C_B$, $T(o) = false$ if $o \notin (C_i \cup C_{-i})$, otherwise $T(o) = true$.

A. Equilibrium strategy

For each stage of SBG involving offer o_i, submitted by P_i, and counteroffer o_{-i}, submitted by P_{-i}, the notion of a *strategy profile* is used; it is defined as a pair of strategies $\langle s_i, s_{-i} \rangle$, where $s_i \in S_i$ and $s_{-i} \in S_{-i}$. We will say that the strategy profile $\langle s'_i, s'_{-i} \rangle$ is the Nash equilibrium of SBG if $\langle \pi_i(s'(k)), \pi_{-i}(s'_{-i}(k+1)) \rangle \succeq \langle \pi_i(s(k)), \pi_{-i}(s'_{-i}(k+1)) \rangle$ for any $s_i \in S_i$ and each player in each stage n; by '$\succeq$' we denote a pairwise comparison operator of 2-element vectors. Strategies s'_i and s'_{-i} are called *equilibrium strategies*; each one is said to provide the *best offer* of the respective player in each stage of the game. When looking for the solution to (1), instead of attempting to find each other's equilibrium strategies each player may just assume that the offers submitted by its opponent in any stage are actually its best responses.

When implementing the simulation experiment we have adopted the following rationale of strategies used by players in various stages of the game. All estimations are made from the point of view of P_i, thus if the move is to be made by P_i, we assume it to choose the offer of the highest payoff from all offers available to it. Alternatively, if the move is to be made by P_{-i}, we assume it to choose the offer of the average payoff calculated for all offers available to it. The reason for the latter is that from the perspective of P_i all offers made by P_{-i} are equally probable, so statistically the payoff that P_i can get is equal to their average value. We calculate that average with the auxiliary function $\rho(o, C', f) = 1/|C'| \sum_{o \in C} f(o)$, where $o \in C'$, and $f : C' \to [0, 1)$ is a function used to valuate offers with the payoff function. For brevity we will skip using the step number argument of the payoff function throughout the rest of the paper.

1) Last move: Strategy $s_i(k_{max}) = \arg\max_{o \in C_{-i}} \pi_i(o)$ is used if P_i makes the last move. Alternatively, if the last move is made by P_{-i}, strategy $s_{-i}(k) = o_{-i}$ is used, where $o_{-i} \in C_i$ satisfies equation $\pi_i(o) = \rho(o, C_i, \pi_i)$. In the first case P_i must choose the best offer from those that have been already presented by P_{-i}, for otherwise (according to rules of *SBG*) it would get the zero payoff. For the same reason, offer o_{-i} made by P_{-i} must be chosen from all available offers in C_i and of the payoff closest to the average payoff of all its offers.

2) Penultimate move: The strategy that has to be used is $s_i(k) = \arg\max_{o \in \{o'_i, o''_i\}}(\pi_i(o'_i), \delta_i \pi_i(o''_i))$, if P_i makes the move, where $o'_i = \arg\max_{o \in C_{-i}} \pi_i(o)$ and $o''_i \in C - C_{-i}$ satisfies $\pi_i(o) = \rho(o, C_i, \pi_i)$. Alternatively, $s_{-i}(k) = o_{-i}$ is used if P_{-i} makes the move, where $o_{-i} \in C - C_{-i}$ satisfies $\pi_i(o) = 0.5(\pi_i(o'_{-i}) + \delta_{-i} \pi_i(o''_{-i}))$, $o'_{-i} \in C_i$ satisfies $\pi_i(o) = \rho(o, C_i, \pi_i)$, and $o''_{-i} = \arg\max_{o \in C_B - C_i} \pi_i(o)$. Unlike in the last move, the player has to accept one of the received offers or make a new one, not yet presented by either player. In the penultimate move only one such offer is left in C_B. If P_i makes the move it finds o'_i giving the best payoff of all offers received already from P_{-i} (as described in p. IV-A1 before) and estimates payoff of offer o''_i that P_{-i} may submit in its next move. The latter may be any offer submitted already by P_i and the last one left in C_B – all equally probable, thus the average payoff is assumed. The choice is between o'_i and discounted o''_i, since the latter concerns continuing the game in the next move. If P_{-i} makes the penultimate move it can either accept offer o'_{-i} from those made already by P_i or reject it by submitting counteroffer o''_{-i}. In order to assess payoff of o'_{-i} player P_i has to calculate the average payoff of all its offers to P_{-i} (in the case when P_{-i} may decide to accept one of them), whereas o''_{-i} is expected to give the highest payoff of the offers made by P_{-i} that P_i may eventually accept. For o'_{-i} the average payoff is calculated, as accepting P_i's offers by P_{-i} is equally probable. The choice is then between o'_{-i} and discounted o''_{-i}, since the latter concerns continuing the game in the next move. The average payoff of o'_{-i} and o''_{-i} is assumed, as both choices by P_{-i} are equally likely to P_i.

3) Intermediate move: In any move $0 < k < k_{max} - 1$ P_i uses strategy $s_i(k) = \arg\max_{o \in \{o'_i, o''_i\}}(\pi_i(o'_i), \delta_i \pi_i(o''_i))$, where $o'_i = \arg\max_{o \in C_{-i}} \pi_i(o)$, and $o''_i \in C'$ satisfies $\pi_i(o) = \max_{o \in C'}(1 - \delta_{-i})\pi_i(o) + \delta_{-i} U_i(s_{-i}(k+1))$. Alternatively, strategy of P_{-i} should be $s_{-i}(k) = o_{-i}$, where $o_{-i} \in C'$ satisfies $\pi_i(o) = 0.5(\pi_i(o'_{-i}) + \delta_{-i} \pi_i(o''_{-i}))$, $o'_{-i} \in C_i$ satisfies $\pi_i(o) = \rho(o, C_i, \pi_i)$, and $o''_{-i} \in C'$ satisfies $\pi_i(o) = (1 - \delta_i)\rho(o, C') + \delta_i \pi_i(s_i(k+1)))$. Rationale for strategies $s_i(k)$ and $s_{-i}(k)$ is the same as in the penultimate move, i.e., offers o'_i and o'_{-i} concern accepting the respective opponent's offer, while o''_i and o''_{-i} concern rejection of the opponent's offer and continuation of the game by submitting new offers. The former are calculated in the same way as in p. IV-A2, whereas calculation of the latter should reflect possible acceptance or rejection during subsequent stages. Therefore, if P_i decides to continue the game the estimated payoff splits in two parts, proportional to discount factor δ_{-i} of its opponent: payoffs that may be get for new offers from C' if accepted, and continuation of the game by player P_{-i} in the next move with strategy $s_{-i}(k+1)$) if not accepted. If P_{-i} decides to continue the game the estimated payoff splits in two parts as well, proportional to discount factor δ_i of its opponent: the average payoff that may be get for new offers from C', if accepted, and continuation of the game by player P_i in the next move with strategy $s_i(k+1)$, if not accepted.

4) Opening move: Strategy of P_1 in move $k = 0$ (starting the game) is $s_i(0) = \arg\max_{o \in C_B}(U_i(o))$. It has no hints concerning its opponent's utility, so the only way to maximize its utility is to choose its most valued offer from C_B.

B. Simulation results

In the experiment we simulated P_1 (proactive documents) using the rationale of selecting best offers in each possible type of step described before. Some generic features of such documents were considered. We distinguished *protected* documents from *open* ones, by taking into account whether they need any secure connection to perform their activity or not, and *heavy* documents from *light* ones, by considering whether they may require CPU power and the amount of RAM above or below some average levels when doing that. Based on that we have defined four classes of documents used in the simulation experiments. Their respective bargaining sets are listed in Table VIII.

TABLE VIII. BARGAINING SETS FOR PROACTIVE DOCUMENTS

Bargaining set:	*Attribute A_i value sets/subsets*				
document class	1	2	3	4	5
C_{ph}: protected & heavy	A_1	$E \cup I$	$[U, R, A, N]4$	$A_4 - P$	A_5
C_{pl} : protected & light	A_1	$E \cup I$	A_3	$A_4 - P$	A_5
C_{oh}: open & heavy	A_1	$E \cup I$	$[U, R, A, N]4$	A_4	A_5
C_{ol} : open& light	A_1	$E \cup I$	A_3	A_4	A_5

For each document class 20 random option trees were generated, each one including offers with attribute values listed in the respective row of Table VIII. Bargaining was performed over each possible pair of options trees, one tree for the particular document class and another for the particular device class; in each case $C_B = C_{doc} \cap C_{dev}$. With five device classes 100 simulations were performed for each document class. In each single simulation the offers in option trees of the respective negotiating parties were sorted in the opposite order to each other, to ensure the maximum possible number of negotiation steps. The discount factor for each party was $\delta = 0.8$. The results are shown in Figure 4.

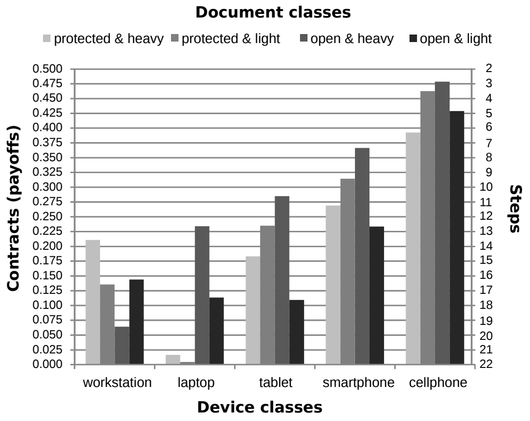


Figure 4. Contracts negotiated by proactive documents and devices

The fair payoff, which may be calculated based on (1) by some hypothetical third party knowing utility functions of the two negotiating parties, was close to 0.5. It may be seen that in general the smaller the number of available options (thus also the bargaining set size) the smaller is the number of negotiation steps required to reach the contract – the document was able to reach the agreement and get close to the fair result with the cellphone device in just a few steps, whereas negotiating with the laptop or workstation devices was possible in about 20 steps.

V. Related work

Our problem of finding the execution context that can satisfy the proactive document intending to perform a given activity on a dynamically changing mobile device, relates to two specific areas: representation of available options as offers, and implementing negotiations as games. One difficulty with the above is the discrete range of offers that the negotiating parties have to search to find the solution. Another is the semi-cooperative setting implied by the fact that partners share the bargaining set but keep their preferences secret. Modeling multiple options with functions as the continuum of offers allows finding contracts using various optimization techniques, e.g., swarm optimization, as demonstrated in [3]. Unfortunately, dealing with non-functional options makes such optimization inherently difficult to implement. In [4] the concept of the Web service modeling ontology (WSMO) has been used to develop a mechanism capable of handling that in a way enabling agents to find the most suitable resources for performing the requested activity; the proposed mechanism also provided for resolving between alternative (conflicting) offers based on the argumentation theory. This approach suits well broker agents recommending services to user agents, however in the case of two agents competing to win as much wealth as possible the game theoretic approach seems to be easier to implement.

A generic approach to the problem of multiple issue negotiation with no information about the opponent has been proposed in the literature [5], but the formal mathematical proof of the convergence of the monotonic concession strategy, which our SBG implements with option trees, was not provided until [6]. It has been shown there that offers and counteroffers, selected by each party according to the amount of concession the party can accept in the current round, are getting closer in the utility space until the contract is agreed. For each party a hyperquadratic utility function was assumed – general enough to simulate negotiation protocols, but imposing an unnecessary limit on how agents may implement valuation of offers and not allowing for discrete values.

VI. Conclusion

Our approach based on option trees does not assume any class of utility functions, except that they should be injective in order to enable sorting the trees and to ensure monotonicity of preferences. The linear additive utility function defined in Section I-B has been used by us just to simplify generation of option trees for simulation experiments described in Section IV-B. The model proposed in the paper has been recently expanded with the learning capability, based on the history of interaction and the concept of policies. Simulation experiments indicate that properly trained document-agents can recognize preferences of devices and guess the contract in just a couple of steps [7]. The necessary 'knowledge' is carried by documents as weights needed to set up a really simple neural network, which the LWE client would be able provide on the execution device.

Acknowledgment

This work was supported in part by the National Science Center, Poland, under grant DEC1-2011/01/B/ST6/06500.

References

[1] M. Godlewska and B. Wiszniewski, "Smart email: Almost an agent platform," in Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering, ser. Lecture Notes in Electrical Engineering, T. Sobh and K. Elleithy, Eds. Springer International Publishing, 2015, vol. 313, pp. 581–589.

[2] J. Kaczorek and B. Wiszniewski, "Augmenting digital documents with negotiation capability," in Proc. 2013 ACM Symposium on Document Engineering (DocEng'13). ACM, Sep. 2013, pp. 95–98, doi:10.1145/2494266.2494305.

[3] Z. Wang and L. Wang, "Adaptive negotiation agent for facilitating bi-directional energy trading between smart building and utility grid," IEEE Transactions on Smart Grid, vol. 4, Mar. 2013, pp. 702–710, doi:10.1109/TSG.2013.2237794.

[4] A. Caballero, A. Muòoz, J. Soto, and J. A. Botía, "Resource assignment in intelligent environments based on similarity, trust and reputation," J. Ambient Intell. Smart Environ., vol. 6, Mar. 2014, pp. 199–214, doi:10.3233/AIS-140253.

[5] G. Lai and K. Sycara, "A generic framework for automated multi-attribute negotiation," Group Decision and Negotiation, vol. 18, Jul. 2009, pp. 169–187, doi:10.1007/s10726-008-9119-9.

[6] R. Zheng, N. Chakraborty, T. Dai, K. Sycara, and M. Lewis, "Automated bilateral multiple-issue negotiation with no information about opponent," in Proc. 46^{th} Hawaii Int. Conf. on System Sciences (HICSS 2013). IEEE, Jan. 2013, pp. 520–527, ISBN: 978-0-7695-4892-0.

[7] J. Kaczorek and B. Wiszniewski, "Document agents with the intelligent negotiation capability," in Proc. Knowledge and Cognitive Science and Technologies (KCST 2015). IIIS, Jul. 2015, in press.

No Outstanding Surprises when Using Social Media as Source for Weak Signals?

First Attempt to Discuss the Impact of Social Media Sources to Detect Surprising Weak Signals

Robert Eckhoff, Mark Markus,
Markus Lassnig, and Sandra Schön
Innovation Lab
Salzburg Research Forschungsgesellschaft
Salzburg, Austria
markus.lassnig@salzburgresearch.at

***Abstract*—Enterprises as well as research institutions are interested to find very early signs for future trends, disruptions or other emerging big changes. Such “weak signals” may also be detected within user-generated social media content. Information technologies support searching, analyzing and interpreting social media data. According to our experiences with an approach called “innovation signals”, none of our detected weak signals was an outstanding surprise for our industry partners. Within this paper, we try to validate these experiences with a look into weak signals theory, and whether similar experiences can be found in the extant literature. While we were not able to find conclusive evidence that our conclusions are common (or that they are not), we present a set of possible explanations of this phenomenon. Our paper has to be seen as a first discussion of the topic, which should be a first step to validate researchers’ experiences and to initiate a potentially controversial discourse about it.**

Keywords-Weak signals; detection; innovation; surprise; criteria.

I. Introduction

Weblogs, discussion forums or mailing lists are seen as a worthy and relatively accessible source for trend research. At Salzburg Research, we developed an approach to detect “weak signals” [1]. Weak signals are find very early signs for future trends, disruptions or other emerging big changes. According to theory, weak signals may also be surprising. Our approach of weak signals detection is a combination of computer-supported analysis and social scientific interpretation that uses social media content as source of primary data. This approach, called “innovation signals”, was used to get insights into three branches from industry partners [2] [3]. Whereas the feedback and customers’ satisfaction was very good, the involved researchers still got the impression that their results were no big surprises for their customers respectively industry partners. Within this paper, we will analyze if other researchers in social media make similar experiences and how this phenomenon might be explained. If it is common that “weak signals” are not surprising, this should be influence theory of as well as counseling in weak signals detection.

Within this paper, we try to validate our experiences with a look into weak signals theory, and whether similar experiences can be found in the extant literature. Our paper has to be seen as a first discussion of the topic, which should be a first step to validate researchers’ experiences and to initiate a potentially controversial discourse about it.

II. Detecting Weak Signals with Social Media

“Weak signals” are seen as potentially important signs for future developments with big impact on companies. According to Ansoff, weak signals are “imprecise early indications about impending impactful events” [1]. All that is known, he proceeds, “is that some threats and opportunities will undoubtedly arise, but their shape and nature and source are not yet known.” Compared with other levels of knowledge about the future, weak signals are the vaguest and possibly earliest kind of information, especially compared with “drivers” or even “trends” [4]. Being able to recognize such weak signals for future trends and developments might be a chance: Organizations can use the time for management decisions concerning innovative adaptations or new developments within the firm, the product or any other impacted unit. The collection and detection of weak signals could “be a key to anticipating change in advance and avoid letting them cause surprise” [5], see Figure 1.

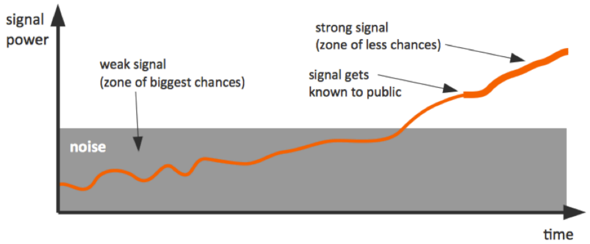

signal power
strong signal (zone of less chances)
weak signal (zone of biggest chances)
signal gets known to public
noise
time

Figure 1. Evolution of a weak signal, building upon Coffman [7] and Steinmüller [20]

A very common approach towards weak signals detection is the use of social media content as a source. Social media are Web tools and services that allow to communicate, to collaborate, and to share information. For example, social networks, discussion forums, Wikis, Weblogs or mailing lists are such applications. Within social media customers, colleagues, experts and others discuss brands, products and services, or related topics and issues. Therefore, social media is not only a way to share and discuss online, but also a good source for research and strategic planning. Information technologies support searching, analyzing and interpreting social media data. Typically, but not always, computer analysis supports the detection of weak signals [6]. This is especially illustrated in a comparison of about 20 social media monitoring tools with regard to their applicability for detecting weak signals [7].

In theory, relatively vague and new topics should pop up when detecting weak signals, but in our experience we did neither find surprising new trends nor previously unknown weak signals. Customers and experts within our search fields were satisfied with our results and were happy with facts, figures and illustrative content (e.g., citations), but we could not detect genuinely surprising weak signals for them.

III. Research question and design of (first) research

Within this contribution, we are interested in whether the limitations we experienced are limited to our approach of detecting weak signals using social media or whether this is a more common phenomenon that we share with other researchers in the field, i.e. whether the problems we encountered are general problems of the research field. Additionally, we try to find explanations for this phenomenon and thus, we try to find answers to the following questions:

- Did others have similar experiences with no outstanding surprises when using social media as a source for the detection of weak signals?
- If this is the case, what are potential explanations for this phenomenon?

This contribution is not a comprehensive study but a first step into a topic that could influence the understanding and practice of weak signals detection in future, if others support our impressions and argumentation. Therefore, we aim to get feedback and to initiate further discussion.

In the following, we will give insights into our first desktop-based research and considerations about our experiences. To start with, we discuss if according to the theory weak signals should be surprising or not. Building on this, we will describe our own setting and the experiences we made with the detection of weak signals. Then, we will present our results of similar experiences we found in the literature. This is followed by a first set of explanations of this phenomenon. Our paper is meant as a first discussion of the topic, which might be of broader interest for researchers in the field.

IV. Should weak signals surprise? What the theory says

According to Coffman [8] weak signals may also be surprising: They are “new and surprising from the signal receiver's vantage point (although others may already perceive it)”. Additionally he wrote that weak signals are often “scoffed at by people who ‘know’” [8]. Both citations can be seen as an explanation, that weak signal may be a surprise (or at least have to be for some), but that does not mean that experts should be surprised by every single detected weak signal. Also, Ilmola and Kuusi see the potential of surprise when they see its bounding to “surprise value”: “We can define that the information content of a signal or new information produced by it depends besides on the relevance of the signal also on its surprise value to the actor” [9, p. 913]. The potential to surprise, “because they are new and even surprising” “can break our prevailing mental models and encourage us to think differently” [5, p. 7]. Kuosa even directly associates weak signals with surprises in his current summary of weak signal detection: “In contemporary futures studies the term weak signal refers to an observed anomaly in the known path of transformation that surprises us somehow” [10, p. 22].

Surprises are also mentioned in weak signals theory as argument why weak signals detection is important: “Collecting and analyzing weak signals could be a key to anticipating changes in advance and avoid letting them cause surprise” [7] [similar 11].

Nevertheless, this explanation does not give a ratio of surprising weak signals or their level of surprise for industry experts. But from the theoretical base it is clear that weak signals should at least have the potential to surprise.

V. The context of our experiences: innovation signals within social media

In this part, we introduce the background of our experiences, the research project “innovation signals” in order be able to compare it similar approaches [2] [3]. The approach of innovation signals and the technology was developed and used within the project “Innovation Signals – Development of a Social Web Innovation Signals Amplifier System”, funded by Austrian Research Promotion Agency.

A. The approach of “innovation signals”

The concept called “Innovation Signals” exploits user-generated content for strategic innovation purposes by combining quantitative data mining [12] and qualitative methods. The Innovation Signals research approach does not rely on technology alone, but unfolds in the development of social media mining technology in unique combination with an interpretative methodology. The process is described as follows.

a) Set-up: The set-up of Innovation Signals research mimics the traditional research design of empirical social science. The main goal is to formulate research hypotheses and define conceptual search terms, which contain between 20 and 50 English and German keywords. Then, 40 to 50 publicly accessible social web sources (forums, communities, blogs, newsgroups) are identified and quickly assessed, according to a catalogue of criteria (e.g., quality of contents, length of contributions, intensity of contribution).

b) Detection and monitoring: The social media mining-based technology provides automatic detection of relevant keywords and topics of interest in sources selected beforehand. It first extracts a large amount of user posts (e.g., 200,000 posts) and then, automatically detects emerging keywords, topics and sentiments from compiled discussions and users' publicly available opinions.

The Innovation Signals technology provides answers to the questions in the context of product development and trend detection such as: How do users talk about existing products? What are critical issues? What issues are discussed very intensively? What are emerging topics? How do topics change over time? The technology enables experts to analyze and interpret detected innovation signals in an easy and intuitive way and also, to save the most important posts for additional manual analysis and coding.

c) Identification and contextualisation of innovation signals: The automated analysis of textual content enables an efficient information processing, but the machine-processed information still remains ambiguous. In order to enable effective research, the interactions in the social web must be structured additionally and analyzed with social science methodology. This means to associate user generated content with relevant statistics, trends and theories to amplify the meaning of the information and to understand the consumers' conversations better and in a broader context.

d) Translation into business opportunities: This phase of the research process utilizes user generated content (in close co-operation with customers/companies) as an additional information source for strategic decision making with regard to the kind of innovation (product, process, business models, strategic innovation fields) to be pursued in order to determine the focus of the product innovation and market strategies and/or to detect new markets and new ideas.

B. No outstanding surprises with"innovation signals"

Three bigger and some smaller practical use cases were delivered within the project "innovation signals" – for different branches and industry partners. Fields of application were the skiing industry, car mobility, and the energy sector.

The general feedback in all three use cases was that our customers said that the results do not surprise them, but rather support their hypotheses. For instance, the social media mining project for a large automobile service provider showed that drivers are increasingly dissatisfied with the costs of mobility. However, this observation did not qualify as groundbreaking news to our client. The analysis in the energy sector could after major contextualization deliver at least some food for thought as we could show that the customer journey towards a solar panel on the roof was paved with negative experiences that an energy provide could relatively easily provide its customers with. Finally, the analysis conducted in the realm of the skiing industry was able to identify some features that users would look for when reviewing new skiing products. However, even though this last analysis was by far the most specific, the results failed to surprise our customer.

Readers might wonder, when and why we came to the conclusion that our research did not result in outstanding surprises for our customers and partners. When we tried to develop guidelines for others that are interested in social media mining for innovation purposes, we quickly realized that in nearly all cases the expected surprise was not reached. One of our most important guidelines, delivered through an expert discussion, is: "Do not expect outstanding surprises" [13]. After writing the respective paper, we wanted to know more about this experience, resulting in this contribution.

VI. Are our experiences individual or common?

Are we alone with our experiences? Might our impression be a fault, misinterpretation or artifact?

Within our analysis of approaches for the detection of weak signals we read all publications available to us on the topic. When reading them, we did not recognize any hints that detected weak signals produced surprising results. Moreover, we failed to find that detected weak signals were surprising at all. To validate this impression, we took a set of current papers with a concrete description of weak signals detection by social media mining [7] [14] and other methods [9] [15] [16]: None of them reported surprising results or surprises when presenting results to the final customer.

We searched within literature databases like Google Scholar and Sciencedirect for "weak signals" and "surprise" and found a long list of hits; most of them refer directly to the idea that the detection of weak signals is seen as a strategy to prevent surprises for enterprises. We scanned all abstracts and where possible (via open access), the papers as such for clues of finding surprising results from weak signals detection.

To sum up, our literature review did not find any statement or even study result that the detection of weak signals produced surprising results or respective surprising weak signals. Of course, this might not be sufficient evidence to argue our point, as "surprise" is not a typical criterion to measure research's quality. It would not be very common in a research paper to address customers' surprise about the research results.

Nevertheless, given that "finding surprising insights" is one of the hallmarks of weak signals detection, the lack thereof in the empirical literature is indeed surprising to us. After reflecting and discussing the issue and other contributions within the field, we decided to try and spark off the debate, even if at this point in time we cannot supply conclusive evidence that our impressions holds true for all other efforts to detect weak signals as well.

VII. Potential explanations for this phenomenon

We are not able to present a good empirical base or data about our impression of a very small (if any) rate for outstanding surprises when detecting weak signals within social media. Nevertheless, we try to collect some explanations for this phenomenon.

A. Characteristic of noisy social media and limits of current mining approaches

Per definition, weak signals are normally hidden in the "noise of the daily produced data" [7]. Typically, weak signals in social media are tricky to detect. Approaches from social media mining typically use combinations of clustering approaches as well as counting algorithms, eventually using semantic analysis, in addition. Following this, singular postings with differing content cannot be detected. Every new topic or issue must be mentioned and discussed from more people within a certain time span, before the approaches might be able to detect such signals. Therefore, even "weak signals" must surpass a threshold, measured as a certain amount of people or postings, to get recognized as "weak signals". Within other approaches and sources, where "weak signals" for example are collected manually as very astonishing or annoying stories of individuals, the potential to detect a single story is potentially given [17]. With data mining approaches and social media as source the probability to find a completely seldom or new incidence seems not possible due to the sheer amount of signals so detected.

B. Filters may avoid the detection of surprising weak signals

So far we did not succeed in automatically filtering those signals that are new and relevant to our context. It is not only complicated to detect such weak signals in the noise, but also to keep and amplify them (i.e. the "real weak signals") for use in a final conclusion to the client. Which of the hundreds of signals is the weak signal that anticipates a future trend? Presenting the client with all possible weak signals isn't a good option (we actually tried this), nor is picking a few that "look promising" good scientific practice. Already Coffman described the issue of "people who 'know'" that scoffed at weak signals [12].

Additionally, "cognitive filters" influence the final detection of what is coined as "weak signals" and which weak signals might be overseen. Ansoff named mental filters that influence the realization of weak signals within enterprises: The "surveillance filter" focuses on special parts of the environment which might deliver data and the "mentality filter" is responsible for the selection that comes to perception in a firm. The third filter, the "power filter" might be the influence of managers that purposely neglect information. As described and empirically shown in [1], such mental filters can be influenced by the setting. Filters can be opened by "virtual process, open question and anonymity" (p. 919). The filters can be deepened through "focused scope, close to the current strategy, strong requirement for plausibility and probability in the social interaction process" and others [1, p. 919].

C. Our customers are experts, not newbies

Typically, customers of weak signals detection are experts. Hence, they of course hear a lot around their key topics, they are aware of all the things going on in their main field of interest. Their wish to detect "weak signals" seems to be driven by their interest in getting more factual knowledge, deeper insights and first figures about the development of topics etc. They want to be the first to know. Of course, their expertise limits the potential of surprise.

Another point that might influence the impression, that customers are not surprised by the found weak signals, might also be explained by a cognitive bias: If I am an expert, I should already know everything (see hindsight bias [18]). Further work might be aware of such psychological influences.

D. Epistemological limits of surprising weak signals detection

On a more philosophical base, we can also argue the epistemological background of the detection of weak signals, especially if it concerns media and technology. "The current mediosphere strongly influences the thinking on media, and therefore the thinking of all, including experts in current study design without possibility to reflect this phenomenon" [19]. Of course, we are limited to what we are able to detect because we see it. Other developments might be blind spots, as we are not aware of them: "Blind spot means, we do not realise it", "it is a spot we cannot see" [19]. Such blind spots of thinking and knowing can be age specific, and related to our cultural techniques and the predominant medium of our society [20]. If someone would be able to imagine our blind spots, deep surprises might be possible. Building on such an argumentation, future hindsight projects might be able to see such early signs of development, which would be pretty surprising for us from today's perspective. But this will only be possible through future knowledge and awareness of a new human age and mediosphere.

VIII. Discussion and conclusion

As described, our discussion paper is meant as a first step of deeper consideration of our experiences that are not necessarily common experiences for those using social media mining for the detection of weak innovation signals.

This first discussion might be a starting point for researchers' and practitioners' who made similar experiences – or even more interesting: other experiences. So we would be happy to hear your stories, if and when your detection of weak signals left surprised recipients behind. To manage expectations at our side as well as our customers' side, today, we do not emphasize the "originality" or "surprise-factor" our detection of weak signals might deliver, until we believe this to be the case after reviewing the first results.

Additionally, our starting point was the failure to find surprising weak signals, which might be related to the usage of social media as our data-mining source. Of course, social media might be additionally limiting for detecting surprises (see section VIII A), but after writing this discussion paper we are hesitating if the source is really of importance for missing surprises or if other factors show to be more important.

As research at weak signals seems to be a very vivid part of current innovation research and futurology, deeper investigations on the theoretically described characteristic of "surprise" should be taken into account.

ACKNOWLEDGMENT

Thanks to Kathrin Parson and Laurenz Giger for their support.

REFERENCES

[1] I. H. Ansoff, Implanting Strategic Management, Prentice/Hall International Inc, 1984.

[2] M. Markus, R. Eckhoff, and M. Lassnig, “Innovation Signals in Online-Communitys – ein komplementärer analytischer Ansatz”, in: M. Blattner, and A. Meier, Eds., Web Monitoring. HMD - Praxis der Wirtschaftsinformatik, 293, 10/2013, pp. 13–21.

[3] M. Lassnig, M. Markus, R. Eckhoff, and K. Wrussnig, “Prospects of technology-enhanced Social Media Analysis for Open Innovation in the Leisure Industries”, in: R. Egger, I. Gula, and D. Walcher, Eds., Open Tourism – Open Innovation, Crowdsourcing and Collaborative Consumption challenging the Tourism Industry. Salzburg, 2013.

[4] T. Kuosa, “Futures signals sense-making framework (FSSF): A start-up tool to analyse and categorise weak signals, wild cards, drivers, trends and other types of information”. In: Futures, Volume 42, Issue 1, February 2010, pp. 42–48.

[5] E. Hiltunen, Weak Signals in Organisational Futures, Aalto University School of Economics, Aalto 2010.

[6] R. Eckhoff, M. Markus, M. Lassnig, and S. Schön, „Detecting Weak Signals with Technologies. Overview of current technology-enhanced approaches for the detection of weak signals“,International Journal of Trends in Economics Management & Technology (IJTEMT), volume III issue V, October 2014.

[7] K. Welz, L. Brecht, J. Kauffeldt, and D. Schallmo, “Weak signals detection: Criteria for social media monitoring tools”, in: Proceedings of the 5th ISPIM Innovation Symposium: "Stimulating Innovation: Challenges for Management, Science & Technology", 09-12 December, 2012; Seoul, Korea.

[8] B. Coffman, “Weak Signal Research”, MG Taylor Cooperation, online available http://www.mgtaylor.com/mgtaylor/jotm/winter97/wsrintro.htm.htm, 1997, downloaded at 31.08.2014.

[9] L. Ilmola and O. Kuusi, “Filters of weak signals hinder foresight: Monitoring weak signals efficiently in corporate decision-making”, Futures, Volume 38, Issue 8, 2006, pp. 908–924.

[10] T. Kuosa, Toward Strategic Intelligence. Foresight, Intelligence and Policy Making, Helsinki, Dynamic Futures 2014.

[11] I. H. Ansoff, “Managing strategic surprise by response to weak signals”, California Management Review XVIII 2, 1975.

[12] R. Zafarani, M. Abbasi, and H. Liu, Social Media Mining. An Introduction, 2014, Cambridge University Press.

[13] R. Eckhoff, M. Markus, M. Lassnig, and S. Schön, “Guidelines for social media mining for innovation purposes. Experiences and recommendations from literature and practice”, Proceedings of eKNOW 2015, The Seventh International Conference on Information, Process, and Knowledge Management, February 22 - 27, 2015 - Lisbon, Portugal.

[14] R. Rohrbeck, N. Thom, and H. Arnold, IT tools for foresight: “The integrated insight and response system of Deutsche Telekom Innovation Laboratories”, Technological Forecasting and Social Change, September 28, 2013.

[15] S. Kim et al., „NEST: A quantitative model for detecting emerging trends using a global monitoring expert network and Bayesian network“, Futures, Volume 52, August 2013, pp. 59–73.

[16] J. Keller and H. A. von der Gracht, “The influence of information and communication technology (ICT) on future foresight processes — Results from a Delphi survey”, Technological Forecasting and Social Change, Volume 85, June 2014, pp. 81–92.

[17] L. Ilmola, and O. Kuusi, “Filters of weak signals hinder foresight: Monitoring weak signals efficiently in corporate decision-making”, Futures 38 (8) (2006), pp. 908–924.

[18] B. Fischhoff and R. Beyth “'I knew it would happen': Remembered probabilities of once-future things.” Organizational Behaviour and Human Performance, 13 (1975), 1-16.

[19] S. Schaffert and C. Schwalbe, “Future Media Adoption in Learning and Teaching: Current Study Design from the Perspective of Cultural Studies”, in: M. Ebner & M. Schiefner (eds.). Looking Toward the Future of Technology Enhanced Education: Ubiquitous Learning and the Digital Native. Hershey: IGI Global, 2009, pp. 1-11.

[20] T. Meyer, “Zwischen Kanal und Lebens-Mittel: pädagogisches Medium und mediologisches Milieu”, in: J. Fromme and W. Sesink (eds.), Pädagogische Medientheorie. Wiesbaden: VS Verlag für Sozialwissenschaften, 2008, pp. 71-94.

[21] K. Steinmüller, “Wild Cards, Schwache Signale und Web-Seismografen. Vom Umgang der Zukunftsforschung mit dem Unvorhersagbaren”, in: Wolfgang J. Koschnick (eds.), FOCUS-Jahrbuch 2012. Prognosen, Trend- und Zukunftsforschung, Focus Verlag München, 2012, pp. 215-240.

Social Networks: Privacy Issues and Precautions

Mohamad Ibrahim Ladan
Computer Science Department
Haigazian University
Beirut – Lebanon
mladan@haigazian.edu.lb

***Abstract*—Social networks, such as Facebook, Myspace, LinkedIn, Google+, and Twitter have experienced exponential growth and a remarkable adoption rate in recent years. These social networks are touching our lives at home and at work by providing attractive means of online social interactions and communications with family, friends and colleagues from around the corner or across the globe; however, this comes with a growing concern regarding the privacy and security risks that accompany the use of such networks. In this paper, we will investigate and discuss the different privacy issues pertaining to social networks, in addition, we will propose some precaution measures that should be applied to tackle these issues.**

Keywords: Social Networks privacy issues; online privacy; information revelation; Social Networks privacy precautions measures.

I. Introduction

The size, growth adoption rate, and popularity of social media networks, such as Facebook, Myspace, LinkedIn, and Google+, are phenomenal. Facebook has reached more than 700 million users and according to a brochure released by Websense, a company specializing in computer security software, Facebook has an annual growth rate of 41% and Twitter is growing at 85% year after year. In 2011 Google released its Google+ social networking offering, first by invitation only and then generally opening the site [1]. In 2014, the largest social network is Facebook and other popular networks include Twitter, Instagram, LinkedIn, and Pinterest. [22]. This fast growth is due in part to the latest advancements in the field information and communication technologies, tablets, mobile smart phones, and other similar mobile computing and communication devices that have become very popular and sometimes necessary home and individual type of appliances to both kids and adults. These types of networks are useful and have a lot of benefits to all kind of users. They were built upon the concept of traditional social networks where you are connected to new people through people you already know. Their goal could be purely social, allowing users to establish friendships or romantic relationships, while others may focus on establishing business connections. They can be used for professional networking and job searches, as a means to increase sales revenue, as a way to reconnect with current and old friends, as a way to make new friends, or as a way to share information and to socialize.

Social networks can be described as web applications that allow users to create their semi-public profile [13], i.e., a profile that some information is public and some is private, communicate with friends, and build an online community. It is based on social relationships among users. Most people join social networks to share their information and keep in contact with people they know. The main feature of social networks is a friend finder that allows social network users to search for people that they know and then build up their own online community. They have changed radically the way people interact with each other regardless of their physical location. They provides every person, regardless of its age, the ability to easily communicate and share data and information of all types, audio, video, or text on one-to-one basis, one-to-many, and many-to-many in a matter of fraction of seconds without any difficulties. These benefits are accompanied with a growing concerns regarding the privacy of the information exchanged or stored on the communication links and servers of those social networks. These information could be sensitive or critical, such as the identification, confidential conversation, personal, and private data, and credit and financial data. So every user of social networks should be concerned one way or the other about these types of privacy risks. Most people became more aware and more concerned about these risks when Facebook inadvertently exposed millions of users' phone numbers and e-mail addresses to unauthorized viewers over years that began in 2012. The major reason for causing such breaches in social network security and privacy emerges from the massive amount of information that these sites process every day, making it much easier to exploit even if there is a little fault in the system. Furthermore, social networking is now no longer restricted to just desktop and laptop computers, the technology is available on smartphones, tablets, and just about anything that is connected to the Internet [4]. This wide spread of smart mobile devices has opened new paths for malware transmission, bringing concerns about information theft, and tracking user's location and preferences.

Despite the fact that most of the privacy risks in using social networks are related to the Information Communication Technology and the internet infrastructure that these social networks are operating on. However, most privacy risks come from the nature of the social network application itself and sometimes from their privacy policies and the way people use the networks. This paper is focused on what we call user-related part of these risks, i.e., risks that are stemmed from user's behavior, privacy policies, and the nature of the application and its default configuration. The application-related part tends to be more technical, i.e., data and communication security, protocols used, enforcing privacy setting, security measures, encryption and decryption will be addressed in another paper.

The rest of this paper is organized as follows: In Section II, the various social networks privacy issues and concerns are presented and discussed. In Section III, different precaution measures and guidelines to maintain an adequate level of privacy on social networks are proposed and presented. In Section IV, discussion of the need for security and privacy legislation and policies to tackle the privacy concerns are discussed. Finally, in Section V, a summary and conclusion of the paper is given.

II. SOCIAL NETWORKS PRIVACY ISSUES

“If you feel like someone is watching you, you're right. If you're worried about this, you have plenty of company. If you're not doing anything about this anxiety, you’re just like almost everyone else.” [15]

The increased pervasiveness and use of information communication technologies have changed many people's lives in terms of how they work, form, and maintain social relations. This rise in social networks or networked societies comes with a lot of concerns, mainly the privacy concern. Social networks usually share three common elements. They allow individuals to construct a public or semi-public profile within a confined system, articulate a list of other users with whom they share a connection, and view and traverse their list of connections and those made by others within the system. While the concept of privacy is not new, modern technological advancements have meant that privacy concerns have evolved. New information communication technologies have transformed our ability to collect, aggregate, and share data. Modern technology has the ability and power to capture, store, aggregate, redistribute, and use data from individual users. The problem is that the owner of this information is often unaware of, or at least unconnected to, its storage and utilization, and that such ubiquitous data collection is harmful to personal privacy [23].

The "State of the Net" research and statistics from Consumer Reports suggest that there is an overall increase in certain digital problems, such as ID thefts, phishing schemes, and security breaches. The most surprising findings however, involve how much Facebook knows about more than 900 million members, and how much we, members, freely offer information that could be extracted by employers, insurers, some government offices, as well as identity thieves and other criminals [17].

Most if not all social network user’s profiles contain real information about users. Sensitive information, such as user’s full name, contact information, relationship status, date of birth, previous and current work, and education background attract hackers. In addition, most social network users share a large amount of their other private information in their social network space, and publish the information publicly without careful consideration. Hence, social networks have become a large pool of sensitive data. Moreover, social network users tend to have a high level of trust toward other social network users. They tend to accept friend requests easily, and trust items that friends send to them.

Most social networks ask users to agree to Terms of Use policy before they can use their services. However, these Terms of Use often contain phrases permitting social networks to store user’s data on their servers and even share it with third parties. The levels of privacy presented for users in social networks vary from one network to the other. Some encourage users to provide real names and other personal information, such as age, family, education, interests, and even relationship status. Facebook has attracted attention over its policies regarding data storage, such as making it difficult to delete an account, holding onto data after an account is de-activated and being caught sharing personal data with third parties [12].

Privacy can be viewed from the perspective of control. Whether it is control over personal data, the choice to disclose data, the physical presence of others, the number of others present in disclosure, or choosing which person to discuss and share issues with, control is central to maintaining privacy. The main privacy issue in social networks is the abuse and the leakage of profile and personal information of the users. Several cases related to privacy issues have surfaced up lately. A report in the *Wall Street Journal* indicates that the Facebook, along with MySpace, and a handful of other social networks, have been sharing users' personal data with advertisers without users' knowledge or consent [6]. The data shared includes names, user IDs, and other information sufficient to enable ad companies, such as the Google-owned DoubleClick to identify distinct user profiles. Moreover, Facebook appears to have gone farther than the other networks when it comes to sharing data. When Facebook's users clicked on ads appearing on a profile page, the site would at times provide data, such as the username behind the click, as well as the user whose profile page from which the click came. In addition, Twitter has admitted that they have scanned and imported their user's phone contacts onto the website database so that they can learn more about their users. Most users were unaware that Twitter is created this way for new users to search for their friends. More than 1,000 companies are waiting in line to get access to millions of tweets from users that are using the popular social networking website. Companies believe that by using data mining technologies they would be able to gather important information that can be used for marketing and advertising [11]. Twitter has stated that they will have their privacy guidelines illustrated more clearly in the future [10].

Although some users may have no problem in revealing their personal information to a large group of people and may not care about the privacy policies and setting of the network, others they care and try to make use of the available security features. However most social networks restore the relaxed default setting after each update. Facebook was criticized due to the perceived carelessness regarding privacy in the default setting for users [8].

Other main issue related to privacy is stemmed in the fact that many social networks provide an Application Programming Interface (API) for 3rd party developers to create applications for the network’s platform. These 3rd party applications are very popular among social network users, and once installed, they are able to access user’s data automatically, and are capable of posting on users’ space or user’s friend’s space, or may access other user’s information without user’s knowledge. In addition, these 3rd party applications are able to track social network user’s activities, or allows advertisement partner to access and collect social network user’s data for commercial and advertising purposes [14].

On the other hand, posting and sharing, directly or indirectly, photos or videos may result in an individual's breach of privacy or to an organization's breach of confidentiality.

Another important issue related to privacy is that potential employers are looking up information about their potential employees using social networks. The information found is used to screen job applicants and may affect or hurt their chances of being employed. Employers may find out that an applicant made a political statement that conflicts with the company ideologies. Facebook and Twitter are being used a lot to screen job applicants. On Facebook and Twitter, we believe employers are trying to get a more personal view of a candidate, rather than the resume-like view they will see on LinkedIn. This means that is important for social networks users to keep in mind that some, if not most, employers use these networks to do some sort of pre-screening their job applicants [16].

Many social networks have responded to criticism and concerns over privacy. It is claimed that changes to default settings, the storage of data and sharing with third parties have all been updated and corrected in the light of criticism, and/or legal challenges. However, many critics remain unsatisfied, noting that fundamental changes to privacy settings in many social networking sites remain minor [9].

III. SOCIAL NETWORKS PRECAUTIONS

The popularity of social networks continues to increase, especially among teenagers and young adults, and the concerns related to privacy risks and issues continue to increase. Fig. 1 shows the worldwide number of active Facebook users from 2008 to 2014. As of the third quarter of 2014, Facebook had 1.35 billion monthly active users [24].

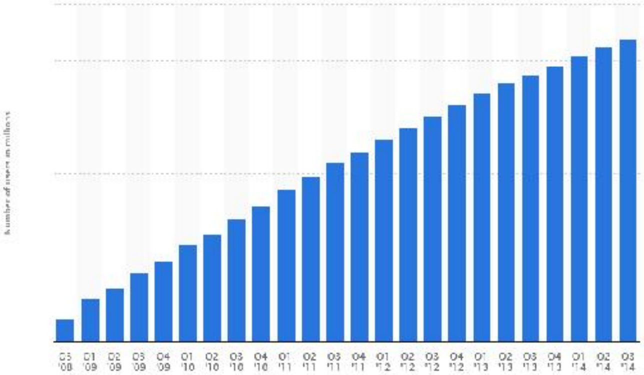


Figure 1: Number of monthly active Facebook users worldwide from 3rd quarter 2008 to 3rd quarter 2014 (in millions)

Therefore, privacy issues become more and more critical and some precautions should be taken to address these issues while using social networks. These precautions can be stated as a set of guidelines and helpful tips that social network users should follow to protect their private information while making use of the benefits of the networks. These guidelines and helpful tips include the following:

- Review the social network’s privacy policy before signing up. If the privacy policy is not clear on how it protects the member’s information, do not sign up or limit your use of such a network.
- Choose a strong password that cannot easily be guessed and different than other passwords you have on different systems or social networks.
- Check and configure the privacy settings. The default settings for some social networks may allow anyone to see your information, these setting

should be changed to allow only those people you trust to have access to the information you post.

- Remember that social networks are based on the internet which is a public resource.
- Keep in mind that once information is posted online, it can possibly be viewed by anyone and may not be withdrawn after that even if you delete the information from your account, cached or saved copies may still exist on other computers on the network. Therefore, confidential information should not be posted or shared. You should only post information you are comfortable revealing to a complete stranger.
- Limit the amount of personal information you post. Do not post information that would make you vulnerable, such as your address or information about your daily schedule or routine.
- Do not post information like your address, mobile phone number or any information that could be used for banking site security questions, such as your mother's maiden name, hometown, favorite car, school name etc. Identity thieves can find out a lot about you just by the information you may already have on your profile, and they can fill in the blanks.
- Do not tag your location and whereabouts, and do not announce that you are on vacation or away for an extended period of time. You don't want to compromise your feeling of safety and security if someone may know you're not home. It also opens up opportunities for cyber stalkers.
- Be careful about installing third-party applications. Some social networks provide the ability to install third party applications, such as games or other entertainment functionality, however some of these applications may be malicious and may have full access to your account and the data you share. In addition, some of these applications may modify your security and privacy settings. Hence, do not install applications unless they come from trusted, well-known sources.
- Don't believe everything you read online. People may post false or misleading information about their own identities. The internet makes it easy for people to misrepresent their identities and goals.
- Limit the people who are allowed to contact you on social networks. If you interact with people you do not know, be cautious about the amount of information you disclose.
- Think twice before clicking a link to another page or running an online application, even if it is from someone you know. Many applications require you to share your information when you use them. Attackers use these sites to distribute their malware.
- Be careful when adding new friends. A "Friend" is anyone on the Facebook network whom you permit to see your personal information, such as birth date, photos, job, comments and list of other Friends. Friends can also see Friends of Friends, which means that you have possibly added strange individuals whom you may never have met as your active%20users%202008individuals have access to your personal and private information.
- Limit the number of your friends. The more friends you have the more people who have access to your information and the more vulnerable your account is.
- Teach children about internet safety and be aware of their online habits. Children are more susceptible to the threats resulted from the use of social networks. Although many of these networks have age restrictions, children may misrepresent their ages so that they can join.
- If you are working in a company and you often communicate with your colleagues using social networks then talk to the manager to put a social Network Use Policy in place for your company to protect the privacy of information exchanged using these networks.

IV. DISCUSSION

Although more people and companies are finding new ways in using social networks, their successes have been faced by major concerns of privacy risks. The control of information or data is lost once it is posted to a social network. Despite the fact the posted info is meant to only go to selected friends may seem protected by the limited distribution to a restricted audience, nothing prevents one of those friends from forwarding that post to someone else outside of the original poster's group of friends [18]. The same can occur within a group of employees collaborating on a project, it only takes one person to become a leak and forward information to outside the group. And if privacy permissions on a social media site are not set correctly, the data may leak out and become public by default. This can easily happen with FaceBook's privacy policy that keeps changing from time to time. As a matter of fact, as the popularity of social networks continues to grow, the concerns over privacy risks and privacy protection becomes more important and more serious for the users. On the other hand, users are having more concerns about not being able to delete permanently their data they have on the social networks. As humans, we can forget. But the Internet never forgets. And once that data is released, there is no getting rid of it [1]. Viviane Reding, the Vice President of the European Commission said: “God forgives and forgets, but the Internet never does” [2]. Privacy advocates are working to change this problem by introducing a “right to be forgotten”. This is being proposed in a new draft of the European Data Protection Directive that “measures will be put in place to allow European citizens’ to have their data deleted by private companies” [3].

Furthermore, as social networks continue to take advantage of mobile devices and location-based services, users will be exposed to even more privacy concerns. Although users will unquestionably enjoy using some of these services, they could possibly be making themselves exposed to more serious privacy risks. Hence, social networks that employ location-based services will have to focus on user privacy concerns to gain people trust.

As people increase the amount of information they share on social networks, some of these giant social networks will have and store a huge amount of personal information about their users. Hence, the need for more privacy controls increases. In addition, without good universal guidelines and overall legislation and privacy laws on how this information can be gathered and used, it could be misused, either intentionally or unintentionally. As a matter of fact, there has been quite a bit of controversy over how much data social networks, such as Facebook and Twitter collect. Facebook lately had a scandal where it used people’s profile information to post ads on Facebook that appeared to be authorized by the person whose profile was pulled. It had their profile picture and text that stated that one should buy this product or use such and such a service. There was a lawsuit and millions of people received an e-mail informing them about this lawsuit and how they could take action and possibly receive a settlement of money. Facebook lost millions of dollars because of this blooper [19].

The privacy legislation process has already started developing in the USA and other part of the world, and it’s likely to gain even more momentum in years to come and that there will be additional towards universal legislation on privacy regulations. An instance of the new privacy legislation process that is taking place in the USA is the Commercial Privacy Bill of Rights Act of 2011 (“CPBR”). It represents one approach to protecting consumer privacy, and it aims to establish a regulatory framework for comprehensive protection of personal data for individuals under the protection of the Federal Trade Commission. The CPBR would require companies that collect consumer data to adhere to certain security practices and would also require consumers to opt-in to the collection of sensitive information. Consumers could also access, correct, and control information that companies have stored. In addition, the bill would limit the data that a company could collect during any given transaction to only data that is necessary for the transaction’s completion. For instance, an online clothes store could not require the consumer to provide personal information, such as his or her birthday if such information is peripheral to the consumer’s purchase of snow suits [20, 21].

Without universal privacy regulation and legislation social networks have been setting their own privacy policies, and there is currently an enormous amount of variation between networks. As a result of that, users are often confused as to what privacy controls are available and how they should be used. Additionally, most people do not really understand how to recognize the potential for information misuse. People often share information innocently because they want to use a specific feature, or because they wish to qualify for a free product or service. When universal legislation is in place, social networks will have standard guidelines and policies to follow, thus creating a more secure, safe, private, and less confusing user experience [5].

V. SUMMARY AND CONCLUSION

With the constantly growing popularity of social networks, such as Facebook, Google+, MySpace, Twitter etc. in the personal scope, and others, such as LinkedIn in business circles, undesirable privacy risk issues have arisen as a result of this unexpected rapid rise and due to the availability of huge amount of sensitive information related to large number of users. Therefore, concerns related to privacy issues and breach of privacy in social networks that can put the individual or a company in a serious risk are increasing.

A privacy issue occurs, in its simplest form, when someone, who may be a hacker or not, gain access to private and confidential information about users who are not careful about what they expose on their Social network accounts. In addition, the potential damage to a user as a result of privacy breach depends on how much this user is actively participating or engaging in the social networks, as well as the amount of information he or she is posting. The more information in general, and private information in particular, is posted the higher the risk and harm.

Moreover, privacy issues in social networks, other than those rising from security issues, are more related to users behaviors and awareness of privacy policies and terms and conditions for using these networks. The more information a user posts, the more information becomes available for a potential misuse by malicious users/hackers. Users who provide private, sensitive, and confidential information about themselves and their friends will be more vulnerable, themselves and their friends, for being attacked or hacked. Information, such as a person's social security number, street address, phone number, financial information, or confidential business information should not be posted and shared online. A well-informed user will not only help to maintain privacy, but will also educate others on these issues. The best solution to social network privacy issues is to limit the amount of shared and posted information.

Social networks developers try to implement different mechanisms and measures to protect their users’ information and data, but attackers will always find new methods to break through those measures. Therefore, social network users should be aware of all these threats, and be more careful when using such networks and to limit the amount of shared and posted information. In addition, until universal privacy regulation and legislation are developed and enforced, social networks are setting their own privacy policies that sometime, intentionally or unintentionally, do not protect the privacy of users date and personal information.

This paper addressed, discussed and presented the different types of privacy issues and risks arising from the use of social networks. In addition, it summarized and presented different privacy precautions tips, measures, and helpful guidelines to be followed to protect the user's private information while making use of the benefits of social networks.

REFERENCES

[1] Shullich, R. Risk Assessment of Social Media: http://www.sans.org/reading-room/whitepapers/privacy/risk-assessment-social-media-33940. Dec. 5, 2011

[2] Berwaerts, P. The right to be Forgotten: http://www.business2community.com/government-politics/the-right-to-be-forgotten-0111815. Dec. 28, 2011

[3] Whittaker, Z. European data protection law proposals revealed: http://www.zdnet.com/blog/london/european-dat-protection-law-proposals-revealed/1365. Dec. 7, 2011

[4] V. Jain, InformationWeek, May 19, 2014

[5] Top Five Social Media Privacy Concerns 2014. http://www.reputation.com/reputationwatch/articles/top-five-social-media-privacy-concerns, 2014.

[6] E. Bangeman, http://arstechnica.com/tech-policy/news/2010/05/latest-facebook-blunder-secret-data-sharing-with-advertisers.ars, May 21 2010

[7] Staying Safe on Social Network Sites, http://www.us-cert.gov/ncas/tips/ST06-003, Feb. 06, 2013.

[8] Kelly, S. Identity 'at risk' on Facebook. BBC News. http://news.bbc.co.uk/1/hi/programmes/click_online/7375772.stm

[9] N. Saint, Facebook's Response to Privacy Concerns: "If you're not Comfortable Sharing, Don't". http://www.businessinsider.com/facebooks-response-to-privacy-concerns-if-youre-not-comfortable-sharing-dont-2010-5, 2010.

[10] Sky news, "Twitter admits peeking at address books, announces privacy improvements". Feb. 16, 2012.

[11] K. Gladdis, "Twitter secrets for sale: Privacy row as every tweet for last two years is bought up by data firm". London: daily mail. Feb. 28, 2012.

[12] Bangeman, E. Report: Facebook caught sharing secret data with advisers. http://arstechnica.com/tech-policy/news/2010/05/latest-facebook-blunder-secret-data-sharing-with-advertisers.ars, 2010.

[13] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," J. Computer-Mediated Communication, vol. 13, no. 1, pp. 210–30. http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html, Oct. 2007.

[14] Online Social Networks," WOSN '08 Proceedings of the first workshop on Online social networks, pp. 37-42. http://www2.research.att.com/~bala/papers/posn.pdf, 2008

[15] Sullivan, B. Social Media Polarizes Our Privacy Concerns. Facebook And Its Competitors Are Challenging Long-Held Perceptions of Privacy. http://www.msnbc.msn.com/id/41995992/ns/technology_and_science/t/study-social-media-polarizes-our-privacy-concerns/#.UMEDzYNtjjU. October 3, 2011

[16] How Employers Use Social Media to Screen Applicants, INFOGRAPHIC.http://theundercoverrecruiter.com/infographic-how-recruiters-use-social-media-screen-applicants/

[17] Rosa Golijan, Consumer Reports: Facebook privacy problems are on the rise, NBC News, http://www.nbcnews.com/technology/technolog/consumer-reports-facebook-privacy-problems-are-rise-749990. 2012.

[18] Associate press, http://www.foxnews.com/us/2011/11/08/judge-rules-teacher-should-lose-job-after-facebook-post/, 2011.

[19] Smith, M.; Szongott, C.; Henne, B.; von Voigt, G.; , "Big data privacy issues in public social media," Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on ,vol., no., pp.1-6, 18-20 June 2012, doi: 10.1109/DEST.2012.6227909

[20] TO TRACK OR NOT TO TRACK: RECENT LEGISLATIVE PROPOSALS TO PROTECT CONSUMER PRIVACY, Harvard Journal on Legislation, Vol 49, 2012.

[21] Danah Boyd & Ezster Hargittai, Facebook Privacy Settings: Who Cares?, FIRST MONDAY. http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3086/2589. Aug. 2, 2010

[22] The U.S. Digital Consumer Report". Featured Insights, Global, Media + Entertainment. Nielsen. Retrieved 25 November 2014.

[23] D. J. Houghtona and A. N. Joinsona, Privacy, Social Network Sites, and Social Relations, Journal of Technology in Human Services, Volume 28, pages 74-94, Issue 1-2, 2010. DOI:10.1080/15228831003770775

[24] Statistica 2014, http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/, access date Dec., 22, 2014.

Risk Assessment Quantification of Ambient Service

Fundamental investigation of the risks of cyber-physical systems

Shigeaki Tanimoto
Faculty of Social Systems Science
Chiba Institute of Technology
Chiba, Japan
shigeaki.tanimoto@it-chiba.ac.jp

Hiroyuki Sato
Information Technology Center
The University of Tokyo
Tokyo, Japan
schuko@satolab.itc.u-tokyo.ac.jp

Atsushi Kanai
Faculty of Science and Engineering
Hosei University
Tokyo, Japan
yoikana@hosei.ac.jp

***Abstract*—Ambient services have attracted attention as a possible ubiquitous, future intelligent infrastructure. An ambient service automatically provides services suited to the user by making sensors and computers cooperate and by gathering and analyzing information about each user. As such, ambient services are related to cyber-physical systems. However, in the process of managing personal information, ambient services are prone to various risks, such as information leakage. Our previous study analyzed the service provision and service use sides. It used the risk breakdown structure (RBS) and risk matrix, which are typical risk management methods of project management, and identified 40 risk factors faced by ambient services and countermeasures thereof. However, we recognized that it was only a qualitative study and that a quantitative evaluation would be needed to make its countermeasures more practical. Hence, in this paper, the risk factors identified in the previous study are analyzed and quantitatively evaluated. Specifically, the values of the risk factors were calculated by using a risk formula used in the field of information security management systems (ISMS). On the basis of these values, the effect of the countermeasures proposed in the previous study was evaluated quantitatively. It was found that the countermeasures in the previous study could reduce their corresponding risk factors by 18% - 36%. The results herein can be used to promote ambient services in the future.**

Keywords- Ambient Service; Cyber-physical System; Risk Assessment; Risk Value Formula; ISMS

I. INTRODUCTION

Ambient services, which use sensors or wireless-communications technology, are now attracting attention [1]. Ambient services offer the possibility of creating a new information society, as follows:

- Computers can be used to gather information from sensors and monitor the user's situation.
- Personal data can be accumulated and analyzed in order to provide services meeting the user’s specific needs.

There are various merits of being able to provide services friendly enough to bridge the digital divide (e.g., to help elderly people unfamiliar with intelligent terminals) through cooperative functioning of computers and sensors [2]-[4]. As such, ambient services are related to cyber-physical systems.

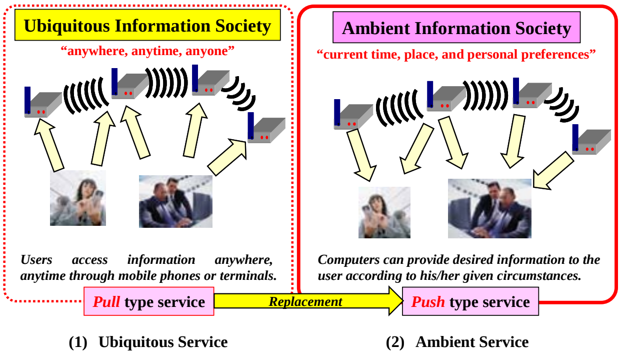
However, an ambient service requires a user's personal information beforehand. Accordingly, there are risks such as leakage of personal information. In fact, leaks could reveal, for example, not only the user’s name, address, and names of other family members, but also his or her current position. Thus, confidentiality of personal information must be guaranteed to ensure that the ambient information society is safe and secure. In this regard, it is important to perform a risk assessment on an ambient service and to take countermeasures in advance against risks. In our previous study, we did a risk assessment of ambient services [5]. In particular, we used the risk breakdown structure (RBS) method to identify risk factors and the risk matrix method to analyze these factors [6]-[7]. We also drew up countermeasures to the identified risks. However, it was only a qualitative study, meaning that a more practical quantitative evaluation still needed to be undertaken.

In this paper, we describe a quantitative evaluation of the risk factors of ambient services obtained in our previous study and the proposed countermeasures. Specifically, a risk value based on the formula is calculated for each risk factor [8]-[10]. Then, on the basis of this value, the effect of the countermeasures on the risks can be quantitatively evaluated. It is shown that the countermeasures in the previous study can reduce their corresponding risk factors by 18% - 36%. We believe that the results of this study will help to promote ambient services.

Section 2 reviews the various ambient services that have been studied so far. In section 3, we describe our previous study and the present problem. Section 4 describes the quantitative evaluation of ambient service's risks. Section 5 discusses related work, and section 6 is a conclusion and describes future work.

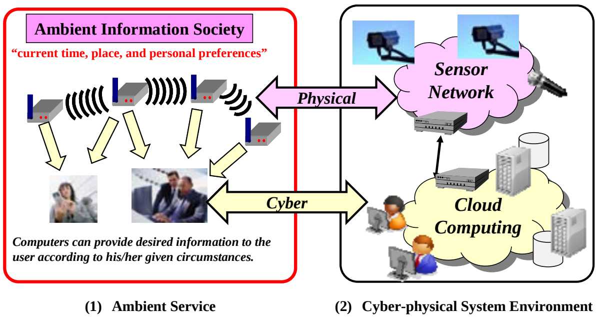
II. AMBIENT SERVICES

In 1998, Eli Zelkha and Brian Epstein of Palo Alto Ventures in the U.S. crafted a presentation on the concept of ambient intelligence in which the future of consumer electronics, telecommunications, and computing is called the "ambient society" [11]. Since then, the idea of ambient services has attracted the attention of researchers as a potential next-generation digitized infrastructure that could replace the ubiquitous information society [9]. For example, the IT strategy of Japan has been transitioning from one of “u-japan” to “i-japan” [12]. Here, “u-japan” refers to a ubiquitous net society, whereas “i-japan” means a movement toward digital inclusion and innovation. The distinction between u-japan and i-japan is depicted in Figs. 1(1) and (2) [2] [12]. Moreover, as shown in Fig. 2, ambient services are also related to cyber-physical systems [13]-[14].


Ubiquitous Information Society
“anywhere, anytime, anyone”
Users access information anywhere, anytime through mobile phones or terminals.
Pull type service
Replacement
Ambient Information Society
“current time, place, and personal preferences”
Computers can provide desired information to the user according to his/her given circumstances.
Push type service

(1) Ubiquitous Service (2) Ambient Service

Figure 1. Transition from Ubiquitous Services to Ambient Servicese


Ambient Information Society
“current time, place, and personal preferences”
Computers can provide desired information to the user according to his/her given circumstances.
Physical
Cyber
Sensor Network
Cloud Computing

(1) Ambient Service (2) Cyber-physical System Environment

Figure 2. Relation between Ambient Services and Cyber Pysical Systems

III. PREVIOUS STUDY: RISK FACTORS AND COUNTERMEASURES OF AMBIENT SERVICES

A. Risk factors of ambient services

Ambient services for a future information society face many problems that could hamper their spread. In the present ubiquitous information society, leaks of personal information due to nefarious schemes or even simple mistakes are a serious problem. Similar problems are of concern in an ambient service. In particular, there are various points of concern that arise in the aspects of privacy protection, disclosure of service content, etc.

In our previous study [5], we employed the risk breakdown structure (RBS) method [6], a typical risk management method for project management, to identify risk factors in ambient services. The results are shown in Table 1. As can be seen, the risk factors were identified from a comprehensive range of viewpoints. A total of 40 risk factors were extracted by the RBS analysis.

TABLE I. RISK FACTORS EXTRACTED BY RBS IN SECURITY PERCEPTION PROBLEM

High division	**Middle division**	**Low division**	**Risk Factor**
1. Service provision side	1.1 System	1.1.1 Software	1.1.1.1 Problem in cooperating with the existing system
			1.1.1.2 Problem with ending Ambient Service
			1.1.1.3 Problem with service entrepreneur's specifications
			1.1.1.4 Problem with service entrepreneur's supervisor
			1.1.1.5 Leaks, etc., by service entrepreneur
			1.1.1.6 Data deleted at end of service use
			1.1.1.7 Problem with requirements for certification
			1.1.1.8 Problem in managing personal information
			1.1.1.9 Data seized by other company
			1.1.1.10 No restoration of missing data
			1.1.1.11 No security management
			1.1.1.12 Leakage and disappearance of data
			1.1.1.13 Lack of internal control or security audit
		1.1.2 Hardware	1.1.2.1 Portability problem with existing hardware
		1.1.3 Network	1.1.3.1 Problem with fulfilling SLA
			1.1.3.2 Insufficient right-to-access management
	1.2 Operation	1.2.1 Information control	1.2.1.1 Insufficient information disclosure by service entrepreneur
			1.2.1.2 Problem with different service specifications and user requirements
			1.2.1.3 Crisis regarding continuation of service
			1.2.1.4 Business continuation plan is insufficient
		1.2.2 Rule	1.2.2.1 Compliance violation
	1.3 Facility	1.3.1 Facility, Equipment	1.3.1.1 Power failure due to increased power consumption
			1.3.1.2 Environmental impacts such as carbon dioxide emissions
			1.3.1.3 Influence of delay or communication failure in real-time distribution
			1.3.1.4 Equipment installation problems.
2. Service use side	2.1 System	2.1.1 Software	2.1.1.1 Complication of operations
			2.1.1.2 Improper management of personal information
		2.1.2 Hardware	2.1.2.1 Portability problem with existing terminal
		2.1.3 Network	2.1.3.1 Problem with security of right to access
			2.1.3.2 Problem with safety of encryption
	2.2 Operation	2.2.1 Personal information	2.2.1.1 Problem in handling personal information
			2.2.1.2 Deletion of personal information
			2.2.1.3 User's incorrect deletion, alteration, etc.
			2.2.1.4 General information disclosure
		2.2.2 Certification	2.2.2.1 Problem with access except for a user
	2.3 Facility	2.3.1 Facility, Equipment	2.3.1.1 Breakage of device due to consumption
			2.3.1.2 Communication failure at base station
3. Other external factors	3.1 Law	3.1.1 Regulation problem arising from revision of law	
	3.2 Disaster	3.2.1 Data center collapses in a disaster	
		3.2.2 Problem compensating user for personal information disclosure, etc.	

25 risk factors

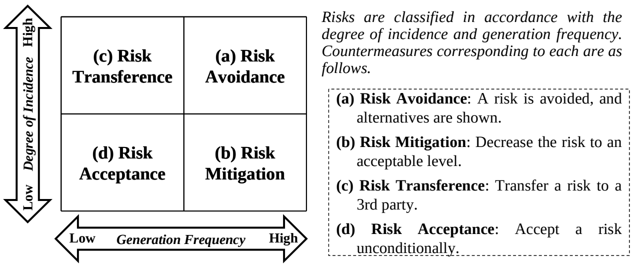
12 risk factors

3 risk factors

40 risk factors

B. Proposed countermeasures against risk factors

Next, we devised potential countermeasures against the identified risks; these are shown in Table 2. The risk matrix method was used to deduce these countermeasures [7]. As shown in Fig. 3, this method classifies risks into four kinds, *i.e., Risk Transference*, *Risk Mitigation*, *Risk Acceptance*, and *Risk Avoidance*, in accordance with their generation frequency and degree of incidence. Furthermore, it gives guidelines to draw up countermeasures. Table 2 lists the classification of the risk matrix methods in correspondence with its proposed countermeasures.



Risks are classified in accordance with the degree of incidence and generation frequency. Countermeasures corresponding to each are as follows.

(a) Risk Avoidance: A risk is avoided, and alternatives are shown.

(b) Risk Mitigation: Decrease the risk to an acceptable level.

(c) Risk Transference: Transfer a risk to a 3rd party.

(d) Risk Acceptance: Accept a risk unconditionally.

Figure 3. Risk Matrix Method

TABLE II. RISK FACTORS EXTRACTED BY RBS AND PROPOSED COUNTERMEASURES

Level 3: Risk Factors	Degree of Influence	Generation Frequency	Countermeasure Classification	Proposed countermeasures
1.1.1.1 Problem in cooperating with the existing system	High	High	Risk Avoidance	Adjustment on the use side
1.1.1.2 Problem with ending Ambient Service	High	Low	Risk Transference	Third-party surveillance
1.1.1.3 Problem with service entrepreneur's specifications	High	High	Risk Avoidance	Adjustment on the use side
1.1.1.4 Problem with service entrepreneur's supervisor	High	Low	Risk Transference	Third-party surveillance
1.1.1.5 Leaks, etc., by service entrepreneur	High	Low	Risk Transference	Application of assurance
1.1.1.6 Data deleted at end of service use	Low	High	Risk Mitigation	User complies with the specification by using the Ambient service.
1.1.1.7 Problem with requirements for certification	Low	High	Risk Mitigation	User complies with the specification by using the Ambient service.
1.1.1.8 Problem in managing personal information	Low	High	Risk Mitigation	User complies with the specification by using the Ambient service.
1.1.1.9 Data seized by other company	High	Low	Risk Transference	Application of assurance
1.1.1.10 No restoration of missing data	Low	Low	Risk Acceptance	Others
1.1.1.11 No security management	High	Low	Risk Transference	Application of assurance
1.1.1.12 Leakage and disappearance of data	High	Low	Risk Transference	Application of assurance
1.1.1.13 Lack of internal control or security audit	Low	Low	Risk Acceptance	Compromise
1.1.2.1 Portability problem with existing hardware	Low	High	Risk Mitigation	User complies with the specification by using the Ambient service.
1.1.3.1 Problem with fulfilling SLA	High	Low	Risk Transference	Third-party surveillance
1.1.3.2 Insufficient right-to-access management	Low	High	Risk Mitigation	Ambient service side adjusts specification
1.2.1.1 Insufficient information disclosure	Low	High	Risk Mitigation	Ambient service side adjusts specification
1.2.1.2 Problem with service specifications and user requirements	High	High	Risk Avoidance	Adjustment on the use side
1.2.1.3 Crisis regarding continuation of service	High	Low	Risk Transference	Application of assurance
1.2.1.4 Business continuation plan is insufficient	High	Low	Risk Transference	Application of assurance
1.2.2.1 Compliance violation	High	Low	Risk Transference	Third-party surveillance
1.3.1.1 Power failure due to increased consumption	High	Low	Risk Transference	Application of assurance
1.3.1.2 Environmental impacts	Low	Low	Risk Acceptance	Compromise
1.3.1.3 Influence of real-time distribution	Low	High	Risk Mitigation	User complies with the specification by using the Ambient service.
1.3.1.4 Equipment installation problems.	High	High	Risk Avoidance	Adjustment on the use side
2.1.1.1 Complication of operations	Low	High	Risk Mitigation	User complies with the specification by using the Ambient service.
2.1.1.2 Improper management of personal information	High	Low	Risk Transference	Third-party surveillance
2.1.2.1 Portability problem with existing terminal	Low	Low	Risk Acceptance	Adjustment on the offer side
2.1.3.1 Problem with security of right to access	Low	High	Risk Mitigation	Ambient service side adjusts specification.
2.1.3.2 Problem with safety of encryption	Low	High	Risk Mitigation	Ambient service side adjusts specification.
2.2.1.1 Problem in handling personal information	High	Low	Risk Transference	Third-party surveillance
2.2.1.2 Deletion of personal information	High	Low	Risk Transference	Third-party surveillance
2.2.1.3 User's incorrect deletion, alteration, etc.	Low	High	Risk Mitigation	Ambient service side adjusts specification.
2.2.1.4 General information disclosure	Low	High	Risk Mitigation	Application of assurance
2.2.2.1 Problem with access except for a user	Low	High	Risk Mitigation	Others
2.3.1.1 Breakage of device due to consumption	Low	Low	Risk Acceptance	Compromise
2.3.1.2 Communication failure	Low	Low	Risk Acceptance	Compromise
3.1.1 Regulation problem arising from revision of law	Low	Low	Risk Acceptance	Adjustment on the use side
3.2.1 Data center collapses in a disaster	Low	High	Risk Mitigation	Application of assurance
3.2.2 Problem providing compensation to user	Low	High	Risk Mitigation	Third-party surveillance

C. Problem of the previous study

The previous study was qualitative; a more practical quantitative evaluation would be needed in order to implement the countermeasures it identifies. The current study thus is a quantitative risk assessment of the risk factors obtained in our previous study and its proposed countermeasures.

IV. QUANTITATIVE EVALUATION OF AMBIENT SERVICE'S RISKS AND PROPOSED COUNTERMEASURES

Here, the validity of a countermeasure is relatively evaluated through a quantification of the risk factors shown in Table 2. First, a risk formula used in the field of information security management systems (ISMS) is shown [8]-[9]. Next, an approximation for calculating a risk value based on our previous qualitative results is described [15]. Finally, a risk value for ambient services is deduced by using the formula and approximation.

A. Risk formula

Each risk value is quantified using (1), which is used in the field of ISMS [8]-[9].

$$\textit{Risk value} = \textit{value of asset} * \textit{value of threat} \\ * \textit{value of vulnerability} \qquad (1)$$

Generally, the calculation of each element of the right-hand side of (1) is very difficult. In this paper, the following approximation is used to simplify these elements [15].

1) Approximation of the Asset Value

Here, the asset value of (1) is approximated in terms of the degree of incidence in the risk matrix, as shown in Fig. 4. Thus, it is assumed that the asset value is the degree of incidence. By the way, references [9]-[10] define the degree of incidence as 1 (low)-5 (high). As a further approximation, these values are mapped in degree of incidence to a risk matrix [15]. As shown in Fig. 4, the degree of incidence of the risk matrix is divided in two. For the sake of simplicity, the maximum degree of incidence (5) is approximated to the higher of the two divisions. Similarly, the minimum degree of incidence (1) is approximated to the lower of the two.

2) Approximation of the Threat Value

The threat value of (1) is approximated in terms of the generation frequency in the risk matrix, as shown in Fig. 4. From references [9]-[10], the generating frequency is defined as a range from 1 (low) to 3 (high). These values are mapped to the generating frequencies of the risk matrix of Fig. 4, as well as the above-mentioned degree-of-incidence approximation. That is, the maximum generating frequency (3) is approximated to the higher of the two divisions, and the minimum (1) is approximated to the lower of the two.

3) Approximation of the Value of Vulnerability

The vulnerability evaluation is defined in reference [9]-[10] as well. It is defined on a three-level scale, 3 (High), 2 (Medium), and 1 (Low), and these levels were approximated in accordance with the classification of the risk matrix of Figure 4. Here, the four domains of the figure are classified into three categories according to the generating frequency and degree of incidence, as follows.

- Risk Avoidance: both the generating frequency and degree of incidence are high. It approximately corresponds to the highest risk classification.
- Risk Transference and Risk Mitigation: either the generating frequency or the degree of incidence is high. It approximately corresponds to the 2nd highest risk classification.
- Risk Acceptance: both the generating frequency and degree of incidence are low. It approximately corresponds to the lowest risk classification.

In the above-mentioned classification, *Risk Avoidance* cases are approximated to 3 (High), *Risk Transference* and *Risk Mitigation* cases are approximated to 2 (Medium), and *Risk Acceptance* cases are approximated to 1 (Low).

B. Calculation of risk value

The risk value before applying countermeasures against a risk was calculated using (1) (see Table 3).

Next, the risk value after applying countermeasures was calculated. The following two measures were chosen from the viewpoint of practicality: "application of assurance" and "third-party surveillance". These countermeasures can be easily implemented, although their costs may be problematic. Table 4 shows the resulting risk values when performing the countermeasures.

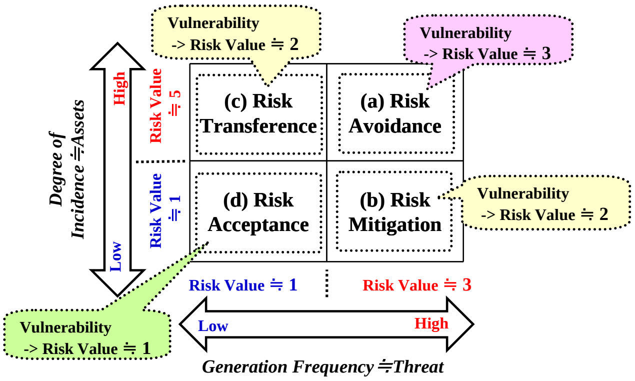


Figure 4. Risk Value Approximation of Risk Matrix [15]

Here, supposing an ideal case, vulnerability was assumed to be 0 as a result using the proposed countermeasures. By the way, supposing an actual case, these countermeasures are not always perfect. For example, in the case of "application of assurance", there may be bankruptcy of an insurance company though its probability is very low. In consideration of such a case, the vulnerability of an actual case is approximated to 1 (the minimum level).

TABLE III. RISK VALUE BEFORE COUNTERMEASURES

Level 3: Risk Factors	Assets	Threat	Vulner-ability	Value of Risk
1.1.1.1 Problem in cooperating with the existing system	5	3	3	45
1.1.1.2 Problem with ending Ambient Service	5	1	2	10
1.1.1.3 Problem with service entrepreneur's specifications	5	3	3	45
1.1.1.4 Problem with service entrepreneur's supervisor	5	1	2	10
1.1.1.5 Leaks, etc., by service entrepreneur	5	1	2	10
1.1.1.6 Data deleted at end of service use	1	3	2	6
1.1.1.7 Problem with requirements for certification	1	3	2	6
1.1.1.8 Problem in managing personal information	1	3	2	6
1.1.1.9 Data seized by other company	5	1	2	10
1.1.1.10 No restoration of missing data	1	1	1	1
1.1.1.11 No security management	5	1	2	10
1.1.1.12 Leakage and disappearance of data	5	1	2	10
1.1.1.13 Lack of internal control or security audit	1	1	1	1
1.1.2.1 Portability problem with existing hardware	1	3	2	6
1.1.3.1 Problem with fulfilling SLA	5	1	2	10
1.1.3.2 Insufficient right-to-access management	1	3	2	6
1.2.1.1 Insufficient information disclosure	1	3	2	6
1.2.1.2 Problem with service specifications and user requirements	5	3	3	45
1.2.1.3 Crisis regarding continuation of service	5	1	2	10
1.2.1.4 Business continuation plan is insufficient	5	1	2	10
1.2.2.1 Compliance violation	5	1	2	10
1.3.1.1 Power failure due to increased consumption	5	1	2	10
1.3.1.2 Environmental impacts	1	1	1	1
1.3.1.3 Influence of real-time distribution	1	3	2	6
1.3.1.4 Equipment installation problems.	5	3	3	45
2.1.1.1 Complication of operations	1	3	2	6
2.1.1.2 Improper management of personal information	5	1	2	10
2.1.2.1 Portability problem with existing terminal	1	1	1	1
2.1.3.1 Problem with security of right to access	1	3	2	6
2.1.3.2 Problem with safety of encryption	1	3	2	6
2.2.1.1 Problem in handling personal information	5	1	2	10
2.2.1.2 Deletion of personal information	5	1	2	10
2.2.1.3 User's incorrect deletion, alteration, etc.	1	3	2	6
2.2.1.4 General information disclosure	1	3	2	6
2.2.2.1 Problem with access except for a user	1	3	2	6
2.3.1.1 Breakage of device due to consumption	1	1	1	1
2.3.1.2 Communication failure	1	1	1	1
3.1.1 Regulation problem arising from revision of law	1	1	1	1
3.2.1 Data center collapses in a disaster	1	3	2	6
3.2.2 Problem providing compensation to user	1	3	2	6
Total				417

TABLE IV. RISK VALUE AFTER COUNTERMEASURES

Level 3: Risk Factors	Proposed countermeasure	Assets	Threat	Vulnerability		Value of Risk	
				Ideal	Actual	Ideal	Actual
1.1.1.1 Problem in cooperating with the existing system	Unapplied	5	3	3	3	45	45
1.1.1.2 Problem with ending Ambient Service	**Third -party Surveillance**	5	1	**0**	**1**	0	5
1.1.1.3 Problem with service entrepreneur's specifications	Unapplied	5	3	3	3	45	45
1.1.1.4 Problem with service entrepreneur's supervisor	**Third -party Surveillance**	5	1	**0**	**1**	0	5
1.1.1.5 Leaks, etc., by service entrepreneur	**Application of assurance**	5	1	**0**	**1**	0	5
1.1.1.6 Data deleted at end of service use	Unapplied	1	3	2	2	6	6
1.1.1.7 Problem with requirements for certification	Unapplied	1	3	2	2	6	6
1.1.1.8 Problem in managing personal information	Unapplied	1	3	2	2	6	6
1.1.1.9 Data seized by other company	**Application of assurance**	5	1	**0**	**1**	0	5
1.1.1.10 No restoration of missing data	Unapplied	1	1	1	1	1	1
1.1.1.11 No security management	**Application of assurance**	5	1	**0**	**1**	0	5
1.1.1.12 Leakage and disappearance of data	**Application of assurance**	5	1	**0**	**1**	0	5
1.1.1.13 Lack of internal control or security audit	Unapplied	1	1	1	1	1	1
1.1.2.1 Portability problem with existing hardware	Unapplied	1	3	2	2	6	6
1.1.3.1 Problem with fulfilling SLA	**Third-party surveillance**	5	1	**0**	**1**	0	5
1.1.3.2 Insufficient right-to-access management	Unapplied	1	3	2	2	6	6
1.2.1.1 Insufficient information disclosure	Unapplied	1	3	2	2	6	6
1.2.1.2 Problem with service specifications and user requirements	Unapplied	5	3	3	3	45	45
1.2.1.3 Crisis regarding continuation of service	**Application of assurance**	5	1	**0**	**1**	0	5
1.2.1.4 Business continuation plan is insufficient	**Application of assurance**	5	1	**0**	**1**	0	5
1.2.2.1 Compliance violation	**Third-party surveillance**	5	1	**0**	**1**	0	5
1.3.1.1 Power failure due to increased consumption	**Application of assurance**	5	1	**0**	**1**	0	5
1.3.1.2 Environmental impacts	Unapplied	1	1	1	1	1	1
1.3.1.3 Influence of real-time distribution	Unapplied	1	3	2	2	6	6
1.3.1.4 Equipment installation problems	Unapplied	5	3	3	3	45	45
2.1.1.1 Complication of operations	Unapplied	1	3	2	2	6	6
2.1.1.2 Improper management of personal information	**Third-party surveillance**	5	1	**0**	**1**	0	5
2.1.2.1 Portability problem with existing terminal	Unapplied	1	1	1	1	1	1
2.1.3.1 Problem with security of right to access	Unapplied	1	3	2	2	6	6
2.1.3.2 Problem with safety of encryption	Unapplied	1	3	2	2	6	6
2.2.1.1 Problem in handling personal information	**Third-party surveillance**	5	1	**0**	**1**	0	5
2.2.1.2 Deletion of personal information	**Third-party surveillance**	5	1	**0**	**1**	0	5
2.2.1.3 User's incorrect deletion, alteration, etc.	Unapplied	1	3	2	2	6	6
2.2.1.4 General information disclosure	**Application of assurance**	1	3	**0**	**1**	0	3
2.2.2.1 Problem with access except for a user	Unapplied	1	3	2	2	6	6
2.3.1.1 Breakage of device due to consumption	Unapplied	1	1	1	1	1	1
2.3.1.2 Communication failure	Unapplied	1	1	1	1	1	1
3.1.1 Regulation problem arising from revision of law	Unapplied	1	1	1	1	1	1
3.2.1 Data center collapses in a disaster	Unapplied	1	3	2	2	6	6
3.2.2 Problem providing compensation to user	**Third-party surveillance**	1	3	**0**	**1**	0	3
Total						**265**	**341**

C. *Results of evaluation*

Table 5 summarizes the results shown in Tables 3 and 4. Although only the "application of assurance" and "third-party surveillance" countermeasures were evaluated in this study, the table shows that the risk can be reduced by from 18% to 36%. These results also show that a detailed numerical expression can treat a risk more specifically by quantifying it and the prospective countermeasure.

D. *Discussion*

As mentioned above, it is not realistic to perform all of the proposed countermeasures on the risks of Table 2. This study thus dealt with only two, i.e., "application of assurance" and "third-party surveillance," chosen on the basis of their practicality. In particular, the "application of assurance" countermeasure is being used in a Cloud user-oriented insurance service that began in 2012 in Japan [16], and "third-party surveillance" is carried out by certification businesses of ISMS.

However, as mentioned above, the problem of cost might also affect these countermeasures. Generally speaking, these countermeasures can become expensive because they need a specialist's knowledge. In the future, we will have to devise a verification considering such costs.

TABLE V. EVALUATION RESULTS

	Before countermeasure against risk factors (①)	After countermeasure against risk factors (②)	
		Ideal case	Actual case
Total risk value	417	265	341
Risk reduction rate = (①−②)/①	−	0.36	0.18

V. RELATED WORK

There has been a lot of research on the security of ambient services. However, each study has been an investigation in regard to the architecture of ambient networks. For example, some of the research targets the implementation of security functions, such as the authentication function [17]-[19], while other research deals with security policies [20].

On the other hand, this paper is a proposal about comprehensive security, which also includes the user side of an ambient service. Such research that takes into account the user side will be important for not only ensuring the security of ambient services but also for spreading new Internet services, such as cyber-physical systems and the IOT (the Internet of Things).

VI. CONCLUSION AND FUTURE WORK

We are interested in promoting ambient services as a next-generation digitized infrastructure by assessing their risks and proposing countermeasures. In our previous study, although countermeasures were developed from a qualitative risk assessment, their effectiveness could not be quantified. Hence, in this study, we performed a quantitative evaluation that used a risk value. It was shown that countermeasures labeled "application of assurance" and "third party surveillance" in the previous study could reduce their corresponding risk factors by 18% - 36%. These results mean that the countermeasures developed in our previous qualitative evaluation can be more specifically evaluated as to their effect by introducing a risk value.

In the future, we will execute further improvement of countermeasures, and verification of cost effectiveness. Furthermore, we will improve the granularity of the quantification. In particular, it is necessary to improve the granularity of the risk matrix to improve the granularity of the quantification. Thus, whereas in this paper, a general four division model was used as a risk matrix, we should improve the model so that it has at least nine divisions in the future.

ACKNOWLEDGMENTS

We would like to thank reviewers for valuable comments on this paper. This work was supported by the Japan Society for the Promotion of Science (JSPS, KAKENHI Grant Number 24300029).

REFERENCES

[1] Ministry of Public Management, Informatin & Communivcations Statistics Database, [Online]. Available from: http://www.soumu.go.jp/johotsusintokei/english/ 2014.12.30

[2] Osaka University, Center of Excellence for Founding Ambient Information Society Infrastructure, [Online]. Available from: http://www.ist.osaka-u.ac.jp/GlobalCOE/indexe_html?set_language=en 2014.12.30

[3] N. Wakamiya and M. Murata, "Introduction to Global COE Project: Center of Excellence for Founding Ambient Information Society Infrastructure," International Workshop on Nonlinear Theoretic Approach to Ambient Network, Oct., 2009. (Invited Talk)

[4] M. Murata, "Global COE Project: Center of Excellence for Founding Ambient Information Society Infrastructure," 14th Academic Exchange Seminar between Shanghai Jiao Tong University and Osaka University, (Shanghai, China), Oct., 2009.

[5] S. Tanimoto, et al., "Risk Management to User Perception of Insecurity in Ambient Service," 13th ACIS International Conference on Software Engineering, pp. 771-776, Aug. 2012

[6] Risk Breakdown Structure, [Online]. Available from:
http://www.justgetpmp.com/2011/12/risk-breakdown-structure-rbs.html, 2014.12.30

[7] Cox's risk matrix theorem and its implications for project risk management, [Online]. Available from:
http://eight2late.wordpress.com/2009/07/01/cox%E2%80%99s-risk-matrix-theorem-and-its-implications-for-project-risk-management/, 2014.12.30

[8] M. S. Toosarvandani, N. Modiri, M. Afzali, "The Risk Assessment and Treatment Approach in order to Provide LAN Security based on ISMS Standard," International Journal in Foundations of Computer Science & Technology (IJFCST), pp.15-36, Vol. 2, No.6, Nov., 2012

[9] H. Sato et al., "Information Security Infrastructure," Kyoritsu Shuppan Co., Ltd., 2010, (in Japanese)

[10] ISMS Risk Assessment Manual v1.4, [Online]. Available from:
https://www.igt.hscic.gov.uk/KnowledgeBaseNew/ISMS%20Risk%20Assessment%20Manual%20v1.4.pdf, 2015.1.4

[11] Ambient Intelligence - Philips and ISTAG, [Online]. Available from:
http://playstudies.wordpress.com/2010/12/01/ambient-intelligence-philips-and-istag/, 2014.12.30

[12] Towards Digital inclusion & innovation, [Online]. Available from:
http://www.kantei.go.jp/jp/singi/it2/kongo/digital/dai9/9siryou2.pdf, 2014.12.30 (in Japanese)

[13] E. A. Lee, "Cyber Physical Systems: Design Challenges," 11th IEEE Symposium on Object Oriented Real-Time Distributed Computing, pp.363-369, May, 2008

[14] A. A. Cardenas, S. Amin, S. Sastry, "Secure Control: Towards Survivable Cyber-Physical Systems," Distributed Computing Systems Workshops, pp.495-500, Jun., 2008

[15] S. Tanimoto, et al., "A Study of Risk Assessment Quantification in Cloud Computing," 8th International Workshop on Advanced Distributed and Parallel Network Applications (ADPNA-2014), pp.426-431, Sep., 2014

[16] Mitsui Sumitomo Insurance, (in Japanese) [Online]. Available from:
http://www.ms-ins.com/news/fy2011/news_0202_1b.html, 2015.1.4

[17] Mahdi Aiash, et al., "A Survey of Potential Architectures for Communication in Heterogeneous Networks," The IEEE Wireless Telecommunications Symposium, April 2012. London, UK.

[18] M. Lebre, et al., "Media Independent Transport Service for Ambient Intelligence," [Online]. Available from: https://ria.ua.pt/bitstream/10773/6601/3/A_Media_Independent_Transport_Service_for_Ambient_Intelligence.pdf, 2015.1.6

[19] A. F. Abate, M. D. Marsico, "MUBAI: multiagent biometrics for ambient intelligence," Journal of Ambient Intelligence and Humanized Computing, Jun. 2011, Vol. 2, Issue 2, pp 81-89, Springer

[20] O.Dohndorf, et al., "Adaptive and Reliable Binding in Ambient Service Systems," IEEE 16th Conference on Date of Conference, pp.1-8, Sept. 2011

Chronomorphic Programs: Using Runtime Diversity to Prevent Code Reuse Attacks

Scott E. Friedman, David J. Musliner, and Peter K. Keller
Smart Information Flow Technologies (SIFT)
Minneapolis, USA
email: {sfriedman,dmusliner,pkeller}@sift.net

***Abstract*—Return Oriented Programming (ROP) attacks, in which a cyber attacker crafts an exploit from instruction sequences already contained in a running binary, have become popular and practical. While previous research has investigated software diversity and dynamic binary instrumentation for defending against ROP, many of these approaches incur large performance costs or are susceptible to Blind ROP attacks. We present a new approach that automatically rewrites potentially-vulnerable software binaries into chronomorphic binaries that change their in-memory instructions and layout repeatedly, at runtime. We describe our proof of concept implementation of this approach, discuss its security and safety properties, provide statistical analyses of runtime diversity and reduced ROP attack likelihood, and present empirical results that demonstrate the low performance overhead of actual chronomorphic binaries.**

Keywords-cyber defense; software diversity; self-modifying code.

I. Introduction

In the old days, cyber attackers only needed to find a buffer overflow or other vulnerability and use it to upload their exploit instructions, then make the program execute those new instructions. To counter this broad vulnerability, modern operating systems enforce “write XOR execute” defenses: that is, memory is marked as either writable or executable, but not both. So exploit code that is uploaded to writable memory cannot be executed. Not surprisingly, attackers then developed a more sophisticated exploit method.

Computer instruction sets are densely packed into a small number of bits, so accessing those bits in ways that a programmer did not originally intend can yield *gadgets*: groups of bits that form valid instructions that can be strung together by an attacker to execute arbitrary attack code from an otherwise harmless program [1][2]. Known as *Return Oriented Programming* (ROP), these types of cyber exploits have been effective and commonplace since the widespread deployment of W⊕X defenses. Software with a single small buffer-overflow vulnerability can be hijacked into performing arbitrary computations using ROP techniques. Hackers have even developed *ROP compilers* that build the ROP exploits automatically, finding gadgets in the binary of a vulnerable target and stringing those gadgets together to implement the attacker’s code [3][4].

This paper presents a fully automated approach for transforming binaries into *chronomorphic* binaries that diversify themselves during runtime, throughout their execution, to offer strong statistical defenses against code reuse exploits such as ROP and *Blind ROP* (BROP) attacks. The idea is to modify the binary so that all of the potentially-dangerous gadgets are repeatedly changing or moving, so that even a BROP attack tool cannot accumulate enough information about the program’s memory layout to succeed.

In the following sections, we discuss related research in this area (Section II) and describe how our prototype Chronomorph tool converts regular binaries into chronomorphic binaries (Section III) and review its present limitations. We then describe an analysis of the safety and security of the resulting chronomorphic binaries, and performance results on early examples (Section IV). We conclude with several directions for future work, to harden the tool and broaden its applicability (Section V).

II. Related Work

Various defense methods have been developed to try to foil code reuse exploits such as ROP and BROP. Some of defenses instrument binaries to change their execution semantics [5] or automatically filter program input to prevent exploits [6]; however these approaches require process-level virtual machines or active monitoring by other processes. Other approahces separate and protect exploitable data (e.g., using shadow stacks [7]), but such approaches incur comparatively high overhead.

To reduce overhead and maintain compatibility with existing operating systems and software architectures, many researchers have focused on lightweight, diversity-based techniques to prevent code reuse exploits. For example, Address Space Layout Randomization (ASLR) is common in modern operating systems, and loads program modules into different locations each time the software is started. However, ASLR does not randomize the location of the instructions *within* loaded modules, so programs are still vulnerable to ROP attacks [8]. Some diversity techniques modify the binaries themselves to make them less predictable by an attacker. For example:

- Compile-time diversity (e.g., [9]) produces semantically equivalent binaries with different structures.
- Offline code randomization (e.g., [10]) transforms a binary on disk into a functionally equivalent variant with different bytes loaded into memory.
- Load-time code randomization (e.g., [11][12]) makes the binary load blocks of instructions at randomized addresses.

These diversity-based approaches incur comparatively lower overhead than other ROP defenses and they offer statistical guarantees against ROP attacks.

ISBN: 978-1-61208-381-0

Unfortunately, these compile-time, offline, and load-time diversity defenses are still susceptible to *Blind* ROP (BROP) attacks that perform runtime reconnaissance to map the binary and find gadgets [13]. So even with compile-time, offline, or load-time diversity, software that runs for a significant period of time without being reloaded (e.g., all modern server architectures) is vulnerable. Some ROP defenses modify the operating system to augment diversity [14][15] provide promising results, but these approaches do not modify existing third-party programs to work on existing operating systems.

Unlike the above diversity techniques, chronomorphic programs diversify themselves *throughout program execution* to statistically prevent code reuse attacks even if the attacker knows the memory layout.

III. Approach

The Chronomorph approach requires changing machine code at runtime, a technique known as *self-modifying code* (SMC). Using SMC, Chronomorph must preserve the functionality of the underlying program (i.e., maintain semantics), maximize diversity over time, and minimize performance costs.

Any SMC methodology requires a means to change the permissions of the program's memory (i.e., temporarily circumvent W$\oplus$X defense) to modify the code and then resume its execution. Different operating systems utilize different memory protection functions, e.g., `mprotect` in Linux and `VirtualProtect` in Windows, but otherwise, the basic instruction set architectures (e.g., x86) are equivalent. In this paper, we describe Chronomorph in a 32-bit Linux x86 setting.

Our approach automatically constructs chronomorphic binaries from normal third-party programs with the following enumerated steps, also illustrated in Figure 1:

Offline:

1) Transform the executable to inject the Chronomorph SMC runtime that invokes `mprotect` and rewrites portions of the binary during execution. This produces a *SMC binary* with SMC functions that are disconnected from the program's normal control flow.
2) Analyze the SMC binary to identify potentially-exploitable sequences of instructions (i.e., gadgets).
3) Identify relocatable gadgets and transform the SMC binary to make those gadgets relocatable.
4) Compute instruction-level, semantics-preserving transforms that denature non-relocatable gadgets and surrounding program code.
5) Write the relocations and transforms to a *morph table* outside the chronomorphic binary.
6) Inject *morph triggers* into the SMC binary so that the program will morph itself periodically. This produces the *chronomorphic binary*.

Online:

7) During program runtime, diversify the chronomorphic binary's executable memory space by relocating and transforming instructions without hindering performance or functionality.

We have implemented each step in this process and integrated third-party tools including a ROP compiler [4], the Hydan tool for computing instruction-level transforms [16], and the open-source `objdump` disassembler. We next describe each of these steps in this process, including the research challenges and the strategy we employ in our Chronomorph prototype implementation. We note relevant simplifying assumptions in our prototype, and we address some remaining research challenges in Section V.

A. Injecting SMC morphing functionality

Before the Chronomorph analysis tool can analyze the binary and compute transformations, it must inject the Chronomorph SMC runtime, which contains functions for modifying memory protection (e.g., `mprotect`), writing byte sequences to specified addresses, and reading the morph table from outside the binary. These Chronomorph functions may *themselves* contain gadgets and have runtime diversification potential, so the SMC-capable binary should be the subject of all further offline analysis.

We identified three ways of automatically injecting the Chronomorph runtime code, based on the format of the target program.

1) Link the target program's source code against the compiled Chronomorph runtime. This produces a dynamically- or statically-linked SMC executable. This is the simplest solution, and the one used in our experiments, but source code may not always be available.
2) Rewrite a statically-linked binary by extending its binary with a new loadable, executable segment containing the statically-linked Chronomorph runtime. This produces a statically-linked SMC executable.
3) Rewrite a dynamically-linked binary by adding Chronomorph procedures and objects to an alternative procedure linkage table (PLT) and global object table (GOT), respectively, and then extend the binary with a new loadable, executable segment containing the dynamically-linked Chronomorph runtime. This produces a dynamically-linked SMC executable.

All three of these approaches inject the self-modifying Chronomorph runtime, producing the *SMC binary* shown in Figure 1. At this point, the self-modification functions are not yet invoked from within the program's normal control flow, so we cannot yet call this a chronomorphic binary.

B. Identifying exploitable gadgets

As shown in Figure 1, the Chronomorph offline analysis tool includes a third-party ROP compiler [4] that automatically identifies available gadgets within a given binary and creates an exploit of the user's choice (e.g., execute an arbitrary shell command) by compiling a sequence of *attack gadgets* from the available gadgets, if possible. The Chronomorph analysis tool runs the ROP compiler against the SMC binary, finding

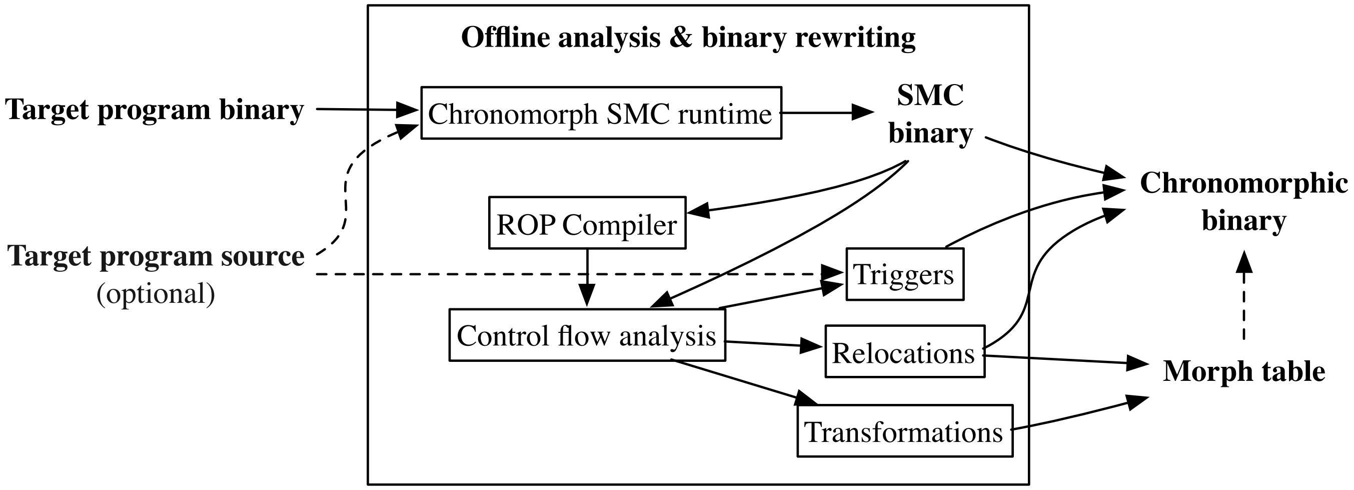


Figure 1. Chronomorph converts a third-party program into a chronomorphic binary.

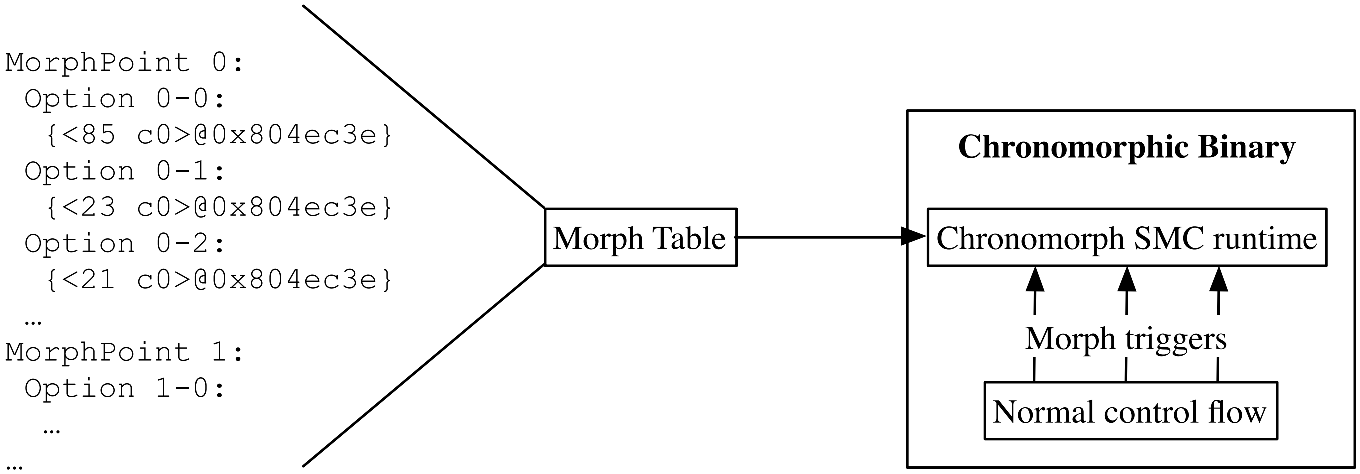

```
MorphPoint 0:
 Option 0-0:
  {<85 c0>@0x804ec3e}
 Option 0-1:
  {<23 c0>@0x804ec3e}
 Option 0-2:
  {<21 c0>@0x804ec3e}
 ...
MorphPoint 1:
 Option 1-0:
  ...
...
```

Figure 2. The resulting chronomorphic binary and its interaction with the morph table.

gadgets that span the entire executable segment, including the Chronomorph SMC runtime.

The ROP compiler prioritizes Chronomorph's diversification efforts as follows, to allocate time and computing resources proportional to the various threat of exploit within the binary:

- Attack gadgets are highest priority. The chronomorphic binary should address these with highest-diversity transforms.
- Available gadgets (i.e., found by the ROP compiler but not present in an attack sequence) are medium priority. These too should be addressed by high-diversity transforms within acceptable performance bounds.
- Instructions that have not been linked to an available gadget are lowest priority, but should still be diversified (i.e., substituted or transformed in-place). Since zero-day gadgets and code-reuse attack strategies may arise after transformation time, this diversification offers additional security.

Our approach attempts all transformations possible, saving more costly transformations, e.g., dynamic block relocations, for the high-risk attack gadgets. The ROPgadget compiler [4] used by Chronomorph may be easily replaced by newer, broader ROP compilers, provided the compiler still compiles attacks and reports all available gadgets. Also, a portfolio approach may be used, running a variety of ROP compilers and merging their lists of dangerous gadgets.

C. Diversity with relocation

We may not be able to remove a high-risk gadget entirely from the executable, since its instructions may be integral to the program's execution; however, the chronomorphic binary can relocate it with high frequency throughout execution, as long as it preserves the control flow.

Relocation is the highest-diversity strategy that Chronomorph offers. Chronomorph allocates an empty *block relocation space* in the binary, reserved for gadget relocation. Whenever the chronomorphic binary triggers a morph, it shuffles relocated blocks to random locations in the block relocation space and repairs previous control flow with recomputed jmp instructions to the corresponding location in the block relocation space.

For each high-risk attack gadget, Chronomorph performs the following steps to make it relocatable during runtime:

1) Compute the *basic block* (i.e., sequence of instructions with exactly one entry and exit point) that contains the gadget.
2) Relocate the byte sequence of the gadget's basic block to the first empty area in the block relocation space.
3) Write a jmp instruction from the head of the basic block to the new address in the block relocation space.
4) Write nop instructions over the remainder of the gadget's previous basic block, destroying the gadgets.
5) Write the block's byte sequence and the address of the new jmp instruction to the morph table.

The morph table now contains enough information to place the gadget-laden block anywhere in the block relocation space and recompute the corresponding jmp instruction accordingly.

Intuitively, diversity of the binary increases with the size of the block relocation space. For a single gadget block g of with byte-size $|g|$, and block relocation space of size $|b|$, relocating g adds $V(g, b) = |b| - |g|$ additional program variants.

If we relocate multiple gadget blocks $G = \{g_0, ..., g_{|G|-1}\}$, the we add the following number of variants:

$$V(G, |b|) = \prod_{i=0}^{|G|-1} (|b| - \sum_{j=0}^{i} |g_j|). \qquad (1)$$

The probability of guessing all of the relocated gadgets' addresses is therefore $1/V(G, |b|)$, which diminishes quickly as the block relocation space increases.

Our Chronomorph prototype has the following constraints for choosing gadget blocks for relocation:

- Relocated blocks cannot contain a call instruction. When a call instruction is executed, the subsequent instruction's address is pushed onto the stack, and if the calling block is relocated, execution would return into an arbitrary spot in the block relocation space.
- Relocated blocks must be at least the size of the jmp to the block relocation space, so that Chronomorph has room to write the jmp.
- Relocated blocks must end in an indirect control flow (e.g., ret) instruction; otherwise, we would have to recompute the control flow instruction at the block's tail at every relocation. Empirically, the vast majority of these blocks end in ret.
- Relocated gadgets cannot span two blocks.

We discuss some improvements in the conclusion of this paper for hardening Chronomorph and removing some of these constraints.

D. Diversity with in-place code randomizations

Chronomorph uses *in-place code randomization* (IPCR) strategies to randomize non-relocated instructions [10]. IPCR performs narrow-scope transformations without changing the byte-length of instruction sequences.

At present, Chronomorph use two IPCR strategies to compute transformations. The first, *instruction substitution* (IS), substitutes a single instruction for one or more alternatives. For example, comparisons can be performed in either order, xor'ing a register with itself is equivalent to mov'ing or and'ing zero, etc. These instructions have the same execution semantics, but they change the byte content of the instruction, so unintended control flow instructions (e.g., 0xC3 = ret) are potentially transformed or eliminated. A single IS adds as many program variants as there are instruction alternatives.

Another IPCR strategy, *register preservation code reordering* (RPCR) reorders the pop instructions before every ret instruction of a function, and also reorders the corresponding push instructions at the function head to maintain symmetry. A register preservation code reordering for a single function adds as many variants as there are permutations of push or pop instructions.

Importantly, RPCR changes the layout of a function's stack frame, which may render it non-reentrant. For instance, if control flow enters the function and it preserves register values via push'ing, and then the chronomorphic binary runs RPCR on the function, it will likely pop values into unintended registers and adversely affect program functionality.

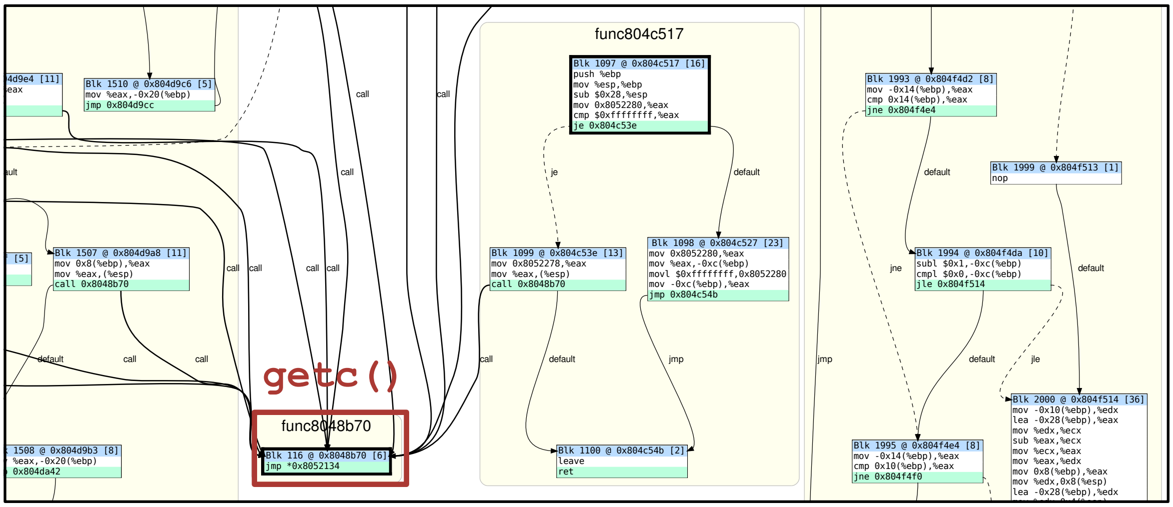
Any stack-frame diversity method such as RPCR should only be attempted at runtime if execution cannot *reenter* the function, e.g., from an internal call, after a SMC morph operation. We enforce this analytically with control flow graph (CFG) analysis: if execution can reenter a function f from the morph trigger (i.e., if the morph trigger is *reachable* from f in the CFG), the stack frame of f should not be diversified. Stack frame diversification is a valuable tool for ROP defense, but it requires these special considerations when invoked during program execution.

E. Writing and reading the morph data

The morph table is a compact binary file that accompanies the chronomorphic binary, as shown in Figure 2. The morph table binary represents packed structs: MorphPoint structs with internal MorphOption byte sequences. Each MorphPoint represents a decision point (i.e., an IS or RPCR opportunity) where any of the associated MorphOption structs will suffice. Each MorphPoint is stateless (i.e., does not depend on the last choice made for the MorphPoint), and independent of any other MorphPoint, so random choices are safe and ordering of the morph table is not important.

The relocation data is a separate portion of the morph table, containing the content of relocatable blocks alongside their corresponding jmp addresses. Like IPCR operations, relocations are stateless and independent, provided the Chronomorph runtime does not overlay them in the block relocation space.

Intuitively, the morph table cannot reside statically inside the binary as executable code, otherwise all of the gadgets would be accessible.



```
4d9e4 [11]
eax

Blk 1510 @ 0x804d9c6 [5]
mov %eax,-0x20(%ebp)
jmp 0x804d9cc

[5]

Blk 1507 @ 0x804d9a8 [11]
mov 0x8(%ebp),%eax
mov %eax,(%esp)
call 0x8048b70

k 1508 @ 0x804d9b3 [8]
 %eax,-0x20(%ebp)
 0x804da42

Blk 116 @ 0x8048b70 [6]
jmp *0x8052134

Blk 1097 @ 0x804c517 [16]
push %ebp
mov %esp,%ebp
sub $0x28,%esp
mov 0x8052280,%eax
cmp $0xffffffff,%eax
je 0x804c53e

Blk 1099 @ 0x804c53e [13]
mov 0x8052278,%eax
mov %eax,(%esp)
call 0x8048b70

Blk 1098 @ 0x804c527 [23]
mov 0x8052280,%eax
mov %eax,-0xc(%ebp)
movl $0xffffffff,0x8052280
mov -0xc(%ebp),%eax
jmp 0x804c54b

Blk 1100 @ 0x804c54b [2]
leave
ret

Blk 1993 @ 0x804f4d2 [8]
mov -0x14(%ebp),%eax
cmp 0x14(%ebp),%eax
jne 0x804f4e4

Blk 1999 @ 0x804f513 [1]
nop

Blk 1994 @ 0x804f4da [10]
subl $0x1,-0xc(%ebp)
cmpl $0x0,-0xc(%ebp)
jle 0x804f514

Blk 1995 @ 0x804f4e4 [8]
mov -0x14(%ebp),%eax
cmp 0x10(%ebp),%eax
jne 0x804f4f0

Blk 2000 @ 0x804f514 [36]
mov -0x10(%ebp),%edx
lea -0x28(%ebp),%eax
mov %edx,%ecx
sub %eax,%ecx
mov %ecx,%eax
mov %eax,%edx
mov 0x8(%ebp),%eax
mov %edx,0x8(%esp)
lea -0x28(%ebp),%edx
```

Figure 3. Small portion of a control flow graph (CFG) automatically created during Chronomorph's offline analysis. Shaded regions are functions, instruction listings are basic blocks, and edges are control flow edges.

F. Injecting morph triggers

We have described how Chronomorph injects SMC capabilities into third-party executables and its diversification capabilities, but Chronomorph must also automatically connect the Chronomorph runtime into the program's control flow to induce diversification of executable memory during runtime.

The injection of these *morph triggers* presents a trade-off: morphing too frequently will unnecessarily degrade program performance; morphing too seldom will allow wide windows of attack. Ideally, morphing will happen at the speed of input, e.g., once per server request or transaction or user input (or some modulo thereof). The location of the morph trigger(s) in the program's control flow ultimately determines morph frequency.

Figure 3 shows a portion of the CFG for the program used in our experiment, calling out the `getc()` input function. Chronomorph can inject calls to the SMC runtime at these points, or at calling functions with stack-based buffers.

Chronomorph also includes an interface for the application developer to add a specialized `MORPH` comment in the source code, which is replaced by a morph trigger during the rewriting phase.

G. Runtime diversification

A chronomorphic binary executes in the same manner as its former non-chronomorphic variant, except when the morph triggers are invoked.

When the first morph trigger is invoked, the Chronomorph runtime loads the morph binary and seeds its random number generator. All morph triggers induce a complete SMC diversification of the in-process executable memory according to IPCR and relocation data in the morph table:

1) The block relocation space is made writable with `mprotect`.
2) The block relocation space is entirely overwritten with `nop` instructions.
3) Each relocatable block is inserted to a random block relocation space address, and its `jmp` instruction is rewritten accordingly.
4) The block relocation space is made executable.
5) Each `MorphPoint` is traversed, and a corresponding `MorphOption` is chosen at random and written. Each operation is surrounded by `mprotect` calls to make the corresponding page writable and then executable. Future work will group `MorphPoints` by their address to reduce `mprotect` invocations, but our results demonstrate that the existing performance is acceptable.

We conducted an experiment with our Chronomorph prototype on a third-party Linux binary to characterize the diversity, ROP attack likelihood, and performance overhead of our Chronomorph approach. We discuss this experiment and its results in the next section.

IV. EVALUATION

We tested our prototype tool on small target binaries of Linux desktop applications, into which we deliberately injected vulnerabilities and gadgets. Here, we discuss results for the `dc` (desktop calculator) program.

The original target binary, with injected flaws, is easily compromised by our ROP compiler. After running the prototype Chronomorph system, the new binary operates as described in Figure 2, and cannot be defeated by the ROP compiler. The dynamically-linked version of the target binary is small (47KB), and after our tool has made it chronomorphic (with a block relocation space of 4KB) it is 62KB.

The rewritten binary is currently able to perform approximately 1000 changes to its own code in less than one millisecond. When the chronomorphic binary is not rewriting itself, it incurs no additional performance overhead, so the overhead is strictly the product of the time for a complete morph (e.g., one millisecond) and the frequency of morphs, as determined by the injected morph triggers. For our experiments of `dc` that

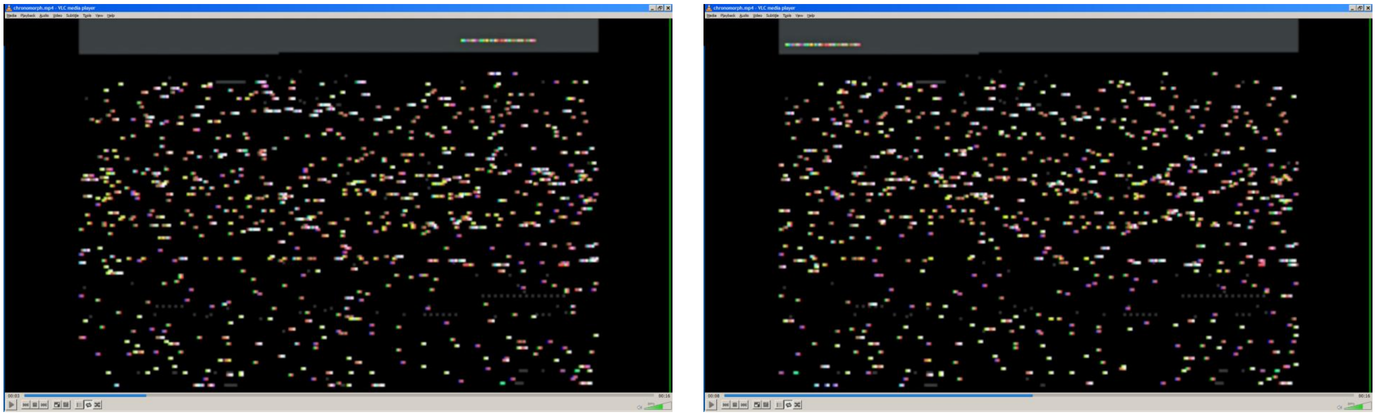


Figure 4. Example memory visualizations illustrating how the executable memory space of the binary changes at runtime.

performed a short regression test, the chronomorphic version of `dc` incurred an additional 2% overhead, but overhead will depend on morph trigger placement for other binaries. Note, however, that a more compute-intensive application might suffer a mild degradation due to cache-misses and branch prediction failures that might not occur in the non-morphing version.

Figure 4 shows two bitmaps illustrating how the binary instructions change in memory, as the program runs. Each pixel of each image represents a single byte of the program's executable code segment in memory. At the top of both images, the gray area is the `nop`-filled block relocation space, with colored segments representing the blocks moved there. Note that the colored segments are in different locations in the two images. Below the gray area, the original binary bits that are never changed remain black, while instructions that are rewritten are shown in different colors, where the RGB is computed from the byte values and `nop` instructions are gray. Again, comparison of the images will show that many of the colored areas are different between the images.

We assessed this example's morph table and estimate that it is capable of randomly assuming any one of approximately 10^{500} variants at any given time during execution. The chronomorphic version of the statically-linked target binary ($>$ 500KB) can assume any one of approximately 10^{8000} variants, using about 8ms to perform all of its rewrites. However, those variant counts do not really accurately characterize the probability that a ROP or BROP attack will succeed.

To do that, we must consider how many gadgets the attacker would need to locate, and how they are morphing. The dynamically linked target contains 250 indirect control flow instructions, and two thirds of those potentially risky elements are moved by the block-relocation phase. With the ROPgadget compiler we used for this evaluation, the original application yielded an exploit needing eight gadgets, of which six were subjected to morphing:

- `inc eax ; ret` – relocated.
- `int 0x80` – relocated.
- `pop edx ; ret` – relocated.
- `pop edx ; pop ecx ; pop ebx ; ret` – reordered (6 permutations).
- `pop ebx ; ret` – relocated.
- `xor eax,eax ; ret` – intact.
- `pop eax ; ret` – relocated.
- `move [edx],eax ; ret` – intact.

Five of the gadgets are relocated dynamically within the block relocation space of size $|b|$, and a sixth gadget is rewritten to one of six permutations. As a result, to accurately locate all eight of those gadgets in the chronomorphed binary, a potential ROP attacker would have to pick correctly from approximately $6{*}|b|^5$ alternatives. For our $|b| = 4$KB example, the probability of a correct guess is approximately $1/10^{18}$, which is extremely unlikely. Needless to say, the ROPgadget exploit was unable to compromise the chronomorphic binary, in thousands of tests. Furthermore, a BROP attack will have no ability to accumulate information about gadget locations, because they change every time a new input is received.

V. Conclusion and Future Work

We have implemented an initial version of an automatic Chronomorph tool and demonstrated that the resulting chronomorphic binaries are resistant to ROP and BROP attacks and retain their initial functionality. Our automatically-generated chronomorphic binary incurred no runtime overhead during normal operation, and only incurred one millisecond overhead to perform over 1000 sequential rewrites to executable memory during a morph operation. However, many research challenges remain for safety and scalability, including:

• **Handling threading —** The morphing behavior must not affect code blocks that are in the middle of execution. For multi-threaded applications, this will require a mechanism to lock the threads out of morphing sections or, more simply, to synchronize the threads in preparation for a morph. If source code is available, adding these sorts of mechanisms is relatively straightforward. To work on pure binaries, more powerful data flow analysis and code injection methods will be required.

• **Protecting the** `mprotect` **—** The system call that allows the chronomorphing code to rewrite executable code is, of course, a dangerous call; if an attacker could locate it and exploit it, he could rewrite the code to do whatever he wants. Therefore, we would ideally like the rewriting/SMC code itself to relocate or transform at runtime; however, the code cannot rewrite itself. We can work around this limitation with a fairly simple trick: we can use two copies of the critical code to alternately rewrite or relocate each other throughout runtime.
• **Protecting the morph table —** While chronomorphic binaries do not rely on obscurity for security, an attacker's chances of success would be higher if he has access to the morph table describing how the binary can change itself. Fairly straightforward encryption techniques should allow us hide and denature the morph table.

These challenges represent areas of future research and development for chronomorphic programs. Our prototype tool and preliminary analyses demonstrate that chronomorphic binaries reduce the predictability of code reuse attacks for single-threaded programs, and we believe that these avenues of future work will improve the safety and robustness of chronomorphic binaries in complex multi-threaded applications.

Acknowledgments

This work was supported by The Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory (AFRL) under contract FA8650-10-C-7087. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

[1] H. Shacham, "The geometry of innocent flesh on the bone: Return-into-libc without function calls (on the x86)," in Proceedings of the 14th ACM conference on Computer and communications security. ACM, 2007, pp. 552–561.
[2] T. Bletsch, X. Jiang, V. W. Freeh, and Z. Liang, "Jump-oriented programming: a new class of code-reuse attack," in Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security. ACM, 2011, pp. 30–40.
[3] E. J. Schwartz, T. Avgerinos, and D. Brumley, "Q: Exploit hardening made easy." in USENIX Security Symposium, 2011, pp. 25–41.
[4] J. Salwan and A. Wirth, "Ropgadget," URL http://shell-storm.org/project/ROPgadget, 2011.
[5] J. Hiser, A. Nguyen-Tuong, M. Co, M. Hall, and J. W. Davidson, "Ilr: Where'd my gadgets go?" in Security and Privacy (SP), 2012 IEEE Symposium on. IEEE, 2012, pp. 571–585.
[6] S. E. Friedman, D. J. Musliner, and J. M. Rye, "Improving automated cybersecurity by generalizing faults and quantifying patch performance," International Journal on Advances in Security, vol. 7, no. 3–4, in press.
[7] L. Davi, A.-R. Sadeghi, and M. Winandy, "Ropdefender: A detection tool to defend against return-oriented programming attacks," in Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security. ACM, 2011, pp. 40–51.
[8] H. Shacham et al., "On the effectiveness of address-space randomization," in Proceedings of the 11th ACM conference on Computer and communications security. ACM, 2004, pp. 298–307.
[9] M. Franz, "E unibus pluram: massive-scale software diversity as a defense mechanism," in Proceedings of the 2010 workshop on New security paradigms. ACM, 2010, pp. 7–16.
[10] V. Pappas, M. Polychronakis, and A. D. Keromytis, "Smashing the gadgets: Hindering return-oriented programming using in-place code randomization," in Security and Privacy (SP), 2012 IEEE Symposium on. IEEE, 2012, pp. 601–615.
[11] R. Wartell, V. Mohan, K. W. Hamlen, and Z. Lin, "Binary stirring: Self-randomizing instruction addresses of legacy x86 binary code," in Proceedings of the 2012 ACM conference on Computer and communications security. ACM, 2012, pp. 157–168.
[12] A. Gupta, S. Kerr, M. S. Kirkpatrick, and E. Bertino, "Marlin: A fine grained randomization approach to defend against rop attacks," in Network and System Security. Springer, 2013, pp. 293–306.
[13] A. Bittau, A. Belay, A. Mashtizadeh, D. Mazieres, and D. Boneh, "Hacking blind," in Proceedings of the 35th IEEE Symposium on Security and Privacy, 2014, pp. 227–242.
[14] C. Giuffrida, A. Kuijsten, and A. S. Tanenbaum, "Enhanced operating system security through efficient and fine-grained address space randomization." in USENIX Security Symposium, 2012, pp. 475–490.
[15] M. Backes and S. Nürnberger, "Oxymoron: making fine-grained memory randomization practical by allowing code sharing," in Proceedings of the 23rd USENIX conference on Security Symposium. USENIX Association, 2014, pp. 433–447.
[16] R. El-Khalil and A. D. Keromytis, "Hydan: Hiding information in program binaries," in Information and Communications Security. Springer, 2004, pp. 187–199.