MOBILITY 2013

The Third International Conference on Mobile Services, Resources, and Users

ISBN: 978-1-61208-313-1

November 17 - 22, 2013

Lisbon, Portugal

MOBILITY 2013 Editors

Josef Noll, University of Oslo & Movation, Norway

MOBILITY 2013

Foreword

The Third International Conference on Mobile Services, Resources, and Users (MOBILITY 2013), held between November 17-22, 2013 in Lisbon, Portugal, continued a series of events dedicated to mobility-at-large, dealing with challenges raised by mobile services and applications considering user, device and service mobility.

Users increasingly rely on devices in different mobile scenarios and situations. "Everything is mobile", and mobility is now ubiquitous. Services are supported in mobile environments, through smart devices and enabling software. While there are well known mobile services, the extension to mobile communities and on-demand mobility requires appropriate mobile radios, middleware and interfacing. Mobility management becomes more complex, but is essential for every business. Mobile wireless communications, including vehicular technologies bring new requirements for ad hoc networking, topology control and interface standardization.

We take here the opportunity to warmly thank all the members of the MOBILITY 2013 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 who dedicated much of their time and efforts to contribute to MOBILITY 2013. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the MOBILITY 2013 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that MOBILITY 2013 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the area of mobile services, resources and users.

We are convinced that the participants found the event useful and communications very open. We hope that Lisbon, Portugal, provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

MOBILITY 2013 Chairs:

MOBILITY General Chair

Josef Noll, University of Oslo & Movation, Norway

MOBILITY Advisory Committee

Petre Dini, Concordia University, Canada & IARIA, USA
Pekka Jäppinen. Lappeenranta University of Technology, Finland
Abdulrahman Yarali, Murray State University, USA

MOBILITY Industry Liaison Chairs

Filipe Cabral Pinto, Telecom Inovação S.A., Portugal
Xiang Song, Microsoft, USA
Xun Luo, Qualcomm Inc. - San Diego, USA

MOBILITY Special Area Chairs on Video

Mikko Uitto, VTT Technical Research Centre of Finland, Finland
Sandro Moiron, University of Essex, UK

MOBILITY Special Area Chairs on Mobile Wireless Networks

Mohammad Mushfiqur Chowdhury, University of Oslo, Norway
Masashi Sugano, Osaka Prefecture University, Japan

MOBILITY Special Area Chairs on Mobile Web / Application

In-Young Ko, Korea Advanced Institute of Science and Technology (KAIST), Korea

MOBILITY Special Area Chairs on Context-aware, Media, and Pervasive

Brent Lagesse, Oak Ridge National Laboratory, USA

MOBILITY Special Area Chairs on Mobile Internet of Things and Mobile Collaborations

Jörn Franke, SAP Research Center - Sophia Antipolis, France
Nils Olav Skeie, University College Telemark, Norway

MOBILITY Special Area Chairs on Vehicular Mobility

Gianluca Franchino, CEIICP - Scuola Superiore Sant'Anna - Pisa, Italy

MOBILITY Special Area Chairs on Mobile Cloud Computing

Chunming Rong, University of Stavanger, Norway
Josef Noll, Center for Wireless Innovation, Norway

MOBILITY Publicity Chairs

Aline Carneiro Viana, INRIA Saclay - Ile de France - Orsay, France
Sarfraz Alam, UNIK-University Graduate Center, Norway

MOBILITY 2013

Committee

MOBILITY General Chair

Josef Noll, University of Oslo & Movation, Norway

MOBILITY Advisory Committee

Petre Dini, Concordia University, Canada & IARIA, USA
Pekka Jäppinen. Lappeenranta University of Technology, Finland
Abdulrahman Yarali, Murray State University, USA

MOBILITY Industry Liaison Chairs

Filipe Cabral Pinto, Telecom Inovação S.A., Portugal
Xiang Song, Microsoft, USA
Xun Luo, Qualcomm Inc. - San Diego, USA

MOBILITY Special Area Chairs on Video

Mikko Uitto, VTT Technical Research Centre of Finland, Finland
Sandro Moiron, University of Essex, UK

MOBILITY Special Area Chairs on Mobile Wireless Networks

Mohammad Mushfiqur Chowdhury, University of Oslo, Norway
Masashi Sugano, Osaka Prefecture University, Japan

MOBILITY Special Area Chairs on Mobile Web / Application

In-Young Ko, Korea Advanced Institute of Science and Technology (KAIST), Korea

MOBILITY Special Area Chairs on Context-aware, Media, and Pervasive

Brent Lagesse, Oak Ridge National Laboratory, USA

MOBILITY Special Area Chairs on Mobile Internet of Things and Mobile Collaborations

Jörn Franke, SAP Research Center - Sophia Antipolis, France
Nils Olav Skeie, University College Telemark, Norway

MOBILITY Special Area Chairs on Vehicular Mobility

Gianluca Franchino, CEIICP - Scuola Superiore Sant'Anna - Pisa, Italy

MOBILITY Special Area Chairs on Mobile Cloud Computing

Chunming Rong, University of Stavanger, Norway
Josef Noll, Center for Wireless Innovation, Norway

MOBILITY Publicity Chairs

Aline Carneiro Viana, INRIA Saclay - Ile de France - Orsay, France
Sarfraz Alam, UNIK-University Graduate Center, Norway

MOBILITY 2013 Technical Program Committee

Jemal Abawajy, Deakin University - Geelong, Australia
Ioannis Anagnostopoulos, University of Central Greece, Greece
Payam Barnaghi, University of Surrey, UK
Mostafa Bassiouni, University of Central Florida - Orlando, USA
Alessandro Bazzi, IEIIT-CNR, Italy
Evangelos Bekiaris, CERTH/HIT, Greece
Paolo Bellavista, University of Bologna, Italy
Rajendra V Boppana, University of Texas - San Antonio, USA
Paolo Bouquet, University of Trento, Italy
Carlos Carrascosa Casamayor, Universidad Politécnica de Valencia, Spain
Ciro Cattuto, Data Science Lab - ISI Foundation, Italy
Ioannis Christou, Athens Information Technology, Greece
Yan Cimon, Université Laval, Canada
Klaus David, University of Kassel, Germany
Claudia de Andrade Tambascia, CPqD Foudation, Brazil
Amnon Dekel, Hebrew University of Jerusalem. Israel
Emanuele Della Valle, Politecnico di Milano, Italy
Raimund Ege, Northern Illinois University, USA
Gianluigi Ferrari, University of Parma, Italy
Randy Fortier, Thompson Rivers University, Canada
Gianluca Franchino, TeCIP - Scuola Superiore Sant'Anna - Pisa, Italy
Xiaoying Gan, Shanghai Jiao Tong University, China
Thierry Gayraud, Université de Toulouse, France
Chris Gniady, University of Arizona, USA
Richard Gunstone, Bournemouth University, UK
Jiankun Hu, Australian Defence Force Academy - Canberra, Australia
Peizhao Hu, NICTA, Australia
Jin-Hwan Jeong, ETRI (Electronics and Telecommunications Research Institute), Korea
Vana Kalogeraki, Athens University of Economics and Business, Greece
Vasileios Karyotis, National Technical University of Athens (NTUA), Greece
Moritz Kessel, Ludwig-Maximilians-Universität München, Germany
Nikos Komninos, Athens Information Technology - Peania, Greece
Ioannis Krikidis, University of Cyprus, Greece
Abderrahmane Lakas, United Arab Emirates University, United Arab Emirates

Jingli Li, TopWorx, Emerson, USA
Xun Luo, Qualcomm Research Center, USA
Dario Maggiorini, University of Milano, Italy
Kirk Martinez, University of Southampton, UK
Barbara M. Masini, CNR - IEIIT, University of Bologna, Italy
Constandinos Mavromoustakis, University of Nicosia, Cyprus
Stefan Michaelis, TU Dortmund University, Germany
Masayuki Murata, Osaka University, Japan
Fatemeh Nikayin, Delft University of Technology, The Netherlands
Ryo Nishide, Ritsumeikan University, Japan
Shumao Ou, Oxford Brookes University, UK
Knut Øvsthus, Høgskolen i Bergen (HiB), Norway
Massimo Paolucci, DOCOMO Euro-Labs, Germany
Evangelos Papapetrou, University of Ioannina, Greece
Symeon Papavassiliou, National Technical University of Athens, Greece
Marco Picone, University of Parma, Italy
Stefan Poslad, Queen Marry University of London, UK
Daniele Puccinelli, University of Applied Sciences of Southern Switzerland (SUPSI), Switzerland
Myriam Ribière, Bell Labs Alcatel-Lucent, France
Joel Rodriques, University of Beira Interior - Covilhã / Instituto de Telecomunicações, Portugal
Anna Lina Ruscelli, TeCIP Institute - Scuola Superiore Sant'Anna, Italy
Djamel Sadok, Federal University of Pernambuco, Brazil
Farzad Salim, Queensland University of Technology - Brisbane, Australia
Stefan Schmid, TU-Berlin, Germany
Christelle Scharff, Pace University - New York City, USA
Minho Shin, Myongji Unversity, South Korea
Behrooz Shirazi, Washington State University, USA
Sabrina Sicari, Università degli studi dell'Insubria, Italy
Andrey Somov, CREATE-NET, Italy
Danny Soroker, IBM T.J. Watson Research Center, USA
Tim Strayer, BBN Technologies, USA
Masashi Sugano, Osaka Prefecture University, Japan
Lars Svensson, German National Library, Germany
Javid Taheri, The University of Sydney, Australia
Vahid Taslimi, Wright State University, USA
Miao Wang, Free University Berlin, Germany
Wei Wang, University of Surrey, UK
Rainer Wasinger, The University of Sydney, Australia
Stephen White, University of Huddersfield, UK
M. Howard Williams, Heriot-Watt University, UK
Hui Wu, University of New South Wales, Australia
Chansu Yu, Cleveland State University, USA
Ting Zhu, State University of New York, USA
Antoine Zimmermann, École Nationale Supérieure des Mines de Saint-Étienne, France

Copyright Information

Table of Contents

Self-Organizing Networks on LTE System:

Antenna Parameters Configuration Effects On LTE Networks Coverage with Respect to Traffic Distribution

Nourredine Tabia, Oumaya Baala, Alexandre Caminada
University of Belfort-Montbéliard (IRTES-SeT laboratory)
Belfort, France
{nourredine.tabia, oumaya.baala, alexandre.caminada}@utbm.fr

Alexandre Gondran
MAAIA Laboratory
Ecole Nationale d'aviation civile (ENAC)
Toulouse, France
Alexandre.gondran@enac.fr

***Abstract*— The new OFDMA-based technology is referred to as the Evolved UMTS Terrestrial Radio Access (E-UTRA) through the Long Term Evolution (LTE) system. This paper proposes to add Self-Organizing functionalities on the antenna architecture, so that the network will be more responsive to changes in traffic and environment. This paper also shows the interest of a robust approach due to the uncertainty of the traffic distribution. First, we develop and validate the interference model based on SINR metric for the deployment of the LTE network, and then we use greedy algorithms to show how the antenna parameters settings such as frequency, tilt and output power, can highly impact the networks coverage due to the traffic changes.**

Keywords-LTE; SON; SINR; interference; parameter setting; optimization; robustness

I. INTRODUCTION

The Long Term Evolution (LTE) is a new air-interface designed by the Third Generation Partnership Project (3GPP) [6]. The 4th generation of mobile system is aimed to achieve additional substantial leaps in term of service provisioning and cost reduction [3]. In addition, to make the network design process time-efficient, Self-Organizing Network (SON) functionalities added within LTE architecture by incorporating automated optimization, can significantly reduce deployment and maintenance costs (CAPital EXpenditure and OPerating Expenditure, respectively). The self-organization capability of a mobile network mainly includes three aspects: self-configuration, self-optimization and self-healing [17]. Focus in this paper is on self-optimization to achieve better network performance goals. The self-optimization phase is preceded by a step of measurements carried out on all the resources to recover data as different as the workflow [18]. This will include assessing network performance toward the user request such as traffic and mobility. Measures may include the characteristics of radio channels, admission, congestion, handover, etc. These measures serve as input data for the self-optimization. In the self-optimization phase, considered as the intelligent phase in SON, effective methods are applied to the previous data to carry out the management of the network based on user needs in traffic demand and quality of service. The optimization involves the physical parameters of antennas (frequency allocation, tilt, azimuth radiation, power, etc.).

In wireless networks, the main key to improve the achievable average throughput by user is to mitigate the inter-cell interference [1, 2] caused by using the same frequency within adjacent cells. Various frequency reuse schemes have been proposed in literature [3, 4, 5].

Resource allocation in radio networks essentially depends on the quality of some reference signals received by the User Equipment (UE). In LTE, they are the Reference Signal Received Power (RSRP) and the Reference Signal Received Quality (RSRQ) corresponding, respectively, to Received Signal Code Power (RSCP) and Ec/No in UMTS (Universal Mobile Telecommunications System). Each user is assigned a portion of the spectrum depending on RSRP and RSRQ. The more complex optimization of reference signal is the RSRQ, which is based on Signal to Interference plus Noise Ratio (SINR) [2, 6]. SINR is an important performance indicator for estimating the achievable throughput, taking into account the interference received from the neighboring cluster of first-tier cells. The estimation and optimization of the SINR are well-known problems in radio communication systems such as the 802.11, the Global System for Mobile Communications (GSM) or the UMTS [11, 12, 13], and LTE needs also a good estimation and control of SINR. Optimizing antenna parameters configuration to meet variant of services and performance requirement is one of main targets of next generation networks. It can significantly improve the coverage and the capacity of the network dealing with the lack of available bandwidth in base stations. Several studies have been done in this direction to understand the impact of parameters on antennal quality of service offered by the network [6, 7]. Didan et al. in [14] have measured the impact of azimuth and tilt inaccuracies on network performance considering three main quality parameters: service coverage, soft handover areas and the ratio of chip energy to interference *Ec/No*. Many simulation results show that azimuth error in the range of ±8 degrees is tolerable to improve the Networks performance.

Tilt parameter has the same effect on Network performance while setting tolerance is just about ±0.5 degrees. The approach of simulated annealing is used by Siomina et al. [7] to study the number of network configuration parameters (the Common PIlot CHanel power, CIPCH, the downtilt and the antenna azimuth) effects toward coverage service in UMTS network. In modern communication systems, further parameters investigations have been developed to meet requirement set. Various combinations of antenna have been studied in term of SINR and throughput performance in LTE case, outlined for example in [6, 15].

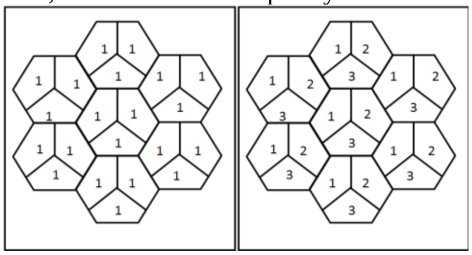
The interference model developed in this paper takes into account the load factor of cells to measure the impact of the traffic on the SINR metric. Our aim is to study the influence of the frequency, tilt and output power parameters on the coverage performance metrics (quality of the SINR, users in outage…), and also emphasize the interest of robust optimization considering frequency, tilt and output power as important design parameters when tuning live network. The choice of the robust approach is mainly due to the uncertainty of the traffic distribution due to the traffic change. Our contribution for SON feature is designed to monitor the performance of the network operation. Following data analysis step, optimization algorithms and corrections will be triggered automatically to make decisions on how to operate the system and this according to the objectives of operators and users needs. We show here the "interest" of using robust approach based on simple greedy descent algorithm. Further work must be done on algorithmic approach.

For this aim, the paper is structured as follows. Section II introduces the system model and basic assumptions. Section III extends this paper and shows the performance metrics and test assumptions. Subsequently, Section IV presents some results to highlight the use of SINR interference model and shows the utility of robust optimization toward the change in the traffic demand. Conclusion and perspectives are drawn in Section V.

II. SYSTEM MODEL

A. *Basic Assumptions*

In this paper, we consider the downlink transmission and illustrate the interference schemes using a theoretical model of seven-cell hexagonal layout as shown in Figure 1. Three sectors are considered in each site (center of the cell) with three evolved Node Bases (eNB), which is the base station in LTE system. Each sector is covered only by only one eNB. Figure 1(a) represents the frequency reuse 1x3x1 pattern where 1 site with 3 sectors use 1 frequency set called sub-band or carrier, that is the same frequency set for all sectors.



a) b)

Figure 1. Seven-cell hexagonal layout. a) 1x3x1 b) 1x3x3

Figure 1(b) represents the frequency reuse 1x3x3 pattern where 1 site with 3 sectors use 3 frequency sets, that is different frequency set for all sectors. The 1x3x3 requires a frequency plan but should avoid interference when the network load becomes too high to apply the 1x3x1 pattern.

The features of our computational model are the following:

- Intra-frequency interference is avoided due to the use of the Orthogonal Frequency Division Multiple Access (OFDMA) technique in downlink transmission. In LTE the orthogonality between subcarriers insures that the interference inside the cell can be ignored.

- The basic resource element in OFDMA is the Physical Resource Block (PRB), which spans both frequency and time dimensions. In this paper, we do not take into account PRB to estimate the inter-cell interference; we only consider the frequency sub-band (or carrier) reuse scheme. Two adjacent cells are scrambling with each other if they are using the same sub-band to transmit data. It gives a fast estimation of the SINR.

- Scenarios are used to represent the traffic distribution of our model. Since LTE network traffic is unpredictable now, we adapt the real UMTS network to simulate the LTE system. The interference model defined here and based on SINR depends not only on antenna parameters (frequency, tilt, azimuth, etc.) but also on traffic load along day; then in our computation the users available in target sector or adjacent sectors can affect the received signal from other users. As we consider that all users communicate at the same time, a scenario of traffic shows the distribution of demand at a given time.

- We introduce a performance indicator to measure occupancy rate on each sector, the load factor. This indicator is a ratio between used bandwidth and the maximum available bandwidth. It is measured for each sector of the network. The feature of the load factor is to determine the bottlenecks of the network and the overloaded sectors. This indicator will be used in the interference model.

B. *Case Study and Problem Formulation*

The considered network for this study consists of tri-sectors sites in the city of Belfort, located in the north-eastern of France. The service area is a 40km x 20km area with a lot of big industrial companies. For our model the service area is divided into a grid of equally sized test points. A test point is a 25x25 meters area. Due to the very small size of the test point, we assume the same signal propagation conditions for all users within the same test point; it means that all users located inside a test point have the same signal and the same RSRP, SINR and RSRQ. Thus, a test point determines the resolution of the computation and the amount of data on signals. A test point is characterized by its number of users and the category of required services for each user (voice, data, etc.). Each sector in the network is equipped with one directive antenna. Each antenna is characterized by its parameters: radiation pattern, azimuth, electrical and mechanical tilts, gain in transmission and reception, frequency and output power in downlink. Due to the dynamic aspect of the network and changes in traffic demand, we introduce the concept of traffic scenarios. A scenario is a given distribution and load of the traffic demand at a given time for each test point of the map.

Several scenarios allow us to compute different situations of network performance to study the robustness problem. The problem formulation is given by the following sets of data, parameters and functions.

Let: $B = \{1,..., n^B\}$, the set of n^B base stations eNB of the network; $T = \{1,..., n^{TP}\}$, the set of n^{TP} test points of the map; and $C_{t,s}$, the number of users located on the test point *t* in scenario *s*.

The interference model based on SINR is thus calculated as defined in (1):

$$\gamma_{b,t,s} = \frac{p^R_{b,t}\, f_b}{\sum_{b'\neq b,\, f_b = f_{b'}} p^R_{b',t}\, f_{b'}\,\delta_{b,s}\,\delta_{b',s} + n_0 w} \qquad (1)$$

where, $\gamma_{b,t,s}$ is the SINR received by the test point *t* and issued from the eNB *b* in scenario *s*; f_b and $f_{b'}$ are binary variable frequencies used by eNB *b* and b', respectively. It means $f_b = f_{b,n} = 1$ as we consider that the eNB *b* is using the carrier *n*. and $f_{b'}$ is a binary variable, which is set to 1 if it uses the same carrier as f_b and 0 otherwise. $\delta_{b,s}$ and $\delta_{b',s}$ are load factors that corresponds to the target eNB *b* and interfering eNBs b' in scenario *s*. Both load factors $\delta_{b,s}$ and $\delta_{b',s}$ are added in the model because we assume that users inside target and interfering cell could impact the quality if the SINR. Base station *b* is said to be saturated in scenario *s* if its load factor $\delta_{b,s}$ is equal to 1. The term *w* represents the total bandwidth used by *b* and n_0 is the thermal noise over the bandwidth *w*. The terms $p^R_{b,t}$ and $p^R_{b',t}$ are the end power received by UE located in test point *t* from, respectively, *b* and *b'*. Note that the load factor in the interference pattern is introduced because we believe that users scramble neighboring cells transmitting on the same frequency, causes more interference and suddenly, penalizes the SINR.

The estimation of this power is based on the Hata propagation model [16] in (2):

$$\begin{aligned} p^R_{b,t}(dBm) &= p_b - PL_{b,t} + g^{MAX}_b \\ &\quad - a^{VER}_b(\theta_{b,t} - t^M_b - t^E_b) - a^{HOR}_b(\phi_{b,t} - a_b) \end{aligned} \qquad (2)$$

where, p_b is the power in *dBm* issued from the eNB *b*. g^{MAX}_b is the antenna gain while a^{VER}_b and a^{HOR}_b are the vertical and horizontal radiation pattern due to the position of the test point from the main beam of the antenna. As shown in Figure 2(a), a^{VER}_b depends essentially on the antenna tilt, which is the angle of the main beam below the horizontal plane. We distinguish two different tilts: the mechanical tilt t^M_b to adjust the physical angle of the antenna brackets and the electrical tilt t^E_b, which does not change the physical angle, but adjusts the radiating currents in the antenna elements to lower the beam in all horizontal directions.

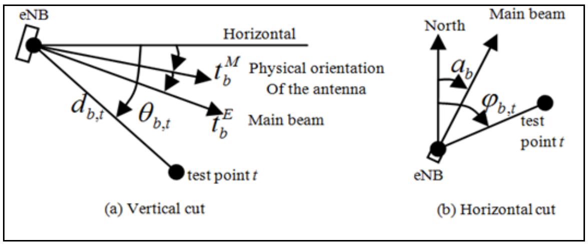


Figure 2. Horizontal and vertical angles

The right part of Figure 3 shows the impact of azimuth parameter on the horizontal radiation pattern. The azimuth is the horizontal angle a_b between the north and the antenna main lobe direction.

In 3GPP LTE tests, we apply the two formulas [6] given by (3) and (4) for the computation of $a^{VER}_b(\theta_{b,t}, t^M_b, t^E_b)$ and $a^{HOR}_b(\varphi_{b,t}, a_b)$:

$$a^{VER}_b(\theta_{b,t}, t^M_b, t^E_b) = -\min\left[12\left(\frac{\theta_{b,t} - t^M_b - t^E_b}{\theta_{3dB}}\right), SLA_v\right] \qquad (3)$$

and $SLA_v = 20dB$

$$a^{HOR}_b(\phi_{b,t}, a_b) = -\min\left[12\left(\frac{\phi_{b,t} - a_b}{\phi_{3dB}}\right), A_m\right] \qquad (4)$$

and $A_m = 25dB$

where, SLA_v is the Side Lobe Attenuation and A_m is the front-to-back attenuation [10]. θ_{3dB} and φ_{3dB} are the half power beam width in vertical and horizontal plan respectively.

Finally, the path loss based on Hata model [16] for urban areas is formulated as in (5):

$$\begin{aligned} PL_{b,t} &= 69.55 + 26.16\log(f_0) - 13.82\log(z_b) \\ &\quad - a(z_t) + \left(44.9 - 6.55\log(z_t)\right)\log(d_{b,t}) \end{aligned} \qquad (5)$$

And for small or medium sized city the value of $a(z_t)$ is:

$$a(z_t) = 0.8 + \left(1.1\times\log(f_0) - 0.7\right) z_t - 1.56\log(f_0) \qquad (6)$$

where, $PL_{b,t}$ is the path loss (dB) in urban area between the eNB *b* and the test point *t*, f_0 is the frequency (MHz), z_b and z_t are the height of the base station *b* and the test point *t* (m), $d_{b,t}$ is the distance between the base station *b* and the center of the test point *t* (m), while $a(z_t)$ is the correction factor for mobile unit antenna height (dB).

In the current work, we will consider three antenna parameters setting. The study focuses on the impact of the frequency, tilt and power parameters on the number of non covered users in the service area. The robust approach uses the mean robustness over three different demand scenarios in a traffic day. The proposed evaluation methodology aims to show the effect of the antenna parameters configuration on the coverage performance metric, with respect to traffic distribution. Considering the study presented in this paper, the formulation of the overall problem is done in the following.

C. Problem formulation

a. Decision variables (parameter settings)

- t_b is the tilt orientation of the eNB $b : t_b \in T_b$ where T_b is the set of possible values of the tilt parameter. For each antenna, we define a set of all possible antennas tilt configurations. We denote this set by $K = \{1,...\mathrm{K}\}$ and assume that the range of possible tilts is the same for all antennas. To describe the current network configuration, we use a set of binary variables, $\Lambda = \{\lambda_b^{(k)}, b \in B, k \in T_b\}$ defined as follows:

$$\lambda_b^{(k)} = \begin{cases} 1 \text{ if antenna } b \text{ uses tilt configuration } k \\ 0 \text{ otherwise} \end{cases}$$

- p_b is the power issued from the eNB b: $p_b \in P_b$ where P_b is the set of possible values of the output power parameter. We denote by $L = \{1,...,\mathrm{L}\}$ as a set of all possible antennas power configurations. We also assume that the range of possible output power is the same for all antennas. We use a set of binary variables, $A = \{\alpha_b^{(l)}, b \in B, l \in P_b\}$ defined as follows:

$$\alpha_b^{(l)} = \begin{cases} 1 \text{ if antenna } b \text{ uses output power } l \\ 0 \text{ otherwise} \end{cases}$$

- $f_{b,n}$ is the variable for carrier assignment to eNB;

$$f_{b,n} = \begin{cases} 1 \text{ if carrier } n \text{ is assigned to eNB } b \\ 0 \text{ otherwise} \end{cases}$$

N, set of available sub-carriers (sub-bands).

b. ***Constraints:*** The main constraints of our model are:

- minimum and maximum number of neighborhood cells for *b*. for this study $|v_b|$ is set to the value 5. $|V_b| = 5$.

$$v_b^{MIN} \le |v_b| \le v_b^{MAX}, \quad \forall b \in B \tag{9}$$

- a test point is associated with exactly one eNB.

$$\forall t \in T, \ \sum_{b \in B} u_{b,t} \le 1 \tag{10}$$

where, $u_{b,t} = \begin{cases} 1 \textit{ if test point t is associated to eNB b} \\ 0 \textit{ otherwise} \end{cases}$

- each eNB *b* can use one and only one carrier *n*.

$$\forall b \in B, \sum_{n \in N} f_{b,n} = 1 \tag{11}$$

- eNB *b* can use only one antenna tilt configuration *k:*

$$\forall b \in B \sum_{k \in T_b} \lambda_b^{(k)} = 1 \tag{12}$$

- eNB *b* can use only one antenna output power configuration *l*.

$$\forall b \in B, \sum_{l \in P_b} \alpha_b^{(l)} = 1 \tag{13}$$

c. Objective functions

- ***Fitness function***

Let $n_{0,s}^C$ be the number of non covered users in scenario *s*.

$$n_{0,s}^C = \sum_{t \in T_0^C} n_{t,s}^C \tag{14}$$

where, $n_{t,s}^C$ is the number of non-covered users in test point *t* for scenario *s*.

- ***Robustness function***

$$f^{Rob} = \sum_{s \in S} n_{0,s}^C \tag{15}$$

where, f^{Rob} is the sum of non-covered users in all considered scenarios together.

- ***Fitness function optimization***: minimize the number of non covered users in one scenario *s:*

$$Min \sum_{t \in T_0^C} n_{t,s}^C \tag{16}$$

- ***Robustness function optimization***: minimize number of non covered users in all considered scenarios:

$$Min \sum_{s \in S} n_{0,s}^C \tag{17}$$

III. TEST ASUMPTIONS AND PERFORMANCE METRICS

The main parameters and assumptions we used are those selected by 3GPP for LTE as shown in Table I. Evaluations are performed by a static snapshot of the network level.

TABLE I. TEST ASSUMPTION FOR LTE DOWNLINK

Parameters	Simulation setting
Network layout	36 sites and 88 sectors
Required service/user	2Mbps
System frequency	1800 Mhz
System bandwidth	20 Mhz
Frequency reuse factor	1x3x1 and 1x3x3
eNB heights range	[17m, 46m]
UE height	1.5 m
Propagation loss model	Hata model [16]
TX power range	[39 dBm, 46 dBm]
Mechanical tilt range	[0°,6°]
Electrical tilt range	[0°,10°]
Azimuth range	[0°,360°]
Horizontal HPBW	+70°
Vertical HPBW	+10°
Antenna gain range	[14dBi , 18.9dBi]
Traffic distribution	Distribution in proportion to UMTS traffic load

In addition to Table I, further assumptions are used for robust optimization tests. Three realistic traffic scenarios are tested (8am, 3pm and 6pm). Frequency scheme reuse 1x3x3 is retained. Deterministic allocation of sub-bands as showed in the right part of Figure 1 is now used. The antennas are grouped by site and stored on the basis of an index in ascending order of the x-axis (the coordinate of the position *(x, y, z)* of the antenna in the network). The performance metric considered is the *Signal to Interference plus Noise Ratio* (SINR).

The SINR, expressed in (1), is an important indicator to evaluate cellular networks. The SINR choice is motivated by the fact that: it takes into account all the parameters of the antenna; it depends on the traffic distribution and the load factor of the network; it resizes the network and determines which base station controls each user; and it allows us to estimate the total throughput of the network. We define two intermediate performance indicators that allow us to evaluate the SINR at each point of the network.

A. *The load factor*

The load factor of the sector/cell is the ratio between the total allocated bandwidth to the cell, which is the required bandwidth and the maximum total bandwidth available in the cell, which are the resources allocated to the cell. Let $\delta_{b,s}$ be the load factor, then: $\delta_{b,s} = w^{S}_{b,s}/w$ where, $w^{S}_{b,s}$ is the total allocated bandwidth to the base station *b* in the reference scenario *s*, and *w* is its maximum available bandwidth. It is worthwhile to mention that load factor is one of the main key indicators in cellular networks. It has been suggested that the downlink cell load for a stable network should not exceed 70% [7]. Huge loaded cells are those for which $\delta_{b,s} > 0.7$ and overloaded cells are those for which $\delta_{b,s} > 1$.

B. *Throughput*

We used the SINR to determine the throughput offered by a base station to the set of users who are located in the cell test points: the higher the SINR, the greater the quality of the channel and the throughput.

TABLE II. MODULATION, THROUGHPUT AND REQUIRED SINR [8]

Index	**Modulation and coding**	**Throughput [Bits/s/Hz]**	**SINRmin [dB]**
0	Outage	0	<0.9
1	QPSK 1/3	0.75	0.9
2	QPSK 1/2	1	2.1
3	QPSK 2/3	1.25	3.8
4	16QAM 1/2	2	7.7
5	16QAM 2/3	2.75	9.8
6	16QAM 5/6	3.25	12.6
7	64QAM 2/3	4	15.0
8	64QAM 5/6	5	18.2

The Table II below gives the current correspondences between SINR, throughput and modulation [8]. The user is in outage if its SINR is below the required threshold for the most robust Modulation and Coding Scheme (MSC).

IV. RESULTS EVALUATION

In order to evaluate the SINR model presented in Section II, we focus on frequency reuse 1 (1x3x1 pattern) and 3 (1x3x3 pattern). The baseline network used for our study is the city of Belfort described in Section II.B. The UE are randomly dropped in each cell in proportion to UMTS traffic load regard to Belfort city. Due to the unavailability of data for LTE networks, we used data from a real GSM/UMTS and adapted it to a 4G-LTE system.

We present now the methodology to evaluate the SINR model taking into account traffic data of the baseline network. Firstly, we assume that the base station is engaged to communicate with a UE if the SINR received by the UE is high enough, i.e., achieve the required SINR threshold to establish a communication. So, the UE is allocated to a base station according to the quality of the received SINR. Then we assume that the load factor is considered to calculate the SINR as mentioned in (1).

A cell is defined as a set of test points of the map; a test point is assigned to the base station, which provides the best SINR. As a first step we assign the test point to the base station on the basis of the best RSRP and define the initial cell coverage for each station. Then we determine the number of UE per cell and the traffic demand as well. We compute the load factor for each station, and finally we estimate the SINR, which depends on the RSRQ. As a second step we assign again the test point to the base station on the basis of the best SINR. From there we estimate again the load factor for each station. This second step is repeated several times (10 times) to try to reach a stable network configuration. The collected traffic data come from a real UMTS network. The tests consider three different scenarios originating from one-day traffic, as shown in Figure 3. The *x* axis unit is a set of time intervals of 15 minutes each.

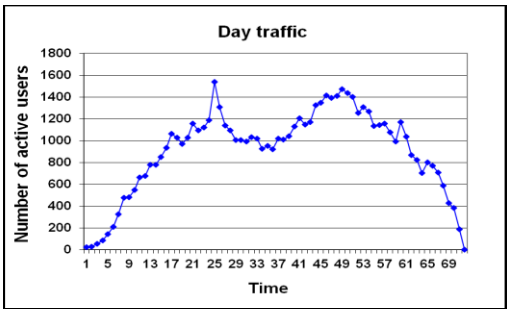


Figure 3. One-day traffic with three chosen scenarios

Three scenarios were selected at different times of the day as follows: a first scenario at 8am with low traffic and 482 users dropped randomly in the network; a second scenario at 3pm with medium traffic and 1,019 users; and a third scenario at 6pm with high traffic and 1,471 users. We are considering that all users are accessing the network at the same time (saturated traffic condition).

A. *Interference model results*

To compare the performances of both reuse 1x3x1 and reuse 1x3x3 patterns, the number of users in outage is used as a performance metric. The minimum required SINR is equal to 0.9 dB. Below this value of SINR, one user cannot establish a communication as defined in (18):

$$\gamma_{b,t,s} < \gamma^{MIN} \qquad (18)$$

where, γ^{MIN} is the required threshold.

Figure 4 depicts the number of users in outage obtained with both reuses pattern: scheme 1x3x1 and scheme 1x3x3. In the network design we test for scheme 1x3x3, the sub-band assignment depends on the azimuth orientation of the sectors. Sectors in opposite direction from neighbor sites are assigned different frequency groups.

The program implementing our model is developed in C++. We run the program 10 times to check the convergence of the SINR computation (the difference is below 5%). An example of three scenarios of traffic is presented in Figure 4 to show the interest of using the reuse scheme 1x3x3 in a real network design and traffic load. We see that the number of users in outage is more important in scheme 1x3x1 (13, 36, and 40) than in scheme 1x3x3 (8, 18 and 17) for the three scenarios of traffic load considered here (8am, 3pm and 6pm, respectively).

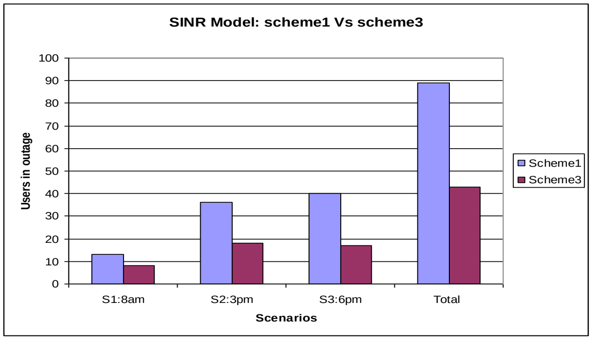


Figure 4. Reuse scheme 1 Vs scheme 3 with 3 scenarios

These results are consistent since the total bandwidth used in pattern 1 can assign more resources, but would jeopardize the rate offered by the network. This is due mainly to the undesirable inter-cell interference generated by neighboring cells using the same frequency set. Such a dense frequency reuse is an obvious pitfall, which limits the throughput at the cell edge. We can also note according to these results that, the higher the traffic, the higher the number of outage users, whatever the frequency pattern solution.

B. *Parameter setting optimization*

The second part of this study is dedicated to robust optimization. It shows that optimizing a number of network configuration parameters like antenna frequency, tilt orientation and power transmission helps considerably to meet variant of services and performance requirement.

A greedy algorithm on frequency, tilt and output power is used to examine how the solution of the antenna configuration, behaves under realistic scenarios.

Algorithm description

In order to show the interest of robust optimization, we present an algorithm able to quickly find a good solution. We were not looking for the best solution at this stage. An iterative algorithm is used; the purpose is not to find an optimal solution but to get the benefit of robust optimization for 3 scenarios in comparison with local solution based on a single scenario. The algorithm with several variants is proposed for each parameter: frequency allocation, tilt and power configuration. We measure the effect of each parameter toward the network coverage performance metric.

-Frequency parameter optimization

It can be proved that frequency assignment problem is NP-hard as it is a graph-coloring problem [19]. For such problems, guaranteeing optimum requires, in the worst case, an enumeration of all possible configurations. The number of possibilities is enormous; in our case study, for 36 cells, 88 antennas and 3 frequency groups, the number of possibilities is $6^{36}= 1.03*10^{28}$. The robust optimization function takes into account the three scenarios considered above. The used algorithm is described in the following. We run the optimization with different varying conditions:

- The scenarios of traffic: several traffic hours.
- The initial frequency assignment to the base stations: deterministic or stochastic per sector from the same site.
- The sites neighborhood search to test the permutation of frequency: sites ranked from the input file or randomly chosen during optimization.

The algorithm starts with a solution using the reuse scheme 1x3x3. The optimization algorithm is run for each scenario to show the best configuration of the frequency parameter setting with respect to the performance metric given by the (14). For each explored site, we evaluate the 6 (3! =6) possibilities of permutations for each sector of the site. The algorithm evaluates 6x88 permutations at each iteration. If a frequency permutation improves the evaluation function of the current solution, the algorithm keeps the last modification and goes through the next sector configuration. The algorithm stops once the current iteration brings no improvement. This is achieved by the following algorithm, which was used for all cases.

Algorithm for antenna parameters optimization

Input parameters

Set B of n^B base stations; Set T of n^{TP} test points; Set S of scenarios: s_1=8am, s_2=3pm and s_3=6pm; Frequency reuse scheme 1x3x3 (3 groups of frequency to assign to base station), tilt and output power operating settings from the real UMTS network.

Variables

-Frequency assignment to base stations.

-Tilt configurations (discrete values from 0° to 12°).

-Output power configurations (discrete values from 36dBm to 46dBm).

Fitness function

Fitness(Config$_p$) = Number of outage users for the current configuration of parameter p, p={*frequency, tilt, output power*}

in s, for non robust optimization

in s_1, s_2 and s_3 ,for robust optimization

Algorithm:

```
Initialize $Config_p$ // $Config_p$ is the initial configuration of a chosen parameter
$Config_p* = Config_p$ // $Config_p*$ is the current best configuration for parameter $p$.
Repeat
Improve=False
      For each site b of the network
            For each possible values of a chosen parameter,
                  Generate the new frequency plan $Config_p$ from $Config_p*$
                  IF $Fitness (Config_p)<Fitness(Config_p*)$
                      $Config_p* = Config_p$
                      Improve =True
                  End IF
            End For
      End For
Until Improve=False
Stopping criteria if there is no improvement
```

Figure 5. Optimization algorithm

-Tilt and power parameters Optimization

The same algorithm is implemented to optimize the tilt parameter and transmission power of the antennas. Different configurations are tested using a discretization of the possible values. We aim at finding good antenna configurations among a range of possible values for both tilt and power. For 36 cells and 88 antennas, we have $13^{88} = 1.06 * 10^{98}$ possibilities. For the tilt parameter, the range of values is [0°, 12°] in degrees, and for power parameter, the range of values are [36dBm, 46dBm]. It is impossible in such case to guarantee an optimum solution; we aim then at finding an acceptable solution and show the tilt and power transmission effects on the traffic demand scenarios. We assume that: three scenarios of traffic are considered; the starting solution on the tilt and output power parameters are those given by the realistic GSM/UMTS network (operating data); the sites in neighborhood search are chosen first in the order of storage as it is in the data file and then, randomly during optimization phase. So, we have two variants of the algorithm for each parameter.

We evaluate the possible values of each parameter for each antenna to meet better number of covered users, using the same starting solution and the same scenarios of traffic (8am, 3pm, and 6pm).

In the first variant of our algorithm, sites are processed in the order of storage in the data file, all possible configurations are tested at each iteration for both tilt and output power parameters. If the current configuration improves the evaluation function of the current solution, we maintain the solution and then process the next neighborhood site chosen in the data file. The algorithm stops once the current iteration brings no improvement (stopping criteria). In the second variant, we keep the same assumptions (same scenarios of traffic and same starting solution) but sites are processed randomly instead of the order in the data file. In this case, the stopping criteria considered here is the running time duration.

Results with non robust and robust optimization

-Frequency optimization results

The results of optimization are shown in the Figures 6 and 7. We emphasize that for the non robust (each scenario tackled alone) and for the robust optimization (the 3 scenarios together) we use the same algorithm but in case of robustness the evaluation function is given by (15) and takes into account the configuration of the frequency considering all the scenarios simultaneously.

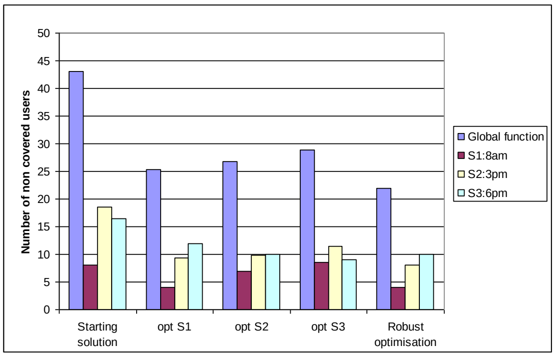


Figure 6. Three scenarios of robust optimization with deterministic frequency allocation

It means that, for each frequency of the network, we evaluate the non-covered users in the three scenarios on one run. So, we run the same algorithm 4 times (one run for each optimization *s1*, *s2*, *s3* and robust case), we use the evaluation function (14) to optimize the 3 scenarios separately; then we use the evaluation function (15) for the robust optimization using the 3 scenarios at the same time.

The x-axis represents the starting solution and the optimization of scenarios s_1, s_2 and s_3 separately and the robust optimization at the end. The y-axis shows the number of users in outage for each scenario s_1, s_2, s_3 and total number. We can note that scenario1 optimization has the smallest number of non-covered users when evaluating s_1 (4 users) comparing to the other cases (8, 7, 9). The same analysis can be done for the scenario s_3 and it is different for s_2 but not far away from the best one. We fixed the run to 20 minutes for random frequency allocation, so there is no guarantee on the solution quality (the convergence is not definitive). We observe that the result of the robust optimization is a trade off between the three scenarios, the best for s_1 and s_2 but not the best for s_3. Finally, the fitness function value of non-covered users for all cases corresponds to the global best solution (blue color in the right part), while in other situations, starting solution and non robust cases, the global function values are 43, 25, 27 and 29, respectively, from left to right part of the Figure 6. The robust optimization does a better compromise between all scenarios. This result shows how the robust approach is important for the remaining of this study. Different variants of the algorithm have been tested by varying several conditions. We run the program 20 minutes for each optimization in Test 2 and Test 3, and keep the best solution for the considered fitness function (so the test conditions are the same for all cases). Test 1 (Figure 6): the initial frequency plan is deterministically assigned and the sites are processed respecting their rank in the input file. Test 2: the initial frequency plan is deterministically assigned and the sites are randomly processed during optimization. The results are similar to the test 1 so we do not plot it. Test 3 (Figure 7): the initial frequency plan is randomly assigned to the co-site sectors and the sites are randomly processed during optimization.

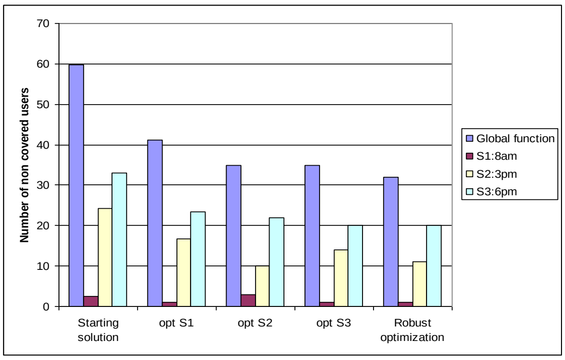


Figure 7. Three scenarios of robust optimization with random frequency allocation

-Tilt optimization results

Robust optimization results for the tilt parameter are highlighted in the right part of each test (Figures 8 and 9). Several tests have been made to show the tilt parameter effect on the number of non covered users, we limit at showing 2 tests varying the neighborhood sites process and the stopping criteria.

Test 1 (Figure 8): the algorithm presented in Figure 5 is implemented. It means that with the same starting solution (3 scenarios of traffic and the operating tilts in real network), the site neighborhood search is done according their storage in the data file. In addition of sites search, the algorithm stops once there is no improvement of the fitness solution. Results are shown in Figure 8. It shows 5 parts, from the right to left: starting solution, optimization of scenarios s_1, s_2 and s_3 separately, and then robust optimization. The starting point of the algorithm is the starting solution (43, 8, 18 and 17), which represents the number of non-covered users in the global function (15) and scenario function (14) for each scenario.

We emphasize that the optimization of each scenario brings better results when each scenario is tackled separately (7, 11, 12) respectively, but impacts the global function. In robust optimization, we can easily note that the configuration found by the algorithm is a better compromise between all scenarios, $f^{Rob} = 31$ comparing with starting solution and the scenarios optimization (43, 41, 33 and 34).

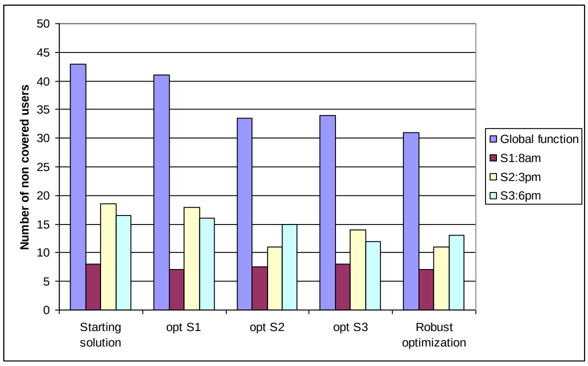


Figure 8. Three scenarios of robust optimization with deterministic tilt configuration

Test 2 (Figure 9): we keep the same initial conditions; but vary the sites neighborhood search (random search) and stopping criteria (an hour of run).

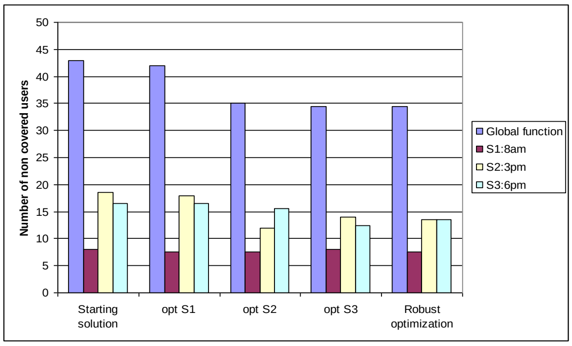


Figure 9. Three scenarios of robust optimization with random tilt configuration

Figure 9 shows the same results as in Figure 8. It confirms that optimizing by means of robust function can improve the network coverage with regard to traffic distribution.

- *Output power optimization results*

Initial output powers used for the starting solution are those operated by the GSM / UMTS network. The same initial conditions, as in the optimization of tilts, were used. Results are shown in the Figure 10.

As in the case of frequencies and tilts, robust optimization of power offers a solution that is a good compromise between the three traffic scenarios (Figure 10). This shows the interest of robust optimization due to the uncertainty of traffic and confirms the results already obtained in the previous cases.

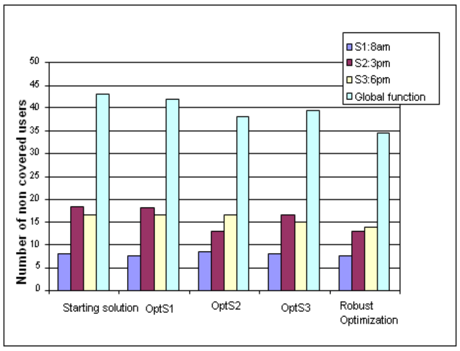


Figure 10. Output power optimization with sites ranked from the input file

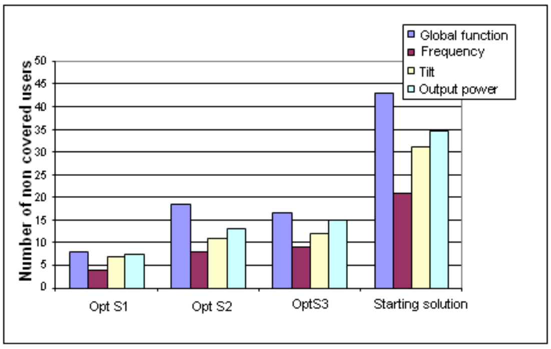


Figure 11. Comparison between frequency, tilt and output power optimization: Sites processed according their storage in data file; stopping criteria when there is no improvement

The Figure 11 represents a comparison between the 3 parameters setting optimizations; with the same initial conditions, using the same greedy algorithm. According to these results, we note that optimizing the frequency parameter provides better results while taking coverage as a performance indicator: less uncovered users for s_1 case (4<7<7.5); s_2 (8<11<13) and s_3 (9<12<15) and also in robust optimization (21<31<34.5). This is due mainly to the undesirable inter-cell interference generated by neighboring cells using the same frequency set. It impacts more the SINR quality comparing to the tilt and output power parameters.

V. CONCLUSION

This paper focuses on the self-organizing networks to automate the configuration of the antenna parameters and shows the interference model and the interest of robust approach with respect of traffic distribution in LTE downlink system. The analysis has been carried out using model radiation pattern and simple model of system performance. We proposed an interference model, which has been validated based on SINR computation and comparing two reuse schemes (1x3x1 and 1x3x3) under realistic scenarios. With respect to coverage, it has been observed that the reference reuse 1x3x3 present best results with respect to the number of covered users, independently on the traffic demand. It shows also that the load factor could impact the quality of signal at end users. It is an important indicator because it highlights the overloaded cells, which represent the bottleneck of the network. Using system simulations, we studied how the frequency, tilt, and output power parameters setting affect the coverage of the macro-cellular scenario. Different combinations of frequency, tilt and output power are used and obtained results show how coverage indicator is sensitive to the combination, and also to the traffic inaccuracies. Simple algorithms used here confirm the interest of robust approach respective of realistic traffic load.

As perspectives, we aim in further studies at analyzing the impact of the parameter settings configuration on the interference model and different performance metrics (throughput, capacity, and coverage). Furthermore, robust optimization approaches like the Variable Neighborhood Search and Tabu Search are under development to highlight the impact of the traffic uncertainty in the deployment of the network.

VI. REFERENCES

[1] X. Mao, A. Maaref, and K.Teo, “Adaptive soft frequency reuse for inter-cell interference coordination in SC-FDMA based 3GPP LTE uplinks,” Proc. IEEE Global Telecommunications Conference (GlobeCom), New Orleans LO, USA, Nov 30-Dec 4. 2008, pp. 1-6.

[2] M. Rahman and H. Yanikomeroglu, “Enhancing cell edge performance: A downlink dynamic interference avoidance scheme with inter-cell coordination,” IEEE Transaction on Wireless Telecommunication, Doha, Qatar. 2010, vol.9, no.4, pp. 1414-1425.

[3] M. Rahman and H. Yanikomeroglu, “Interference Avoidance With Dynamic Inter-Cell Coordination for Downlink LTE System,” Proc. IEEE Wireless Communication and Networking Conference (WCNC), Budapest, Hungary. 2009, pp. 1-6.

[4] 3GPP R1-050738, “Interference Mitigation Considerations and results on frequency reuse,” Siemens, 2005.

[5] 3GPP R1-050507, “Soft Frequency Reuse Scheme for UTRAN LTE,” Huawei, 2005.

[6] O.N.C. Yilmaz, S. Hamalainen, and J. Hamalainen, “System level analysis of vertical sectorisation for 3GPP LTE,” Proc. IEEE 6th International Symposium On Wireless Communication System (CSWCS), Tuscany, Italy. 2009, pp. 453-457.

[7] I. Siomina, P. Varbrand, and D. Yuan, “Automated optimization of service coverage and base station antenna configuration in UMTS networks,” Proc. IEEE Wireless Communications, 2006, vol 13, no. 6, pp. 16-25.

[8] R. Schoenen, W. Zirwas, and B.H Walke, “Capacity and coverage analysis of a 3 GPP-LTE multihop deployment scenario,” Proc. IEEE International Conference On Communications Workshops, Beijin, China. 2008, pp. 31-36.

[9] H. Holma and T. Toskala “LTE for UMTS OFDMA and SC-FDMA Based Radio Access,” John Wiley & sons Ltd Edition, 2009.

[10] O.N.C. Yilmaz and S. Hamalainen, “Comparaison of remote electrical and mechanical antenna downtilt performance for 3GPP LTE,” Proc. IEEE 70th Vehicular Conference Fall (VTC-2009 FALL), Anchorage, AK, USA. 2009, pp. 1-5.

[11] A Gondran, O. Baala, A. Caminada, and H. Mabed, “Interference management in IEEE 802.11 frequency assignment,” Proc. IEEE Vehicular Technology Conference (VTC Spring), 2008, pp. 2238-2242.

[12] H. Mabed and A. Caminada. “Geometric criteria to improve the interference performances of cellular network,” Proc. 64th IEEE Vehicular Technology Conference (VTC Fall), 2006, pp. 1-5.

[13] Z. Altman, J.M. Picard, S. Ben Jamaa, B. Fourestie, A. Caminada, T. Dony, J.F Morlier, S. Mourniac, “New challenges in automatic cell planning of UMTS networks,” Proc. 56th IEEE Vehicular Technology Conference (VTC Fall), 2002, pp. 951-954.

[14] I. Didan and A. Kurochkin, “The impacts of antenna azimuth and tilt installation accuracy on UMTS Network Performance,” Bechtel Telecommunications Technical Journal, 2006. Vol. 4, No. 1.

[15] F. Athley and M. Johansson, “Impact of Electrical and Mechanical Antenna Tilt on LTE Downlink System Performance,” Proc. IEEE Vehicular technology conference, Taipei, Taiwan, 16-19 May. 2010, pp. 1-5.

[16] M. Hata, “Empirical formula for propagation loss in land mobile radio services,” Proc. IEEE Transactions on Vehicular Technology, 1980, vol.29, no. 3.

[17] NGMN, “Next Generation Mobile Networks, Use cases related to Self Organising Network, Overall Description,” May 31, www.ngmn.org, 2007.

[18] J.L. Van Den Berg, R. Leitjens, A. AInsenblatter, M. Armijoo, O. Linnell, C. Blondia, T. Kurner, N. Scully, J. Oszmianski and L.C Schmelz, “Self-Organization in Future Mobile Communication Networks,” Proc. of ICT Mobile Summit, Stockholm, Sweden, June 10-12, 2008.

[19] A. Gamst and W. Rave, “On frequency assignment in automatic telephone system,” Proc. IEEE Global Communication Conference GLOBCOM’82, Miamy, USA, Nov 29-Dece 2. 1982, pp. 309-315.

On the use of Gibbs Sampling for Inter-Cell Interference Mitigation under Partial Frequency Reuse Schemes

K. Koutlia, J. Pérez-Romero, R. Agustí
Dept. of Signal Theory and Communications (TSC)
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
Email: {katkoutlia, jorperez, ramon}@tsc.upc.edu

M. Žiak
Department of Telecommunications and Multimedia
University of Žilina
Žilina, Slovakia
Email: {ziak.mato@gmail.com}

***Abstract*—Fourth Generation (4G) cellular networks present a number of improvements in the overall network performance. However, and despite the advanced technologies that are being employed, Inter Cell Interference (ICI) remains a constraining factor. ICI Coordination techniques target the minimization of ICI and have gained ground in the literature. The introduction of dynamicity in these schemes results in even better bandwidth utilization and enhances the overall performance. In this work, we propose a distributed algorithm that performs dynamic channel allocation to mitigate the ICI in cellular scenarios applying Partial Frequency Reuse (PFR). In particular, the algorithm is based on a Gibbs Sampler mechanism that allows achieving an optimized performance. Simulation results have shown that the proposed solution reduces the network interference up to 13 dB with respect to classical PFR. In addition, benefits have also been observed in the user capacity, where our scheme achieves improvements of up to 43% in terms of average user capacity and up to 17% for the users located at the cell edge.**

Keywords—Dynamic Frequency Allocation; Partial Frequency Reuse; Gibbs Sampler

I. INTRODUCTION

The new era of mobile communications is dictated by the usage of smartphones, tablets and laptops and their demand for high data rate applications and seamless connections. The introduction of the fourth generation (4G) cellular systems has been a crucial point in the history of mobile communications evolution, targeting improved coverage, enhanced capacity and robust, high speed data transfer. Third Generation Partnership Project (3GPP) has adopted Orthogonal Frequency Division Multiple Access (OFDMA) as radio access technology in 4G networks, resulting in better spectral efficiency and in the reduction of the Intra-Cell Interference, due to the orthogonality of the users. However, despite its significant contribution to the overall network performance, Inter-Cell Interference (ICI) can degrade the achievable capacity. Especially for the edge users, which are located close to the cell borders, ICI becomes a constraining factor resulting in a considerable capacity reduction for these users.

In order to cope with the above mentioned problem, 4G systems make use of ICI coordination (ICIC) techniques. These schemes allow the allocation of the available resources to the edge users with higher reuse factors, mitigating in this way the network interference [1]. ICIC techniques usually follow the general concept of Fractional Frequency Reuse (FFR) [2], where the cell is divided in two areas, the inner and the outer, and the same strategy is applied to the available bandwidth.

Different schemes have been proposed for FFR, such as Soft Frequency Reuse (SFR) [3] and Partial Frequency Reuse (PFR) [4][5]. This work focuses on the PFR scheme that splits the cell in two regions, the inner and the outer, as illustrated in Figure 1. In the same way, the bandwidth is divided into the inner band, assigned with a reuse-1 factor (Full Reuse) so that it is common to all the cells, and the outer band, which is assigned with a higher reuse factor (Partial Reuse), e.g., reuse-3, as it can be seen in the right part of Figure 1, where the frequency bands assigned to each inner/outer cell are presented. In addition, PFR allows the possibility for different powers to be used for the downlink (DL) transmissions in the inner/outer parts of the cell.

However, despite of the advantages of the classical PFR schemes, the allocation of the resources follows a static principle. As a result, since the network traffic conditions vary over time, a static allocation will not be able to adapt to these changes [6]. Moreover, these schemes may not be optimal in irregular deployments. As such, research has been focused on the optimization of the ICIC schemes though the introduction of dynamicity. A Dynamic Fractional Frequency Reuse scheme has been presented in [7] making use of a graph-based framework to re-allocate resources depending on cell load variations. In [8], the authors presented a Dynamic Frequency Reuse scheme that mainly deals with uneven traffic loads. Two algorithms are used, one for resource allocation and another one for power control, which significantly improved the network capacity and the energy efficiency. In [9], an adaptive PFR scheme has been developed based on an off-line genetic algorithm enhancing the performance in terms of edge user throughput. Recently the FFR concept has also been proposed for interference management in heterogeneous networks involving both macro and femtocells as for example in [10].

Under the above presented framework, in this work we propose a novel dynamic allocation scheme based on the Gibbs Sampler [11][12] concept as optimization tool. The proposed solution is applied in tri-sectorial PFR deployments targeting the minimization of the downlink ICI. The rationale behind this selection is that this mechanism accomplishes such an interference minimization in a natural way and it is an efficient tool for distributed optimization. In particular, this paper investigates how to optimally assign a set of frequencies in the inner and the outer parts of the different cells in the scenario. The proposed mechanism is suitable for 4G systems with special focus on 3GPP LTE since the partitioning of the available bandwidth in sub-bands and the X2 interface used for coordination purposes constitute it an ideal candidate.

Gibbs sampler-based algorithms for optimization purposes have been widely used in the literature under a variety of situations. Two fully distributed algorithms that follow the concept of Gibbs Sampler have been used in [13] in order to perform channel selection and user association in unmanaged WiFi networks. In [14], the authors have adopted this methodology to improve the performance of homogeneous cellular networks. The optimization targets the power control and the user association. Finally, in [15] a Gibbs sampler-based mechanism has been applied to perform joint optimization in heterogeneous networks.

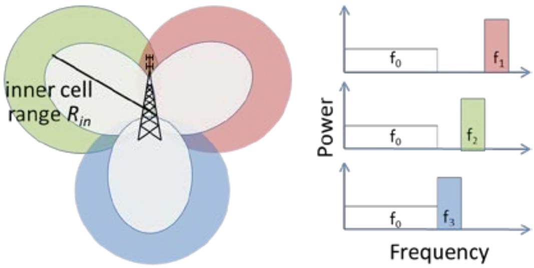


Figure 1: Partial Frequency Reuse Scheme

The rest of the paper is organized as follows. In Section II, a description of the system model and the definition of the notation used throughout the text are given. The optimization model and the algorithm formulation are presented in Section III. In Section IV, the simulation model along with the evaluation of the algorithm performance are presented. Finally, important conclusions and the future work are given in Section V.

II. SYSTEM MODEL

The system model of this work consists of a cellular network where N Base Stations are spatially divided in three sectors (cells) with the use of directional antennas, resulting in a set of cells X.

Users are randomly distributed in the scenario and each user is associated with the cell with which experiences the minimum Path Loss described by the following equation:

$$L_{u,x}(dB) = l_A + l_B \log d_{u,x}(km) - B(\phi_{u,x},\theta_{u,x}) + S_{u,x} \quad (1)$$

where $d_{u,x}$ is the distance between user u and cell x, l_A and l_B are parameters of the propagation model that depend on the considered environment, $S_{u,x}$ is a Gaussian random variable representing the Log-Normal Shadowing between user u and cell x and $B(\phi_{u,x},\theta_{u,x}) = B_H(\phi_{u,x}) + B_V(\theta_{u,x})$ is the antenna pattern decomposed into the horizontal $B_H(\phi_{u,x})$and the vertical $B_V(\theta_{u,x})$ patterns calculated using the following formulas in dB [16]:

$$B_H(\phi_{u,x}) = -min\left(B_o, 12\cdot\left(\frac{\phi_{u,x}-\Phi}{\Delta_\phi}\right)^2\right) \quad (2)$$

$$B_V(\theta_{u,x}) = -min\left(B_o, 12\cdot\left(\frac{\theta_{u,x}-\Theta}{\Delta_\theta}\right)^2\right) \quad (3)$$

where $\phi_{u,x}$ and $\theta_{u,x}$ are the azimuth and elevation angles, respectively, between user u and cell x. Moreover, Φ and Θ are the azimuth and downtilt orientations of the antennas, respectively, Δ_ϕ is the horizontal antenna beam width, Δ_θ is the vertical antenna beam width and B_o is the backward attenuation.

The set of users that is associated with cell $x \in X$ is denoted as U_x. Each user is classified as inner or outer according to a Path Loss Threshold L_{th} that is related to the inner cell range R_{in} as follows:

$$L_{th}(dB) = l_A + l_B \log R_{in}(km) \quad (4)$$

Specifically, a user u associated to cell x belongs to the inner part of the cell if its path loss $L_{u,x}$ is lower than L_{th} . Otherwise, it belongs to the outer part of the cell. Note that L_{th} is the average path loss that it would be observed by a user located at distance R_{in} in the direction of maximum radiated power by the antenna $\phi_{u,x} = \Phi$ and $\theta_{u,x} = \Theta$. As such, the set of users is further split into the inner set $U_{x,in}$ and the outer set $U_{x,out}$. Since the calculation of each users Path Loss includes also the Shadowing, the consideration of a user being inner or outer is not only related to its distance from the base station, but it also accounts for the randomness in the propagation that is inherent to practical wireless scenarios.

Let us consider a set of C frequency channels or sub-bands to be shared among the set of X cells. The bandwidth of each channel $c \in C$ is B_c. For simplicity reasons we assume that each cell can be assigned only one channel for the inner and another one for the outer part; however this work can be easily extended to assign a group of frequencies to each cell. Then, at a given point of time each cell x is characterized by its state $c_x = (c_{x,in}, c_{x,out})$ that is given by the channel $c_{x,in} \in C$ assigned to the inner part and the channel $c_{x,out} \in C$ assigned to the outer part. In order to avoid that the same channel is shared by inner and outer users the allocation will ensure that different channels are assigned to the inner and the outer parts, that is $c_{x,in} \neq c_{x,out}$.

Moreover, we consider that the transmit power of a given cell x in channel c is:

$$P_{x,c} = \begin{cases} P_{x,out} & \text{if } c_{x,out} = c \\ P_{x,in} & \text{if } c_{x,in} = c \\ 0 & otherwise \end{cases} \quad (5)$$

where $P_{x,in}$ and $P_{x,out}$ are the transmit power (in W) of cell x for the inner and the outer parts, respectively.

Based on the above, the Signal to Interference and Noise Ratio (SINR) for an inner user $u_{x,in} \in U_{x,in}$ of cell x is then expressed by the following equation:

$$SINR_{u_{x,in}} = \frac{\dfrac{P_{x,c_{x,in}}}{L_{u_{x,in},x}}}{P_N + \sum_{k \neq x} \dfrac{P_{k,c_{x,in}}}{L_{u_{x,in},k}}} \tag{6}$$

where P_N is the noise power (in W) and $k \in X$ denotes the interfering cells. $L_{u_x,x}$ and $L_{u_x,k}$ respectively denote the Path Loss (in linear units) of user u_x with its serving cell x and with the interfering cell k.

We assume that the bandwidth of one channel in the inner/outer part is equally shared between all users of the inner/outer part. This would correspond to, e.g., a round robin scheduling. In that case, the total capacity (b/s) seen by an inner user of cell x (using Shannon capacity) is:

$$C_{u_{x,in}} = \frac{B_{c_{x,in}}}{\left|U_{x,in}\right|} \log_2\left(1 + SINR_{u_{x,in}}\right) \tag{7}$$

where |.| denotes cardinality. Note that the same expressions (6) and (7) apply for the outer users by simply replacing the *in* sub-index by *out*.

Then, the average capacity per user in the scenario is:

$$C_{user,avg} = \frac{\sum_x \sum_{u_{x,in}} C_{u_{x,in}} + \sum_x \sum_{u_{x,out}} C_{u_{x,out}}}{\sum_x \left(\left|U_{x,in}\right| + \left|U_{x,out}\right|\right)} \tag{8}$$

The target of this work is to find the optimal allocation of frequencies to the inner and the outer parts of the cells (i.e. the optimal cell states c_x) that results in the minimization of the network inter-cell interference and thus, it enhances the capacity. For this purpose a Gibbs sampler-based methodology is proposed. In the following Section, a thorough description of the optimization model is given.

III. Gibbs Sampler for Channel Allocation

The Gibbs Sampler uses the notion of *energy function* which is the optimization target and thus it should be defined in accordance with each specific problem [12]. Therefore, the following sub-sections present the formulation of the energy function considered in this paper for minimizing the ICI in accordance with the system model defined in previous section, and the distributed Gibbs-sampler based algorithm to achieve the minimization.

A. Optimization Model

The target of the proposed optimization approach is to find the states c_x (i.e. the channel allocation) for each cell that minimize the overall inter-cell interference. For that purpose, we define the *global energy* to be minimized as the total interference of the network which will be the sum of the total noise and the interference measured by all the cells:

$$\varepsilon = \sum_x \left[P_N + \frac{1}{\left|U_{x,in}\right|} \sum_{u_{x,in}} \sum_{k \neq x} \frac{P_{k,c_{x,in}}}{L_{u_{x,in},k}} + \frac{1}{\left|U_{x,out}\right|} \sum_{u_{x,out}} \sum_{k \neq x} \frac{P_{k,c_{x,out}}}{L_{u_{x,out},k}} \right] \tag{9}$$

Note that in the above expression we consider for each cell the average interference seen by its served users. Then, the *energy function* can be rewritten as:

$$\varepsilon = \sum_x P_N + \sum_{(x,k)} \left(f(k,x) + f(x,k) \right) \tag{10}$$

where $f(k,x)$ is the interference generated by cell k to cell x:

$$f(k,x) = \frac{1}{\left|U_{x,in}\right|} \sum_{u_{x,in}} \frac{P_{k,c_{x,in}}}{L_{u_{x,in},k}} + \frac{1}{\left|U_{x,out}\right|} \sum_{u_{x,out}} \frac{P_{k,c_{x,out}}}{L_{u_{x,out},k}} \tag{11}$$

As such, the *energy function* derives from the following *potential function* $V(v)$:

$$\varepsilon = \sum \left\{ V(\nu) \middle| \nu \subseteq X \right\} \tag{12}$$

where v represents any possible subset of cells that can be formed with the elements of X and $V(v)$ is given by

$$V(\nu) = \begin{cases} P_N & \text{if } \nu = \{x\} \\ f(k,x) + f(x,k) & \text{if } \nu = \{x,k\} \\ 0 & \text{if } |\nu| \geq 3 \end{cases} \tag{13}$$

A *global energy* which derives from the *potential function* (13) can be optimized using Gibbs with the following *local energy* for each cell x [11][15]:

$$\varepsilon_x = \sum \left\{ V(\nu) \middle| x \in \nu, \nu \subseteq X \right\} = P_N + \sum_{k \neq x} \left(f(k,x) + f(x,k) \right) \tag{14}$$

The Gibbs sampler will compute the *local energy* for each possible state of cell x, $c_x = (c_{x,in}, c_{x,out})$, as follows:

$$\begin{aligned} \varepsilon_x\left(c_{x,in}, c_{x,out}\right) = P_N &+ \sum_{k \neq x} \left(\frac{1}{\left|U_{x,in}\right|} \sum_{u_{x,in}} \frac{P_{k,c_{x,in}}}{L_{u_{x,in},k}} + \frac{1}{\left|U_{x,out}\right|} \sum_{u_{x,out}} \frac{P_{k,c_{x,out}}}{L_{u_{x,out},k}} \right) + \\ &+ \sum_{k \neq x} \left(\frac{1}{\left|U_{k,in}\right|} \sum_{u_{k,in}} \frac{P_{x,c_{k,in}}}{L_{u_{k,in},x}} + \frac{1}{\left|U_{k,out}\right|} \sum_{u_{k,out}} \frac{P_{x,c_{k,out}}}{L_{u_{k,out},x}} \right) \end{aligned} \tag{15}$$

The *local energy* function actually includes the measurement of the interference that users of cell x will experience from the other cells if cell x state is c_x (second term of the equation), as well as the interference that cell x will cause to the neighboring cells (third term of the equation).

B. Algorithm

The minimization of the *energy function* given by (9) is achieved by means of the execution of the procedure indicated in the algorithm presented in Figure 2 at each cell.

Assuming that the system starting time is *t*=0, each cell is assigned with an exponentially distributed timer with mean t_a. When a cell's timer expires, the algorithm is executed and the state c_x selection (i.e. the set of channels $(c_{x,in}, c_{x,out})$) is carried out by sampling a random variable λ using the probability distribution of (16). The latter represents the probability of selecting state c_x among the set of all possible states denoted as C_S. The set C_S includes all the combinations $(c_{x,in}, c_{x,out})$ composed by $c_{x,in} \in C$ and $c_{x,out} \in C$ with $c_{x,in} \neq c_{x,out}$.

$$\pi(c_x) = \frac{e^{\left(-\varepsilon_x(c_{x,in},c_{x,out})/T\right)}}{\sum_{c' \in C_S} e^{\left(-\varepsilon_x(c'_{x,in},c'_{x,out})/T\right)}} \qquad (16)$$

where T is the *temperature* parameter and is calculated as:

$$T = \frac{T_0}{\log_2(2+t)} \qquad (17)$$

In this formula T_0 is a constant and t is the age variable representing the time passed since *t*=0.

```
if cell x timer ($T_x$) expires at time t
    calculate the temperature parameter T      (17)
    for each state $c_x \in C_S$
        calculate the Local Energy $\varepsilon_x(c_{x,in}, c_{x,out})$ (15)
        calculate the Selection Probability $\pi(c_x)$ (16)
    end for
    sample a random variable $\lambda$ with law $\pi(c_x)$
    assign channels $(c_{x,in}, c_{x,out})$ according to the outcome of $\lambda$
    sample an exponential random variable $\mu$ with mean $t_a$
    assign a new timer ($T_x$=t+$\mu$)
end if
```

Figure 2: Algorithm of the Gibbs Sampler Procedure

After the state selection is performed for a given cell, a new timer is generated to schedule the subsequent execution of the algorithm. The probability distribution described above favors the lower energy states and with $T \to 0$ it will converge to a steady state that minimizes the *global interference*.

IV. SIMULATION RESULTS

The performance of the proposed algorithm has been evaluated by means of system level simulations. In this section we present the simulation model and the parameters used, as well as the most important results.

A. Simulation scenario and parameters

The cellular network deployment consists of 4 tri-sectorial base stations, thus having a total of 12 cells, as depicted in Figure 3. The set of channels that can be assigned are $C = \{f_0, f_1, f_2, f_3\}$ with the restriction that $c_{out} \neq c_{in}$. As such, there is a total of 12 possible states to be selected for each cell. Figure 3 depicts the classical PFR scheme that is used as reference for the comparison and evaluation of the results. Moreover, for the Gibbs sampler-based algorithm it is assumed that the initial allocation considered in the beginning of each simulation is also the one shown in Figure 3.

Each cell serves 10 users uniformly distributed in a circular area with range *R*=1 *km*. The total simulation time is 12000·t_a and the simulation step is t_a/24. T_0 is set to 0.7 and the energy values in (15) are given in dBW. Simulations are performed for different values of the inner cell range R_{in}.

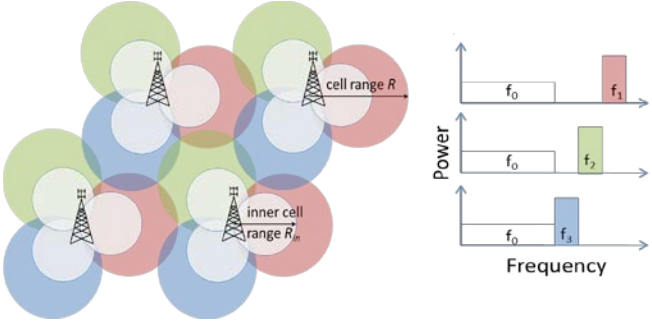


Figure 3: System Topology and frequency assignment for the reference case

The outer transmit power ($P_{x,out}$) is kept constant to 43 dBm in all the simulations, while the inner transmit power ($P_{x,in}$) is set according to the inner cell range (equivalently L_{th}) as follows:

$$P_{x,in}(dBm) = P_{x,out}(dBm) - l_A - l_B \log R(km) + L_{th} \qquad (18)$$

The rationality of this expression is to have the same average received power level for an outer user located at a distance equal to the cell range *R* and for an inner user located at a distance equal to the inner cell range R_{in}.

The rest of simulation parameters are indicated in Table I.

TABLE I. SIMULATION PARAMETERS

Simulation	Parameters
Antenna Pattern	Δ_ϕ=70º, Δ_θ=10º, B_o=20 dB, Φ=120º, Θ=0º
Shadowing Std. Deviation	10 dB
Path Loss Parameters	l_A= 128.1 dB , l_B = 37.6
Bandwidth per channel B_c	5MHz
Noise Power P_N	-100 dBm

B. Numerical Results

We evaluate our scenario under two different criteria. In the first part, an analysis of the performance of the algorithm is given according to the energy reduction (interference minimization) it provides. Similarly, in the second part, we analyze the effect of the proposed algorithm to the network and edge user capacity. Additional information will be presented related to the performance of the algorithm in terms of convergence and feasibility for real time execution.

It has to be noted that for each inner cell range the result is the average of 500 experiments with different random user distributions. For the comparison of the results we use as reference the PFR scheme presented in Figure 2.

1. *Global Energy Reduction:* Figure 4 shows the comparison of the global energy (in dBW) between the proposed solution and the reference scheme of Figure 3, where the Gibbs Sampler is not applied.

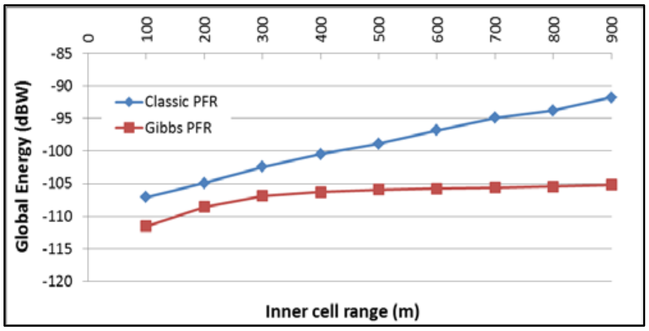


Figure 4: Global Energy

By studying the behavior of the energy, one can notice that there is a significant benefit from the execution of the algorithm, especially when considering inner cell ranges from 400 m and above. The highest gain of 13.43 dB is observed for the range of 900 m, while the average gain for all the inner cell ranges is 8.47 dB. If we focus on the reference scheme, it can be seen that the inner parts of all the cells of the network are assigned the same frequency (f_0). As such, the amount of the interference that these users experience, especially for high inner cell ranges, is quite high. This justifies the increasing behavior of the global energy when increasing the inner cell range. However, after the execution of the algorithm, the inner parts are assigned different frequencies resulting in this way in a significant interference reduction. It can also be observed that the global energy level for inner cell ranges above approximately 500m is kept at a very similar level.

2. *Capacity Improvement:* The benefits of the proposed PFR algorithm in ICI reduction are also reflected in terms of capacity improvement. This can be observed in Figure 5 that shows the average capacity for the users located at the cell edge, which are those more sensitive to ICI. For this computation, users are considered to be at the cell edge if they are located at a distance above 0.9*R* (note that edge users can be outer or inner users depending on the considered inner cell range in each simulation and also depending on shadowing conditions). As it is reflected from the figure, the edge user capacity is significantly improved compared to the classic PFR scheme. The maximum gain is observed for the inner cell range of 900 m, which reaches a capacity increase of 17%, and the average gain for all the inner cell ranges is 11.64%.

Furthermore, Figure 6 shows the comparison in terms of average capacity per user taking into consideration all the users in the cell. Similarly to the edge users, the highest gain is observed for the inner cell range of 900 m and reaches the value of 42.67%. The average gain in this case is 22.53%.

It can also be observed in Figure 5 and Figure 6 that, while the maximum average user capacity occurs for an inner cell range of 400m, when considering cell edge users, the maximum occurs for larger values. This is due to the fact that, for large inner cell ranges, in addition to the ICI reduction brought by the algorithm, the edge users share the available capacity with fewer users, since the outer area is reduced. Then, the optimal setting of the inner cell range would result from the trade-off between average and cell-edge capacity, in accordance with network operator policies.

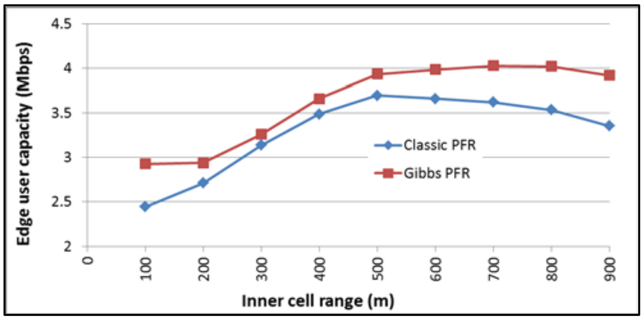


Figure 5: Average Edge User Capacity

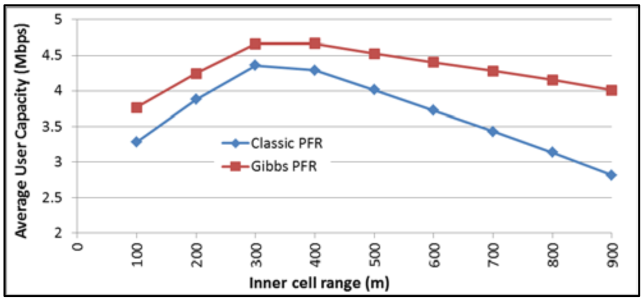


Figure 6: Average User Capacity

3. *Convergence of the algorithm:* In order to analyze the convergence of the algorithm, in this work we consider that the algorithm is executed either until a convergence criterion or the total simulation time ($12000 \cdot t_a$) is reached. The convergence criterion in this paper is that all the cells have reached a selection probability according to (16) above 0.99 for one of the possible states (then this state determines the assigned frequencies).

In Figure 7, we present the average number of the experiments that have met the convergence criterion of this paper as a result from the execution of 500 experiments for each inner cell range. It can be noticed that above the 400 m a significant amount of experiments has met the criterion. For smaller inner ranges however, it can be seen that this number very small. This was expected, since for these ranges the number of inner users is very small and in some cases there are cells with no inner users. Correspondingly, there exist actually multiple solutions that are optimal (e.g., for a cell without inner users the allocation of the inner channel does not affect the received inter-cell interference). In these situations, it has been observed that the algorithm does not converge towards a high probability for a given state but it keeps similar probability levels for all the optimum states. A similar effect is also observed for the larger cell ranges in the particular cases that the convergence criterion is not met. The algorithm keeps similar probabilities for some states that exhibit the lowest energy. This behavior reflects the good operation of the algorithm.

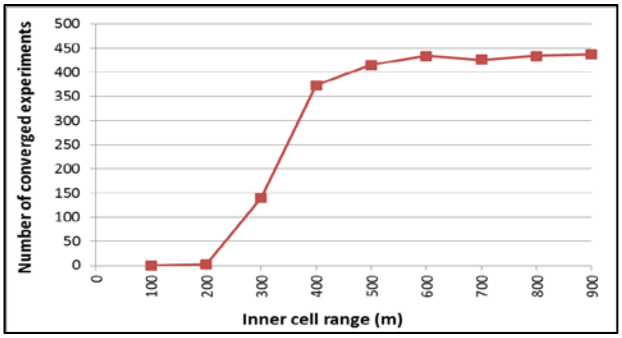


Figure 7: Number of experiments that have met the convergence criterion considered in this paper

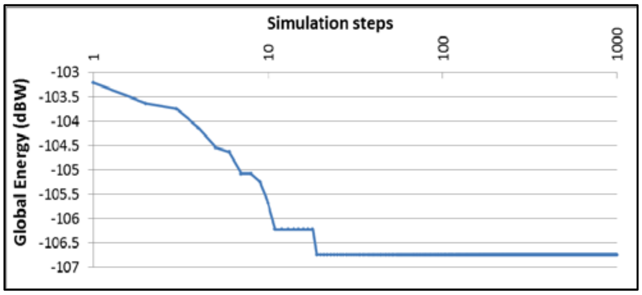


Figure 8: Global Energy Evolution

Another important aspect to evaluate the performance of the algorithm is the required number of executions in order to reach convergence. In that respect, Figure 8 presents the time evolution of the global energy of a random experiment, until the convergence criterion was met. In this particular experiment the algorithm reached convergence after 19 executions. Moreover, as it can be seen from the figure, the reduction of the energy is carried out continuously during the simulation time, suggesting this way the possibility of the online implementation of the algorithm.

V. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a new distributed algorithm based on Gibbs sampling that performs dynamic channel allocation in a PFR cellular deployment. Through simulations it has been proved that the proposed solution outperforms the classical PFR scheme in terms of interference and capacity. Results have shown the reduction of the network interference of up to 13 dB. Moreover, the proposed scheme provides a significant capacity improvement for the edge users with gains up to 17%, and up to 42.67% when considering the average capacity of all the users in the cell.

Further details about the specific implementation are envisaged as part of our future work, including the measurements involved and the information exchange between cells. Moreover, future work is envisaged to include the capacity optimization explicitly in the formulation, as well as the extension of the algorithm to adjust the transmit power of the cell. Finally, heterogeneous networks will be investigated, including macro and small cells.

ACKNOWLEDGMENT

This work has been supported by FP7 NEWCOM# project (grant number 318306) and by the Spanish Research Council and FEDER funds under ARCO grant (ref. TEC2010-15198).

REFERENCES

[1] G. Boudreau, J. Panicker, N. Guo, R Chang, N. Wang, and S. Vrzic, “Interference coordination and Cancellation in 4G Networks”, IEEE Communications Magazine, vol. 47, no. 4, April 2009, pp. 74-81.

[2] K. Begain, G.I. Rozsa, A. Pfening, and M. Telek, “Performance analysis of GSM networks with intelligent underlay-overlay”, ISCC , July 2002, pp. 135-141.

[3] 3GPP R1-050507 “Soft Frequency Reuse Scheme for UTRAN LTE”, 3GPP TSG RAN WG1 Meeting #41, May 2005.

[4] R1-050738, “Interference mitigation – Considerations and Results on Frequency Reuse”, RAN WG1#42, Aug./Sept. 2005.

[5] M. Sternad, T. Ottosson, A. Ahlen, and A. Svensson, “Attaining both coverage and high spectral efficiency with adaptive OFDM downlinks”, IEEE VTC, Oct. 2003, pp. 2486-2490.

[6] D. Lopez-Perez, A. Juttner, and J. Zhang, “Dynamic Frequency planning versus Frequency Reuse Schemes in OFDMA Networks”, IEEE VTC 69th, April 2009, pp.1-5.

[7] R. Chang, Z. Tao, J. Zhang, and C.-C. Jay Kuo, “Dynamic Fractional Frequency Reuse (D-FFR) for multicell OFDMA Networks using a graph framework”, Wirel. Commun. Mob. Comput., Jan. 2011, pp. 12-27.

[8] X. Shui, M. Zhao, P. Dong, and J. Kong, “A novel dynamic soft frequency reuse combined with power re-allocation in LTE uplinks”, WCSP, Oct. 2012, pp.1-6.

[9] G. Koudouridis, C. Qvarfordt, T. Cai, and J. Johansson, "Partial Frequency Allocation in Downlink OFDMA Based on Evolutionary Algorithms", IEEE VTC 72nd, Sep. 2010, pp.1-5.

[10] C. Kosta, A. Imran, A.U. Quddus, and R. Tafazolli, “Flexible Soft Frequency Reuse Schemes for Heterogeneous Networks (Macrocell and Femtocell)”, IEEE VTC 73rd, May 2011, pp.1-5.

[11] P. Bremaud, Markov Chains: Gibbs Field, Monte Carlo Simulation, and Queues, Springer Verlag, 1999.

[12] G. Winkler, “Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction (Stochastic Modelling and Applied Probability)”, Springer-Verlag, New York, Inc., Secaucus, NJ, USA, 1995.

[13] B. Kauffman, F. Baccelli, A. Chaintreau, V. Mhatre, K. Papagiannaki, and C. Diot, “Measurement-based Self Organization of Interfering 802.11 Wireless Access Networks”, IEEE INFOCOM, May 2007, pp.1451–1459.

[14] C.Chen and F. Baccelli, “Self - Optimization in Mobile Cellular Networks: Power Control and User Association”, IEEE ICC, May 2010, pp. 1-6.

[15] C. S. Chen, F. Baccelli, and L. Roullet, "Joint Optimization of Radio Resources in Small and Macro Cell Networks," IEEE VTC, May 2011, pp.1-5.

[16] I. Viering M. Dottling, and A. Lobinger, “A mathematical perspective of self – optimizing wireless networks”, IEEE ICC Conference, 2009, pp. 1-6.

Towards a Semantic-aware Radio Resource Management

Luis Guillermo Martinez Ballesteros, Cicek Cavdar, Pietro Lungaro, Zary Segall
Mobile Services Lab
KTH Royal Institute of Technology
Kista, Sweden
lgmb@kth.se, cavdar@kth.se, pietro@kth.se, segall@kth.se

Abstract— **In this paper semantic-aware model for radio resource management in wireless networks is introduced and studied through simulation. By semantic-awareness, the network can selectively manage the radio resource allocation based on the evaluation of transferred content, and its associated processing, and prioritize users that are close to experience interruptions, in order to improve the wireless resource utilization and the user's Quality of Experience (QoE). Different radio resource management (RRM) strategies are proposed and investigated, considering buffer capacity at the terminals and the experience of the users in time while watching a video and waiting for resource allocation. The simulation results show that the users can reduce the total duration, frequency and length of the interruptions during a playback by applying semantic-awareness in the radio resource allocation, which might affect positively user's QoE.**

Keywords-Radio Resource Management; resource allocation

I. INTRODUCTION

Recent introduction of new generation of wireless infrastructures is being accompanied by an increase in both the number of users and their interest in multimedia content. This growth has been driven in the last decade by the popularity of multimedia content (e.g. video-sharing websites, social networks, video on demand sites, mobile IPTV, etc.), that according to the tendency will generate much of the mobile traffic growth through 2016, showing, at the same time, the highest growth rate of any mobile application ([1]). Before this scenario, a common approach to reach the goal of high quality information delivery has been the implementation of resource management schemes and scheduling algorithms to optimize resource allocation and traffic distribution as function of network parameters ([2]-[13]). Solutions have evolved from a perspective mainly centred on the evaluation of network based constraints (e.g., Signal to Noise Ratio or instant data rates) deprived of knowledge about the transferred content [3], to a perspective where inherent characteristics of the content are considered to improve network performance. In some cases ([11] [12]), the video distortion level is used to calculate the rates to deliver a multimedia content in a gradient-based scheduling and resource allocation scheme. Even though these studies consider the evaluation of content status to allocate resources, their objective is to maximize the average peak signal-to-noise ratio (PSNR) of all video users, which not always impacts positively the QoE. In [13], a resource allocation scheme that considers both the average rate achieved so far and the future expected rate is proposed with the goal of maximizing sigmoid function of the average bit rate. Prediction does not consider what happens to the content in the terminal by establishing a direct relation between the bit rate and the QoE. In ([2][9]), authors improve system throughput by allocating resources according to predefined utility functions to measure the QoE and QoS respectively, without considering how the content is processed at the terminal side. However, the idea of maximizing performance through infrastructure improvements and adjustment of network parameters is usually not optimal with respect to user perceived quality for multimedia applications [2]. In this paper, we want to investigate the effect of using semantic information (i.e., buffer capacity, player data rate, waiting times) provided by users terminals on the radio resource allocation in the downlink transmissions (base station (BS) to device) in mobile networks and its impact on the user's perception. In particular, some QoE related parameters (i.e., duration, the length and the frequency of the interruptions) are quantified to provide an initial measure of the effect of incorporating semantic-awareness to wireless infrastructures.

The rest of the paper is organized as follows: In Section II, we present the semantic-aware proposed architecture and RRM schemes considered in this study. In Section III, we describe the simulation settings and performance measurements considered in the paper. In Section IV, we present the results obtained with the simulation scenario. We conclude the paper in section V.

II. SYSTEM MODEL

A. Network Description and Service Model

Semantic-aware networks are infrastructures with the capacity to selectively manage the information flow depending on its importance from an application point of view. Unlike the concept of content-awareness, where the network management is considering the type of content in a static way, semantic-aware approach requires infrastructures with the capability to exchange information dynamically with the terminal and evaluate the content related information provided by users (i.e., both specific details regarding transferred content and the status of their processing in the terminal) in order to selectively manage the information flow, and distribute resources depending on the applications performance. In this regard, two elements are essential for the operation of the semantic-aware engine: a semantic-aware resource manager (SRM) and a semantic client (SC) (Figure 1).

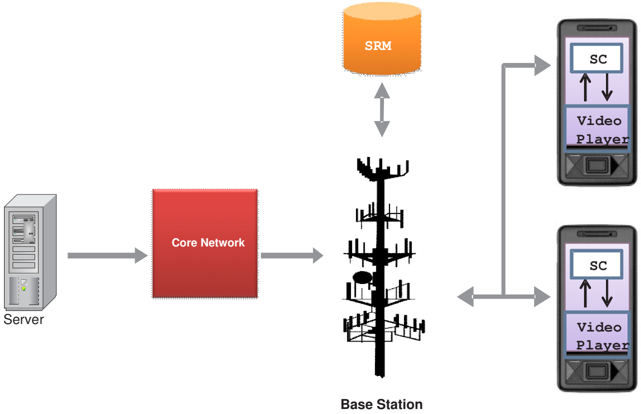


Figure 1. Example of the main elements composing the Semantic-aware system.

The SRM is centrally located in the base station and collects the reports provided by the user terminals. It keeps track of the terminals and the processing content current status. The SCs report information regarding the buffer capacity at the terminal, the player data rate consumption, and the users waiting times to the semantic-aware resource manager. SCs are software applications with collecting and sensing functionalities installed in the mobile terminals. Once the gathered data is passed to the SRM, this selects the time instants more appropriate for allocating resources and delivering content to individual users according to the operator's goal, by applying a scheduling policy.

At the fixed side of the infrastructure, the BS is connected to a multimedia server with capacity to store different multimedia files of size S_v bits. With regard to the wireless side of the architecture, we assume that the total user population of size N_u is uniformly distributed and downloading multimedia files, v. Each user $l \in \{1,\dots, N_u\}$ is interested in receiving a maximum of M items, all of them with size S_v and duration in seconds T_v. Once a user l requests content, information will be downloaded at an instantaneous data rate from the BS to the user l at time t, $R_l(t)$. Downloaded information is placed into a buffer B_l of infinite capacity before being effectively consumed by user l. An initial buffering time of b seconds is considered. This time b, counted only once by a multimedia file, corresponds to the interval between the first request time and the time at which the effective consumption of bits by user l from the buffer B_l starts. Once the time b has elapsed, the playback starts and bits from the buffer B_l will be consumed at time t with a data consumption rate C_l bps. C_l is a constant value that depends on the size and the duration of time that needs to be spent to watch the multimedia file v requested by user l, S^l_v and T^l_v respectively. So,

$$C_l = \frac{S^l_v}{T^l_v} \tag{1}$$

Content processing will continue until the file stored in the server has been consumed by user l. Duration of processing one multimedia file will determine how long users will last in the system, and the request for a new multimedia file will stay active until M multimedia files has been processed by user l.

B. *Semantic-aware Resource Management Schemes*

In our implementation, all proposed semantic-aware schemes follow a similar procedure to assign resources every slotted time interval n of duration Δt. The SRM starts by detecting how close the users are of experiencing a lack of resources in the buffer B_l that can affect the correct processing of the information and the user perception. Proximity of shortage is measured in terms of the video watching time left in the buffer at time t, T_b seconds, given the $C_l(t)$ rate. So,

$$T_b = \frac{B_l(t)}{C_l} \tag{2}$$

This identification process based on the evaluation of the T_b value lead to a classification of the users in two queues, one with those users with imminent shortage and other filler with those with no imminent shortage, called X and Y, respectively. If size of queue X is equal to one, the user l in that queue receives the resource with no consideration of the users present in the queue Y. In contrast, if the length of the queue X is more than one, users in the queue will be ranked in descending order considering the utility function of the RRM scheme. Then the scheduler allocates the resource to the user in the top of the ranking. If queue X is empty, the procedure described before will be executed with the users placed in the queue Y. Supported by T_b other values extracted from the semantic information, the scheduler will look first at the users of X group. If size of X is equal to one, that user l receives the resource. In contrast, if there is more than one in the set X, users will be ranked in descending utility order considering the criteria of the semantic aware RRM scheme. Then the scheduler allocates the resource to the user in the top of the ranking. Different RRM used in this study, including the reference case, and the utility functions linked to them are described below:

1) Proportional Fair (PF): This RRM strategy represents the reference case or no semantic-aware scheme. In this case, users are not classified according to T_b value. By contrast, there is no buffer capacity consideration. At each time interval n, this scheme assigns the resource to the user with the largest ratio $\left({}^{R_l}/_{A_l}\right)$ where R_l is the instant download data rate achievable by user l and A_l is the average download data rate of user l.

2) Buffer based (BB): This RRM strategy tries to allocate resources to the user with the smallest video watching time from the buffer based on the evaluation of T_b by considering current buffer capacity. This scheduler tries to allocate a time slot to the user with the higher imminence of having an interruption. Utility function for grading the users is the inverse of the T_b, $\left({}^{1}/_{T_b}\right)$.

ISBN: 978-1-61208-313-1

3) *Inactive online time based (IB):* This RRM uses the time a user has been active in the system but with no wireless resource assigned to download bits. We evaluate the total time a user *l* has been active in the system, with outstanding bits in the server to download, but without a wireless resource assigned to place bits in the buffer or T^{l}_{wait} in seconds. Utility function for grading the users is the current $\left(1/_{T^{l}_{wait}} \right)$.

4) *Active online time based (AB):* This scheduler looks at the time a user has been selected by the scheduler while it is consuming bits from the buffer B_l or T^{l}_{dwld}. Utility function for grading the users is the current $\left(T^{l}_{dwld} \right)$.

5) *Mixed criteria based (MB)*: In this RRM the utility function for the user *l* is computed as the sum of the individual values T^{l}_{wait} and T^{l}_{dwld} divided by the value of T^{l}_{b}.

III. Investigation

A. Simulation Settings

To investigate the performance of using semantic-aware RRM in a wireless infrastructure we performed extensive simulations of an HSDPA network focusing on the downlink connection between one 3-sector BS, with 300m of cell radius, and the user devices requesting for the streaming of content stored in a multimedia server. Propagation model is the 3GPP model, where path loss is $L = 128.1 + 37.6\ log_{10}\ (R), R\ in\ Km$. We assumed that the backbone is lossless and the transmission delay from the media server to the BS is negligible. Maximum BS transmission power $\widehat{P_T} = 20W$, and maximum data rate of 14.4 Mbps. One user is allocated in a time slot of 0.25s. The basic system level assumptions used in the simulations are summarized in Table I. In our system, we played with users densities ranging between 5 and 25 users. In each case the user requests of multimedia content have been modelled with expected inter-arrival time equal to 1 minute. In all cases, users are supposed to pick a video of 57.76 Mbytes, representing a 1080p YouTube-like video of 5 minutes duration. In the simulation, users will watch up to 20 videos of the same size and resolution (homogeneous scenario), being a realization the completion of this number of videos by all users present in the system.

B. Performance Measures

From the operators' perspective, performance is evaluated considering the average Total Duration of Interruption (TDI) obtained by using each one of the proposed schedulers. The maximum and minimum length of interruptions represents the values of the longest and shortest interruptions experienced by user *l* during the playback. Finally, the average frequency of the interruptions will represent how often a user *l* can experience cuts during the playback.

IV. Results

Figure 2 demonstrates the average TDI comparison of different RRM schemes for different number of users watching HD videos. With the implementation of the MB scheme there is reduction in the average TDI during the streaming session compared to the PF scheme. Reduction goes from 74%(with 5 users) to 8% (with 25 users). The other scheme that looks at the buffer capacity (BB) also guarantees a reduction in the TDI that goes between 53% (5 users) to 3% (25 users) regarding the PF. In contrast, schemes focused on the evaluation of online times (inactive/active) reduce the performance of the system increasing TDI.

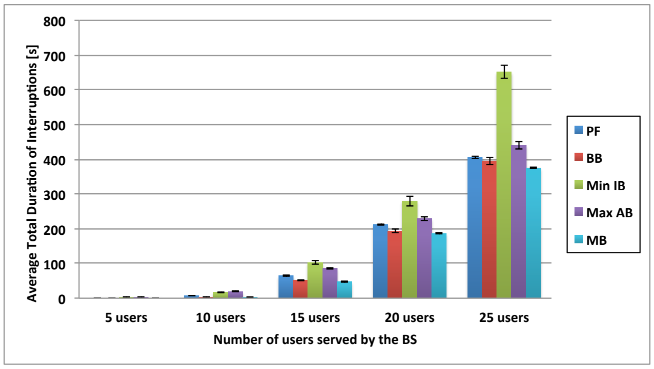


Figure 2. Average TDI by number of user for the different resource schedulers in a HD scenario. Error bars indicate 95% confidence intervals.

Figure 3 shows the length of the shortest interruption experienced by different number of users when the proposed RRM schemes are used. The figure reflects that schemes PF, MB and BB can guarantee that the shortest interruption in any case will be less than 1.3s. Schedulers that do not consider T_b in its allocation criteria will generate as shortest interruption duration values between 2s to 400s, which will generate a negative impact on the QoE perceived by users in a real scenario. Figure 4 shows the average maximum length of interruptions perceived by different number of users with different RRM schemes. In this case network reflects a better performance when PF, BB and MB are used as resource allocation schemes. Although PF shows the best performance when maximum length of interruptions is considered, observing the frequency of the interruptions in Figure 5, PF shows around 25% more interruptions during the playback than the best of the other considered RRM schemes. This recurrence in the number of interruptions will affect more the user's perceived quality, according to the results obtained in [14]. In summary, semantic-aware schemes that use buffer-related information will reduce the TDI through a reduction in the frequency and the length of interruptions.

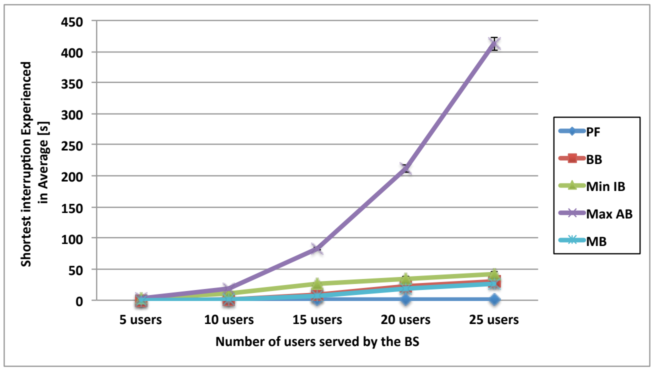


Figure 3. Average minimum length of interruptions by number of users for the different resource schedulers in a HD scenario.

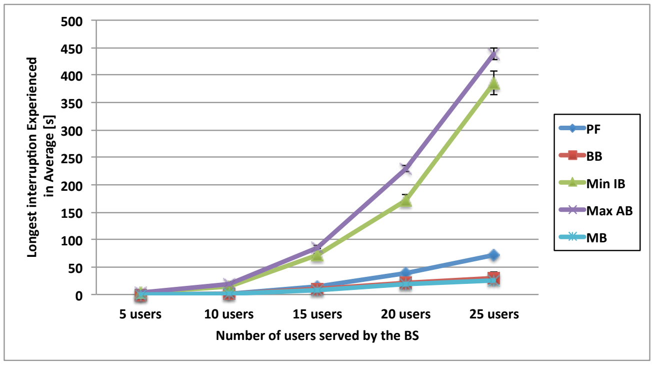


Figure 4. Average maximum length of interruptions by number of users for the different resource schedulers in a HD scenario.

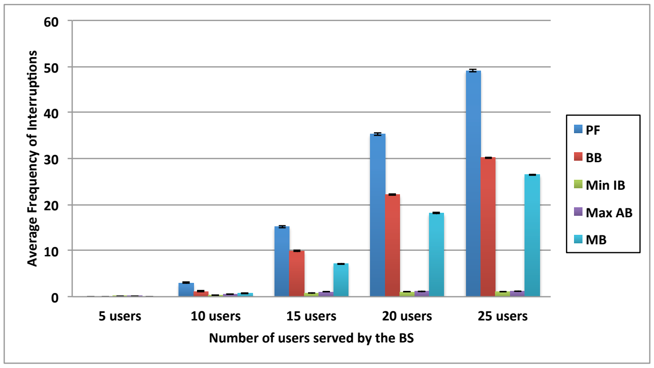


Figure 5. Average frequency of interruptions by number of users for the different resource scheduler in a HD scenario.

V. CONCLUSION

The concept of Semantic Radio Resource Management has been introduced in this paper as an alternative to improve the user service perception in video streaming services. The considered solution simply requires the introduction of software agents both at the network and terminal side, capable of monitoring applications behaviours. Different RRM strategies were simulated and results show that by using semantic-aware schemes evaluating user's buffer capacity, it is possible to improve the total duration of video stalling, and impact the length and frequency of the interruptions users can experience during the video playback. This indicates a potential of proposed solution to generate improvements in terms of the final QoE perceived by a user in comparison to the ”classical” RRM. As future work, the extension of the proposed scheme considering more semantic elements to make resource allocation decisions is planned.

REFERENCES

[1] Cisco, “Cisco visual networking index: Global mobile data traffic forecast update, 2011–2016.” Cisco, White Paper, 2012.

[2] S. Thakolsri, S. Khan, E. Steinbach, and W. Kellerer, “Qoe-driven crosslayer optimization for high speed downlink packet access,” Journal of Communications, Special Issue on Multimedia Communications,Networking and Applications, Vol. 4, No. 9., 2009, pp. 669–680.

[3] H. Yin and H. Liu, “An efficient multiuser loading algorithm for ofdm-based broadband wireless systems,” in Global Telecommunications Conference, 2000. GLOBECOM ’00. IEEE, vol. 1, 2000, pp. 103 –107.

[4] G. Aristomenopoulos, T. Kastrinogiannis, V. Kaldanis, G. Karantonis, and S. Papavassiliou, “A novel framework for dynamic utility-based qoe provisioning in wireless networks,” GLOBECOM 2010, 2010 IEEE Global Telecommunications Conference, pp. 1–6, 2010.

[5] K. Piamrat, C. Viho, J.-M. Bonnin, and A. Ksentini, “Quality of experience measurements for video streaming over wireless networks,” Information Technology: New Generations, 2009. ITNG ’09. Sixth International Conference on, pp. 1184 – 1189, 2009.

[6] S. Rabiul Islam and M. Hossain, “A wireless video streaming system based on ofdma with multi-layer h.264 coding and adaptive radioresource allocation,” in Image Information Processing (ICIIP), 2011 International Conference on, nov. 2011, pp. 1 –6.

[7] M. Shehada, S. Thakolsri, Z. Despotovic, and W. Kellerer, “Qoe-based cross-layer optimization for video delivery in long term evolution mobile networks,” in Wireless Personal Multimedia Communications (WPMC), 2011 14th International Symposium on, oct. 2011, pp. 1 –5.

[8] P. Dutta, A. Seetharam, V. Arya, M. Chetlur, S. Kalyanaraman, and J. Kurose, “On managing quality of experience of multiple video streams in wireless networks,” in INFOCOM, 2012 Proceedings IEEE, march 2012, pp. 1242 –1250.

[9] S.-P. Chuah, Z. Chen, and Y.-P. Tan, “Energy-efficient resource allocation and scheduling for multicast of scalable video over wireless networks,” Multimedia, IEEE Transactions on, vol. 14, no. 4, aug. 2012, pp. 1324–1336.

[10] H. Adibah Mohd Ramli, K. Sandrasegaran, R. Basukala, R. Patachaianand, M. Xue, and C.-C. Lin, “Resource allocation technique for video streaming applications in the lte system,” in Wireless and Optical Communications Conference (WOCC), 2010 19th Annual, may 2010, pp. 1 –5.

[11] P. Pahalawatta, T. Pappas, R. Berry, and A. Katsaggelos, “Content-aware resource allocation for scalable video transmission to multiple users over a wireless network,” in Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, vol. 1, april 2007, pp. I–853 –I–856.

[12] X. Ji, J. Huang, M. Chiang, G. Lafruit, and F. Catthoor, “Scheduling and resource allocation for svc streaming over ofdm downlink systems,” Circuits and Systems for Video Technology, IEEE Transactions on, vol. 19, no. 10, , oct. 2009, pp. 1549 –1555.

[13] C. Yang and S. Jordan, “Power and rate allocation for video conferencing in cellular networks,” in Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on, sept. 2011, pp. 127 –134.

[14] Acision, “Seizing the opportunity in mobile broadband - a global perspective,” Acision, Tech. Rep., 2011.

A3-Based Measurements and Handover Model for NS-3 LTE

Budiarto Herman, Dmitry Petrov, Jani Puttonen, Janne Kurjenniemi
Magister Solutions Ltd.
Jyväskylä, Finland
Emails: {budiarto.herman, dmitry.petrov, jani.puttonen, janne.kurjenniemi}@magister.fi

***Abstract*—This paper presents a Long Term Evolution (LTE) handover algorithm implementation based on Reference Signal Received Power (RSRP) measurements and Event A3 on top of the LTE module of NS-3 network simulator. Many simulation scenarios in various research projects rely on user mobility. However, until recently complete realisation of relevant functionality has been missing in free and open source tools such as NS-3. Detailed modeling of RSRP measurements, including sliding window averaging, time-to-trigger and hysteresis evaluations are considered in this paper. Received simulation results are verified through comparison with other publication, suggesting a promising direction for further studies of dynamic scenarios.**

Long term evolution; handover; simulation; ns-3

I. INTRODUCTION

Long Term Evolution (LTE) is a standard for radio technology developed by the 3rd Generation Partnership Project (3GPP). It is a significant change from the previous 3G/UMTS technology, providing higher data rates, simplified network architecture, and improved user mobility.

Simulation of a realistic LTE wireless network requires User Equipment (UE) mobility modeling, including UE measurements and handover. However, detailed modeling of this procedure has only been found in commercial or proprietary LTE simulators. This adds considerable cost to academia and researchers who are interested in the subject. Moreover, the source code of these simulators are not publicly available, imposing difficulties for professional community to study and verify the produced simulation results.

At least three prominent free and open source LTE simulators have been available at the moment of publication. Each of them has its own limitations in modeling user mobility. For instance, link and system level simulators from University of Vienna [1] have not modeled mobility. LTE-Sim [2] has included a modeling of handover procedure, but it has been based on Signal-to-Interference-Ratio (SIR) and location, which is not in accord with 3GPP specification. The LTE module of NS-3 has been developed within LENA project [3] and its recent development version is featuring a handover algorithm based on Reference Signal Received Quality (RSRQ) measurements and Event A2. This feature, however, has not been revealed in the stable release of NS-3 because it is still in the development phase.

We have extended the handover modeling in NS-3 by utilizing Reference Signal Received Power (RSRP) measurements and Event A3, as designed in [4]. In the case of RSRP as the measurement of choice, *Event A3* is defined as a reporting triggering event which is fired when there exists a neighbouring cell which measured RSRP is better than measured RSRP of the serving cell by certain offset [5]. The aim of this extension is to develop a measurements and handover model according to 3GPP specification, thus enabling detailed mobility studies with NS-3.

Moreover, the functionality we are introducing into the simulator is vital for Self-Organizing Networks (SON) and Cognitive Networks (CN), which have been under intense study in recent years. Many of SON and CN algorithms depend on information from Radio Access Network (RAN) signal levels and coverage. In addition, according to the 3GPP specification, UE must have the ability to provide RSRP and RSRQ measurements in Evolved UTRAN LTE [6]. Therefore, the developed measurement and reporting mechanisms can be used in the future for the study of SON and CN features.

LTE is based on distributed architecture, where Evolved NodeBs (eNodeB) are responsible for handover decision. Handover algorithms have serious impact on the cellular network performance. Taking into account that LTE is targeted to operate in different propagation environments, UE-based measurements should be carefully studied to enable robust handover in wide range of UE speed. Furthermore, an optimal tradeoff between number of handovers (signaling load, amount of connection disruptions) and signal quality in large realistic scenarios should be found. This requires utilisation of system level simulators with detailed implementation of UE measurement procedure.

The rest of the paper is organised as follow. Section II elaborates on the design of RSRP- and Event A3-based handover model developed for NS-3. In order to test the model, several simulation scenarios have been studied, as described in Section III. Finally, Section IV presents the conclusion and several ideas for future research.

II. HANDOVER MODELING

In contrast with 3G/UMTS standard, handover in LTE is specified as hard handover or “break-before-connect”. It is a UE-assisted and eNodeB-triggered procedure [7]. The handover model considered in this paper is based on this specification, and is presented in Figure 1 as a series of operations and message exchanges between UE, source eNodeB, and target eNodeB. Several steps of the procedure have been already provided out-of-the-box by NS-3. This paper is reusing this functionality while focusing on the first stages, which include the *measurement reports* and *handover decision*.

ISBN: 978-1-61208-313-1

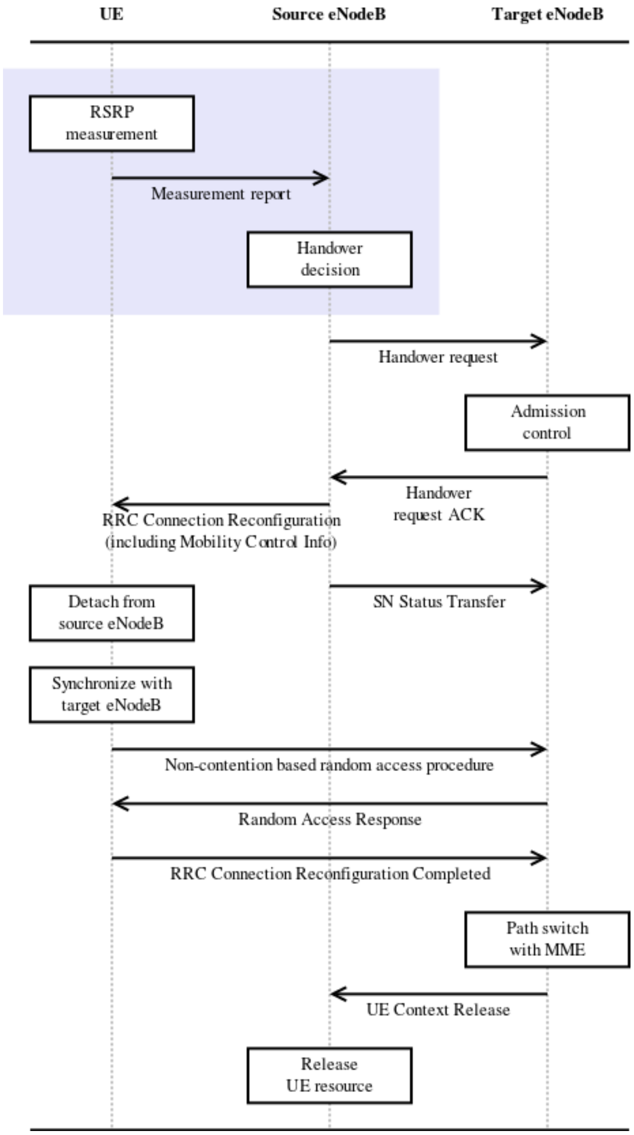


Fig. 1. Modeling of handover procedure in NS-3, where the shaded box indicates the part studied in this paper.

Measurement reports in the model are generated as follows. UE makes periodical measurements of RSRP at every time period T_m from each identified cell over the whole bandwidth. These measurement samples are then forwarded from the physical layer (PHY) (*RSRP ChunkProcessor* in Figure 2) to the Radio Resource Control (RRC) layer. RRC applies time averaging to the measurements from every specific cell (*SlidingWindow*). Sliding window always holds the averaged value from $n = \frac{T_f}{T_m}$ measurements within the time window T_f. Thus, every time a new measurement sample comes, the oldest one is discarded. The objective of this averaging is to reduce the influence of channel fading component on RSRP measurements. As a result, the rate of ping-pong handovers in the system is expected to decline.

In event-triggered handover procedure, each UE evaluates the Event A3 condition every time a new averaged measurement sample is available (*A3Evaluator*). The evaluated condition is the entering condition of Event A3: whether the RSRP measured from a neighbouring cell becomes an offset better than the RSRP measured from the serving cell [5]. The offset is represented as *hysteresis*. This condition must stay true for at least a certain duration of time, which is called the *time-to-trigger*.

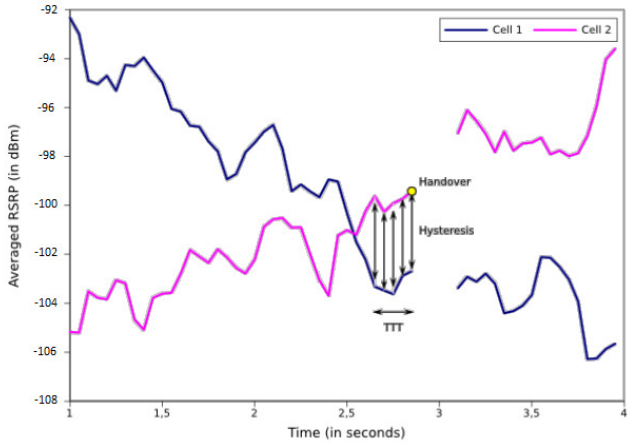


Fig. 3. Sample RSRP trace.

TABLE I. VARIABLE PARAMETERS

Parameter	Values
Time-to-trigger	50, 200, 400 ms
Hysteresis	1, 3, 6 dB
UE speed	3, 30 kmph
Sliding window size (T_f)	200, 400 ms

The actual Event A3 is triggered immediately after the time-to-trigger. The UE generates a measurement report and transmit it as an RRC message to the serving cell. This report typically contains measurement results of at least the serving cell, but is extendable with measurement results of neighboring cells. The whole process is demonstrated in Figure 3, which shows the trace of averaged RSRP measurements from serving *Cell1* and neighbouring *Cell2* before and after handover.

In practice, the eNodeB is responsible for deciding whether or not a handover is needed. In our case, we assume that an Event A3-triggered measurement report indicates that handover is really needed. Upon receiving this report, source cell immediately prepares a handover to the target cell. The rest of the handover procedure is performed as illustrated in Figure 1.

III. SIMULATION RESULTS

We conducted a simulation campaign in order to validate the developed measurement and handover model. The main focus of the study was on confirming whether the available handover-related parameters, shown in Table I, behave as theoretically expected. Simulation assumptions were loosely based on 3GPP case 1 [8], as summarised in Table II.

The number of handovers and number of ping-pong handovers were the metrics collected from each simulation. The simulation regarded ping-pong handover as two consecutive handovers by a UE, which occurred within a short period of time (in this particular case, 2 s), provided that the first one is a handover from cell A to cell B, while the second one is from

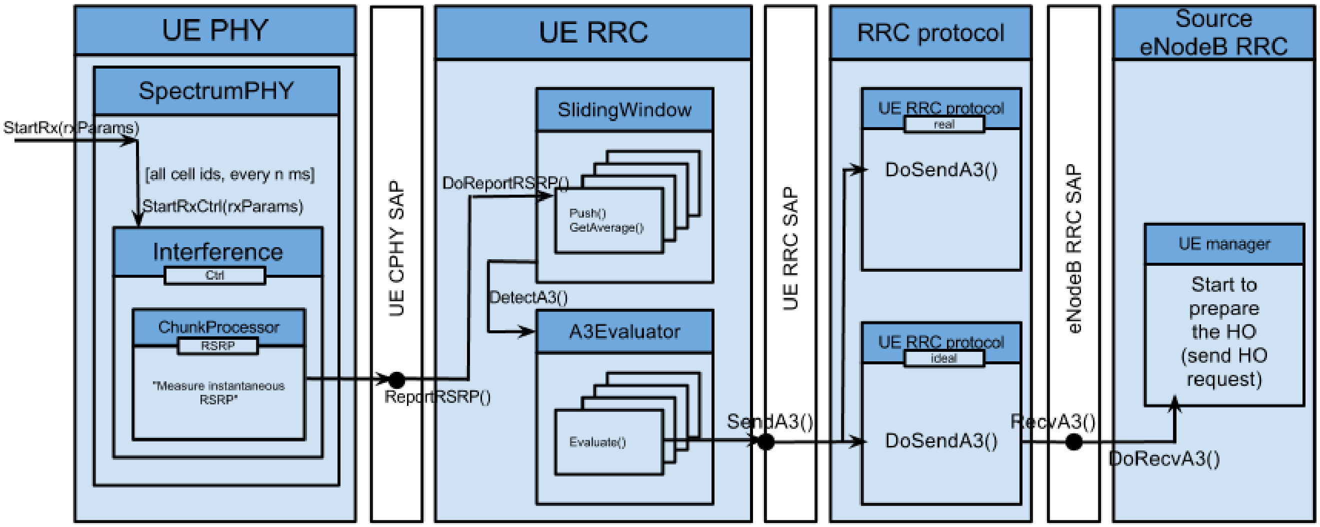


Fig. 2. Implementation of measurements and handover model.

TABLE II. SIMULATION PARAMETERS

Parameter	Value
Cellular layout	7 three-sectored sites in hexagonal layout (21 cells in total)
Inter-site distance	500 m
Cell Tx power	30 dBm
Path loss model	$L = 128.1 + 37.6 \cdot \log_{10} R$
Channel fading	Typical urban
Carrier frequency	2 GHz
System bandwidth	5 MHz (25 RBs)
Traffic	Only control messages, no data traffic
Error model	None
UE distribution	10 UEs distributed randomly in front of each eNodeB (210 UEs in total)
UE movement pattern	Constant speed to random direction (changing direction every 5 s)
UE measurement interval (T_m)	50 ms
Simulation duration	70 s

cell B back to cell A. Consequently, a series of handovers within cells A–B–A–B counts as two ping-pongs.

Figure 4 and 5 show the average number of handovers per user per second for UE speed of 3 kmph and 30 kmph respectively. It is obvious that the increase in hysteresis parameter value significantly reduces the number of handovers. This behaviour was also observed in [9].

The number of handovers is also sensitive to time-to-trigger. As seen from the same figures, the increase of time-to-trigger has the effect of reducing the number of handovers. [9] and [10] observed the same behaviour.

Time-to-trigger variation is also known as one of the means to manipulate the number of ping-pong handovers [9]. This is demonstrated in Figure 6, which depicts the proportion of ping-pongs over the total number of handovers in the simulation.

Filtering period has also been identified as a parameter for tuning handover rates. Fast moving UE typically requires shorter filtering period than slow moving UE. It has been confirmed that longer filtering period reduces the number of handovers [11]. In our study, the sliding window size T_f plays the same role as filtering period, and its effect to handover rate is presented in Figure 7.

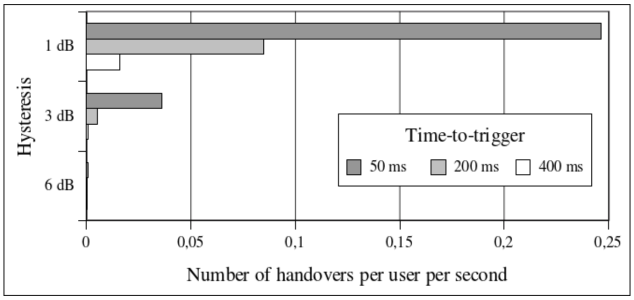


Fig. 4. Simulation results with UE speed of 3 kmph and sliding window size T_f of 200 ms.

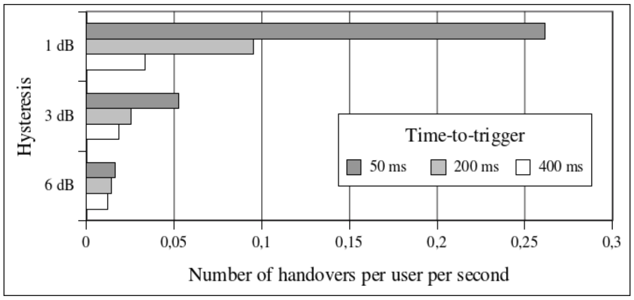


Fig. 5. Simulation results with UE speed of 30 kmph and sliding window size T_f of 200 ms.

IV. CONCLUSION AND FUTURE WORK

In this paper, we have described the measurement and handover modeling on top of NS-3 LTE module. Simulation results have been presented in order to verify the implementation. The effect of handover parameters such as hysteresis, time-to-trigger, and filtering can be clearly seen from the results. For

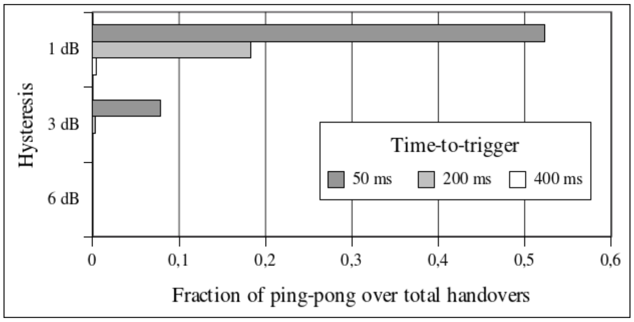


Fig. 6. Rate of ping-pong handovers in simulations with UE speed of 30 kmph and sliding window size T_f of 200 ms.

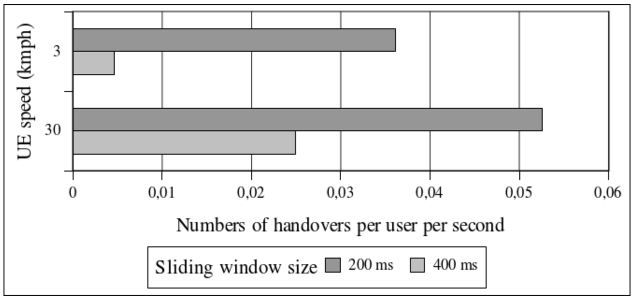


Fig. 7. Effect of T_f to handover rate in simulations with hysteresis of 3 dB and time-to-trigger of 50 ms.

instance, simulation runs with small hysteresis and short time-to-trigger produced large number of handovers, especially of ping-pong type. The number of handovers can be substantially reduced with proper parameterisation. This proves that the simulated behaviour demonstrated by our model is sensitive to these parameters and in accord with results from several other published research works in the field. In the future, it will also be possible to conduct a comparison study of our results with alternative realisation of measurement and handover model, which is under development in NS-3.

It is necessary to mention that we have not yet considered some of the important mobility related statistics in LTE. In particular, features such as Radio Link Failure (RLF) and proper modeling of handover failures provide important input for SON and CN studies. For example, sleeping cell detection relies on analysis of UE measurements and RLF occurrences in the network [12]. We are aiming to incorporate RLF into NS-3 LTE module on top of the measurement and handover models presented in this paper, which altogether will be utilised in our CN research.

References

[1] C. Mehlführer, J. Colom Ikuno, M. Šimko, S. Schwarz, M. Wrulich, and M. Rupp, "The Vienna LTE simulators - enabling reproducibility in wireless communications research," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–14, 2011.

[2] G. Piro, L. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE cellular systems: An open-source framework," *Vehicular Technology, IEEE Transactions on*, vol. 60, no. 2, pp. 498–513, 2011.

[3] N. Baldo, M. Miozzo, M. Requena-Esteso, and J. Nin-Guerrero, "An open source product-oriented LTE network simulator based on ns-3," in *Proceedings of the 14th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems*, ser. MSWiM '11. New York, NY, USA: ACM, 2011, pp. 293–298.

[4] K. Dimou, M. Wang, Y. Yang, M. Kazmi, A. Larmo, J. Pettersson, W. Muller, and Y. Timner, "Handover within 3GPP LTE: Design principles and performance," in *Vehicular Technology Conference Fall (VTC 2009-Fall), 2009 IEEE 70th*, 2009, pp. 1–5.

[5] *LTE; E-UTRA; RRC; Protocol specification (release 11)*, 3GPP Std. TS 36.331, 2013.

[6] *LTE; E-UTRA; Physical layer; Measurements (release 11)*, 3GPP Std. TS 36.214, 2013.

[7] *LTE; E-UTRA and E-UTRAN; Overall description; Stage 2 (release 11)*, 3GPP Std. TS 36.300, 2013.

[8] *E-UTRA; Further advancements for E-UTRA physical layer aspects (release 9)*, 3GPP Std. TR 36.814, 2010.

[9] M. Anas, F. Calabrese, P. Mogensen, C. Rosa, and K. Pedersen, "Performance evaluation of received signal strength based hard handover for UTRAN LTE," in *Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th*, 2007, pp. 1046–1050.

[10] C.-C. Lin, K. Sandrasegaran, H. A. M. Ramli, R. Basukala, R. Patachaianand, L. Chen, and T. Afrin, "Optimization of handover algorithms in 3GPP long term evolution system," in *Modeling, Simulation and Applied Optimization (ICMSAO), 2011 4th International Conference on*, 2011, pp. 1–5.

[11] M. Anas, F. Calabrese, P.-E. Ostling, K. Pedersen, and P. Mogensen, "Performance analysis of handover measurements and layer 3 filtering for UTRAN LTE," in *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on*, 2007, pp. 1–5.

[12] F. Chernogorov, J. Turkka, T. Ristaniemi, and A. Averbuch, "Detection of sleeping cells in LTE networks using diffusion maps," in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, 2011, pp. 1–5.

Towards an Integrated Mobility Simulation for Communications Research

Kamill Panitzek*, Pascal Weisenburger†, Immanuel Schweizer* and Max Mühlhäuser*
*Telecooperation Lab, †Department of Computer Science
Technische Universität Darmstadt
Darmstadt, Germany
e-mail: {panitzek@tk, pascal_w@rbg, schweizer@tk, max@tk}.informatik.tu-darmstadt.de

***Abstract*—Mobile devices and new communication technologies gain ever more importance. Evaluating such technologies for critical domains, like disaster recovery, is a difficult task and is usually done by using simulations. Until today, research focused on the impact of mobility patterns on the investigated communication technology. We argue that the influence of the communication technology on the mobility patterns is also important to produce realistic simulations and meaningful results. In this paper, we present our agent-based mobility model and our simulation framework. This framework can be connected to a network simulator to execute mobility and network simulations in parallel, thus, influencing each other. The evaluation of our mobility model with real world experts in the field of disaster recovery indicated our approach to be accurate and also revealed potential for improvements.**

***Keywords*—mobility model; behavior-based; software agents; simulation; integration; disaster recovery**

I. Introduction

Simulators for mobility and communication are popular research tools to investigate the applicability of new communication technologies under dynamic conditions. Especially in critical domains, like disaster recovery, new communication technologies can help saving human lives more efficiently. For example, a scalable communication network can be achieved by interconnecting hand-held mobile devices carried around by first responders to a peer-to-peer like network as proposed by Bradler et al. [4]. However, testing such prototypes during real disaster recovery missions is very dangerous and negligent because, if the prototype fails during the mission, human lives are threatened. Also, the testing possibilities under real conditions are very limited and results might not be easily reproducible due to many factors influencing the investigations. Using simulations to investigate new communication technologies is, therefore, much safer and more convenient.

Until today, the focus in communications research usually resided on the new communication technology under investigation. Random-based mobility models are commonly used to produce movement traces [6], [13] in a first step. Afterward, these movement traces are used as input data for network simulations [1]. We believe this approach to be insufficient, at least for the scenario of disaster recovery, since communication between first responders directly influences their movement. For example, if the network suffers a malfunction and, thus, new commands from the headquarter cannot be transmitted to first responders in the field, they cannot react on these commands. Conversely, if first responders communicate using wireless network devices, their movement directly influences the network since messages cannot be transmitted if first responders move out of each other's communication range.

We, therefore, propose a framework for mobility simulation to execute mobility simulations together with network simulations in parallel, enabling them to influence each other. Our mobility model is based on software-agents allowing us to simulate different kinds of scenarios by implementing different agent types and scenario specific rule sets. We implemented rule sets for different disaster relief forces because of our expert knowledge and experience in this field. Our simulation framework is targeted on interaction between agents and the environment as well as on communication between agents. These interactions are influencing their movement directly. We evaluated our approach using questionnaires targeted on real world experts most of which already participated in rescue missions. Our evaluation showed that, considering the level of abstraction, most of our assumptions are accurate and our approach is suitable in general. However, we were able to identify some weak spots in our mobility model which we fixed after evaluating the survey. The contributions of this paper are threefold:

1) we present our mobility and interaction model that generates realistic movement and communication patterns for simulations,
2) we describe the general architecture of our framework for integrated mobility and network simulations, and
3) we discuss the evaluation of our model using questionnaires targeted on real world first responders.

Our mobility model and simulation framework enable us to investigate the influence of different communication technologies on the efficiency of first responders during their missions in the future.

The remainder of this paper is organized as follows: in Section II, we discuss the related work on mobility models and focus on models targeted on disaster recovery in more detail. We then present our mobility model and our integrated simulation framework in Section III. The evaluation of our mobility model is discussed in Section IV and Section V presents concluding remarks as well as an overview on future work.

II. Related Work

To simulate the movement of nodes in a mobile ad-hoc network (MANET) scenario, many mobility models have been proposed during the last two decades. In general, mobility models can be separated into two categories: trace-based models and synthetic models [6]. In trace-based models, movement traces are gathered and collected from real world systems. These traces are then replayed in simulations to correctly simulate mobility of network nodes. However, they are bound to the specific environment where they were gathered. Also, these models are not flexible enough to simulate variations of a single scenario.

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-313-1

To simulate other environments or different variations of a single scenario, synthetic models can be used. These models create artificial movement traces for mobile nodes without replaying recorded real world movement traces. However, these models are supposed to create movement traces similar to real world traces and, thus, simulate realistic movement patterns. To accomplish this goal the behavior of single nodes is reproduced in a realistic fashion.

We categorize the approaches for synthetic mobility models into random mobility models and behavior-based mobility models.

A. Random Mobility

Until today, many mobility models based on random movement have been proposed. Although these models are rather simple, they are particularly popular and have been examined extensively in the past [12]. The three most commonly used mobility models that generate random movement traces are:

- *Random Walk Model [7].* In this model, every node picks a direction of the interval $[0, 2\pi)$ at random and a random velocity. Then, the node moves for a random time span before it picks a new direction and a new velocity.
- *Random Direction Model [13].* Here, every node picks a random direction of the interval $[0, 2\pi)$ and a random velocity. The node then moves to the border of the simulation environment where it picks a new direction and velocity.
- *Random Waypoint Model [10].* In this model, every node picks a random point in the simulation environment as well as a random velocity. It then moves until the point is reach and picks a new waypoint and velocity.

Although these simple random mobility models above are commonly used they have different unexpected properties and create a behavior that is not usually intended. Yoon et al. [17] showed that in these models the average movement speed of nodes decreases over time because slow nodes need more time to reach their destination than faster nodes. Furthermore, mobility models exist that try to incorporate a behavior oriented component into the otherwise random models:

- *Pursue Model [14].* Using this mobility model, nodes can be simulated that follow a specific moving target defining the movement as: $p_{new} = p_{old} + a(p_{target} - p_{old}) + v_{random}$. The vector v_{random} defines the influence of the random number generator on the node movement. The acceleration $a(x)$ describes the influence of the moving target on the node movement.
- *Column Model [14].* In this mobility model, nodes move on predefined lines. These lines move forward toward a random direction from the interval $[0, \pi)$ and a random distance. This model is suited for search-like activities where nodes move forward forming a front line.
- *Reference Point Group Model [9].* This mobility model arranges nodes into groups. The center of a group moves on a random path and the group members move randomly around a predefined reference point which depends on the logical group center.

A more detailed overview and a simulation based comparison of different mobility models can also be found in the work of Camp et al. [6].

B. Behavior-Based Mobility

Our main goal is to reproduce real world mobility in a simulated environment. As we look on disaster recovery missions as a concrete scenario it is clear, that a mobility model based on random movement cannot reproduce real world movement and behavior of first responders during rescue missions. Especially the different organizations like police, fire fighters and paramedics have very specific roles when entering a disaster area [12]. Therefore, a mobility model is needed that incorporates these different roles so that the nodes move and act according to their specific behavior. We now highlight three types of behavior-based mobility models capable of reproducing real world behavior and movement of first responders.

1) Role-Based Mobility: Nelson et al. describe a generic event and role-based mobility model [12]. Every node is assigned a role or a set of roles generating actions the node will perform on a given event. The entire movement pattern of a node in disaster recovery is then described by a triple (r, e, a): role r reacts on an event e by performing the activity a. By instantiating the triples with the characteristics for different node types operating in disaster recovery a movement pattern for the scenario can be generated. This way it is also possible that a node follows different movement patterns during one simulation. Four different categories of actions are assumed:

- *Repelling*: This role is mainly used for civilians during a disaster. Also, this role can include a property describing the curiosity of the civilian defining the possibility the civilian stops at the periphery of a disaster area, thus, simulating watchers.
- *Attracting*: This role is typically used for police men and fire fighters moving fast towards one or more events.
- *Oscillating*: This role is mainly used to model ambulances moving between the disaster area and hospitals. The nodes move towards an event and directly after arriving there they move to a predefined location and repeat the movement pattern continuously.
- *Immobile*: This role is used to model naturally static objects like hospitals. But also nodes that become immobile after an event (e.g. injured persons) can be modeled with this role.

2) Gravity-Based Mobility: If several independent disaster areas are simulated, a mobility model based on gravity (proposed by Nelson et al. [12]) can be used to describe how nodes move towards or away from the individual areas. In this model, every disaster area is modeled as a gravity source. The force F a disaster area has on a node can be described by the intensity I as $F = I/d^2$ with d being the distance between the node and the disaster area. I also describes whether the node moves towards a disaster area or away from it. For a particular node, the force vectors $\vec{F}$ from all disaster areas are then combined to the vector sum $\vec{F}_{total}$ describing it's resulting speed and direction of movement.

3) Zone-Based Mobility: Aschenbruck et al. [2] describe a partitioning of the entire disaster area according to handbooks for first responders in Germany [11], [15]. This partitioning then influences how nodes move. The disaster area is divided into four different zones:

- The *incident site* is the zone where the actual disaster occurs. In this zone casualties are to be expected and the effects of the disaster have to be combated (e.g., fire).

- In the *treatment zone* the injured wait for their treatment after they have been rescued from the incident site. Paramedics give first aid to injured and bring them to the transport zone. Usually, the treatment zone is close to the incident site.
- In the *transport zone* transportation vehicles such as ambulances and helicopters wait to take injured to the hospitals.
- The *hospitals* (or the *hospital zone*) are typically further away from the incident site and are not part of the actual disaster area. Injured are being transported here via transportation vehicles for further treatment.

In this model, every node belongs to one of the different zones. Some transportation nodes move between zones, others only move inside one zone, for instance, fire fighters.

The zone-based mobility model, as well as some of the random models described above, are implemented in the software BonnMotion [1], a rich Java-based software that generates synthetic movement traces to investigate mobility in different scenarios. The generated movement traces can also be exported as input data for several supported network simulators (cmp. Figure 3a). However, BonnMotion, and also other mobility simulation frameworks, lack an online interface for network simulators because they are not intended to be executed with such simulators in parallel. Also, none of the mobility models above consider communication between simulated nodes to influence the resulting movement.

III. Mobility Simulation Framework

In the last section, we presented different mobility models to be used for general purpose and for disaster recovery simulation. In this section, we introduce our approach for integrated simulation using disaster recovery missions as an example. In our approach, we focus on a deep interaction between network simulation and mobility simulation to realistically simulate complex scenarios where movement of nodes is also based on communication between them.

We first introduce the environment where the simulated nodes move and operate in. These simulated nodes are implemented as software agents based on rule sets which we present afterward. Third, we describe the key concept of our approach: the communication between agents and the interaction of agents with the environment. Finally, we present the general system architecture of our simulation framework.

A. Environment Model

The basis for mobility simulation is always formed by the environment where mobile nodes move in. Typically, this environment is modeled as a two dimensional plane where mobile nodes and obstacles are placed on. During the mobility simulation, the nodes are moved on this plane, thus, generating movement traces. In many cases, this movement is based on a random mobility model. However, in realistic scenarios, like disaster recovery missions, nodes (e.g., first responders) move according to certain properties in their surrounding environment. First responders use vehicles to move on streets to, from, and between incident sites, fire fighters extinguish fires in buildings or in the environment, and so on. Therefore, information about the environment is essential to disaster simulation and for realistic movement patterns in general.

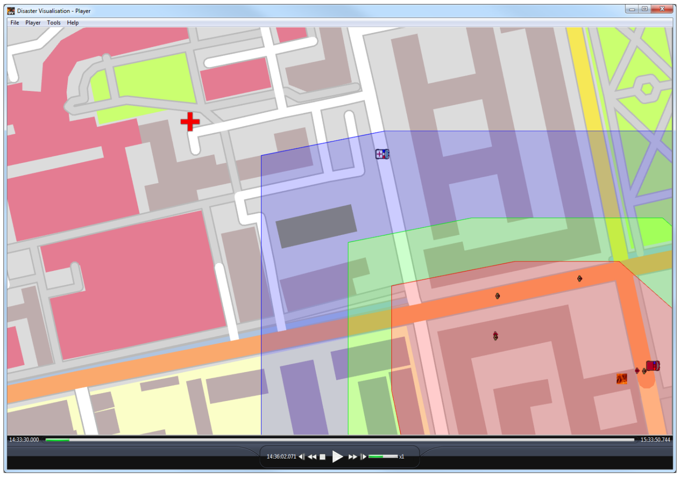


Figure 1. The graphical representation of the environment

For this reason, we use data from the OpenStreetMap project [8] to generate the environment. This data provides information about streets, buildings, fields, woods and other environmental properties. If nodes move on streets, their speed can be adapted according to the streets' speed limit, for instance. Information about hospitals can be used in the simulation as targets for ambulances to transport injured to. This data greatly helps to generate realistic movement traces.

Based on this environmental data, seats of fire can be placed manually on the map as a starting point for the disaster simulation. During the simulation, fire spreads with an average speed of $0.25m^2$ per minute (according to DIN 18232) or it gets extinguished by fire fighters. Also, we use the state of the art work by Aschenbruck et al. [2] to create the three aforementioned virtual zones *incident site* (red), *treatment zone* (green), and *transport zone* (blue) surrounding the fire places, also depicted in Figure 1. The information about hospitals is used to create the *hospital zones* (red cross in the top left corner of Figure 1).

B. Mobility Model

The implementation of our mobility model is based on the concept of software agents [3]. This means, every node is represented by a software agent that generates the movement patterns for that particular node. Agents, on the other hand, are instances of an agent type. Every agent type contains a set of rules describing how to react in specific situations and on different events. In the case of disaster recovery, agent types are *fire fighter* and *police car*, for example, and the individual fire fighters or police cars are then instances of the *fire fighter* or *police car* agent types, respectively.

Each agent has limited knowledge about the simulated world including other agents and the environment. In particular, the knowledge of each agent consists of the environmental map, that is the road network and the positions of the associated institutions such as hospitals or police and fire stations. At the beginning of each simulation, every first responder agent has initial information about the type and location of the disaster areas, injured people in these areas, the partitioning into zones, and the positions of their colleagues. Based on its current information, every agent chooses specific targets in the environment and moves towards them considering the

ISBN: 978-1-61208-313-1

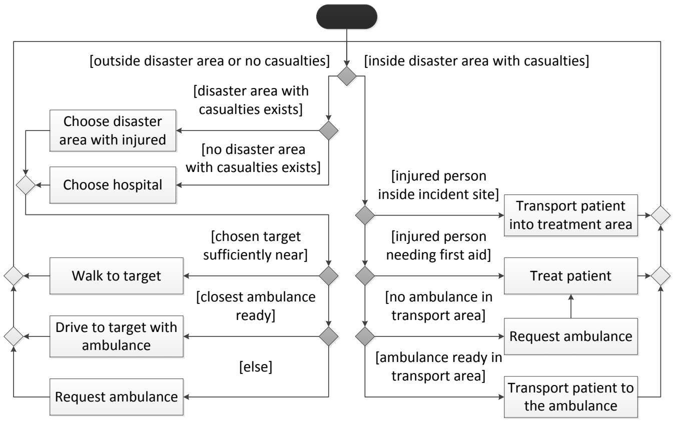

[outside disaster area or no casualties]
[inside disaster area with casualties]
[disaster area with casualties exists]
Choose disaster area with injured
[no disaster area with casualties exists]
Choose hospital
[injured person inside incident site]
Transport patient into treatment area
[chosen target sufficiently near]
[injured person needing first aid]
Walk to target
Treat patient
[closest ambulance ready]
[no ambulance in transport area]
Drive to target with ambulance
Request ambulance
[ambulance ready in transport area]
[else]
Request ambulance
Transport patient to the ambulance

Figure 2. Rule set of a paramedic

available geographical information or retains its position for a specified time span. The choice of targets determines the movement pattern characteristic to that particular agent type. Our behavior-based mobility model differentiates not only between the various types of first responders such as police men or fire fighters, but also between the ways of moving such as *by motor vehicle* or *by foot*. Therefore, agents are moving specifically to their roles in order to obtain most realistic behavior-based movement patterns.

The precise movement of the agents roughly follows the simple high-level actions *repelling*, *attracting*, *oscillating*, and *immobile* used in the generic event- and role-based mobility model described by Nelson et al. [12]. However, targets selected by the agents will depend on a substantially more fine-grained behavioral simulation which simulates the performed tasks, that are specific for the respective agent type, directly at the level of the agents. For example, fire fighters have two primary tasks. First, casualties have to be transported from the *incident site* (where other first responders cannot operate due to the dangers) to the *treatment zone* (where the injured can be treated by paramedics). The second task is to fight disasters, for example, to extinguish fire.

The task of medical personnel during rescue missions is to bring injured people out of danger to the *treatment zone*, to treat them there, and then bring them to the *transport zone*, where they are picked up and transported to a *hospital*. For that, a paramedic is moving together with the injured person from the *treatment zone* to the *transport zone*. This injured person is then transported to a *hospital* for further treatments and the paramedic heads for another injured person in order to prepare him for transportation. Figure 2 shows an example rule set for the paramedic agent type.

The police has the task to secure disaster areas. This means that the traffic hubs (cross roads) need to be secured to prevent civilians from entering the disaster area. Police officers might have to patrol between several intersections and possibly expel civilians who are already within the disaster area.

The movement of civilians is based on the physics-based gravitational mobility model described by Nelson et al. [12], that allows nodes to respond to the presence of multiple disaster events. Agents can flee from several independent disaster events or approach them.

This agent-based concept provides us with the opportunity to simulate other scenarios as well, for instance, a public transportation system. Buses or trams and trains in cities can be modeled using our mobility model by implementing the rule sets for such transportation agents.

C. Interaction and Communication

The rules above clearly show that the movement as well as the behavior of agents are directly based on interaction between agents and with the environment. For example, fire fighters extinguish fires in the environment (e.g., in buildings) and paramedics carry injured people to ambulances to be transported to the hospitals.

Furthermore, first responders can send requests to other rescuers, if they need help or a transportation vehicle. For example, paramedics request ambulances to allow patients to be transported to a hospital. An ambulance, that is available for transportation, can then acknowledge the request and move to the requested location. The transportation of first responders to the disaster area and back to the headquarters is carried out in the same manner, as well. Also, first responders inform their colleagues about their new position as soon as they proceed to a new location. This helps to keep the information on the colleagues' positions up to date for all first responders. First responder agents use this information to decide who should be assigned to which disaster event.

Finally, first responders inform their colleagues and the commander about the progress of the rescue mission. Especially, if the fire is extinguished, the fire fighter agents inform their colleagues about the finished task. Also, paramedics inform their colleagues about rescued persons. Based on this information, first responders know what tasks are remaining and who is available for a new task.

This shows how the movement of agents also depends on the communication between them. Only, if the information exchange between first responder agents is reliable, the rescue mission is accomplished successfully and efficiently. Conversely, the communication of the agents also depends on their movement and the communication technology used. If a real network based on wireless technology is considered for message transportation, the agents need to be in transmission range to each other for messages to arrive at the destination. Hence, the network simulation cannot be launched after the mobility simulation, but both simulations have to run together in parallel to account for the influence of both simulations.

D. System Architecture

Our simulation framework focuses on these influences between mobility and network simulations. It is based on a discrete event-driven simulation engine. This means, events occur at definite points in time during the simulation and the simulation can be described as a chronological sequence of events. In a typical state of the art simulation setup, the mobility simulation is executed first, creating movement traces to be imported into the network simulation afterward (cmp. Figure 3a). Our system, on the other hand, allows for an integrated simulation where network and mobility simulations are executed in parallel, thus, influencing each other (cmp. Figure 3b). However, this requires the mobility model to not only generate basic movement but communication patterns, as well (as described above). Assuming such a model, this provides the advantage of simulating communication between nodes more realistically, since their communication can directly influence their movement (e.g., nodes calling for help).

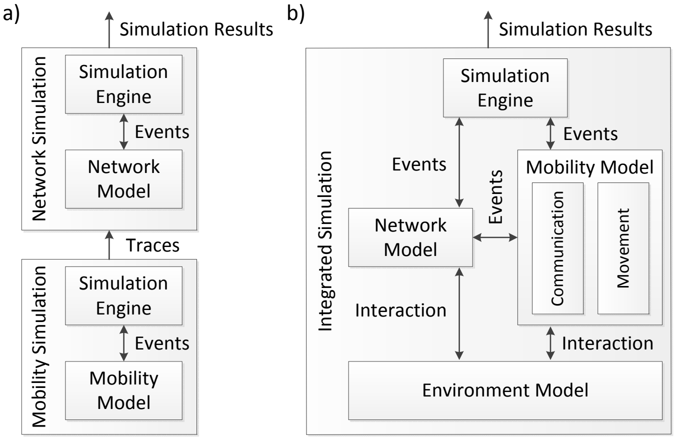


Figure 3. a) Typical simulation setup – b) Integrated simulation framework

To connect our proposed mobility model to a network simulator, a shared simulation engine is needed to dispatch mobility and network related events. This does not necessarily require events to be compatible with both, the network and the mobility simulators. However, for simplicity the events are compatible with both simulators in our current implementation. We connected our framework to the state of the art discrete event-driven network simulator PeerfactSim.KOM [16]. This simulator can now be used to simulate different communication technologies in combination with our mobility model.

In the future, using the disaster recovery example, this integrated simulation enables us to investigate the influence of different modern communication technologies on the works of first responders during their missions. A simulation without a network simulator is possible, as well. In this case, we assume that messages are being transmitted immediately and without any packet loss. This can be used to very basically mimic traditional radio communication, however, not realistically. Furthermore, due to the generic design, other event-driven simulators can be connected to our simulation framework as well, for instance, a road traffic simulator.

We used our first response communication sandbox [5] as a starting point for the mobility simulation framework. However, many improvements and changes were necessary to integrate our new mobility model with the network simulator and the environment. We implemented a graphical user interface called *DisVis*, short for Disaster Visualization, to create and modify scenarios for simulation and to visualize simulation results afterward (depicted in Figure 1). In the future, we intend to investigate different communication technologies for disaster recovery missions using our integrated simulation framework.

IV. Evaluation

To evaluate our mobility model described in the last Section we have drawn on expert knowledge. We created a questionnaire targeted at experts from *police departments*, *fire fighting departments*, and *medical facilities*. With this survey, we intended to extract different types of information. This resulted into three similar questionnaires, one for each of the first responder groups listed above. Each questionnaire was separated in four parts to get information about *a)* activities of first responders during missions, *b)* communication between them, *c)* details about the works of each group, and *d)* the four disaster zones described above.

Our call to fill out the survey was followed by 84 individuals most of which already participated in disaster recovery missions. This number divides into 44 individuals from medical facilities, 32 individuals from fire fighting departments, and 8 individuals from police departments. As the number of participating members from police departments shows, their role during disaster recovery missions is not as big as that of the other groups. Usually fire fighters are the first to arrive and take over certain roles associated with the police. The main role of the police is investigation and criminal prosecution which usually is a long term activity and starts after the disaster recovery mission is finished. This was also reflected in the answers by both, the participants associated with the police departments as well as the participants associated with the fire fighting departments.

A. Activities

First, we presented the atomic activities we implemented in our mobility model to abstract the behavior of the three first responder agent types. The goal was to determine the priorities and rules for these atomic activities as well as to identify important but missing activities.

Our assumption about the priorities was accurate in general. However, participants emphasized the first activity to be *examining the situation on-site*. This activity is not explicitly modeled in our rule set. But, as stated before, the agents know about the situation in the disaster areas at the beginning of the simulation and incorporate this knowledge into their decisions and activities. Therefore, this activity is modeled implicitly in our rule set.

B. Communication

Second, we checked our assumptions made about the importance of communication between first responders. This also included the organizational structures. We found that our assumptions about the highly hierarchical organizational structures of first responders is accurate in general and that communication is very important in disaster recovery missions. In fact, most activities are only executed by command. Commands are either given in a briefing at the beginning of the rescue mission or during the mission via (radio) communication.

Furthermore, the survey exhibited the explicit importance of small first responder groups. For example, a team of fire fighters and a fire engine form a group that moves and operates together. The same applies to police officers and paramedics. This effects the movement of every unit that is part of a group. It also impacts the communication, which is simulated hierarchically so that the members of these groups primarily communicate among themselves and only the group leaders communicate between different groups. We implemented this behavior into the rule sets of our mobility model.

C. Details

Third, we intended to get details about typical walking and driving distances. We also tried to get a rough estimate about the average time paramedics need to stabilize patients and about the average time fire fighters need to extinguish a fire in apartments. In general, these details can only be estimated very roughly and depend on various factors. However, we found that typically distances of up to 400 meters are walked by foot. To stabilize patients paramedics assess 2 to 10 minutes. Fire fighters asses roughly 30 to 90 minutes to extinguish a fire

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-313-1

in apartments, although, the participants explicitly pointed out that the time is highly situation dependent and cannot be easily generalized. This also is hard to be mapped to our simulation. Therefore, fire fighters in our model extinguish fire at a rate of $1.5m^2$ burning area per minute.

D. Disaster Zones

Finally, we evaluated the assumptions made about the partitioning into different zones and their sizes. The existence of different zones was confirmed by the participants, as expected. Usually, first responders define locations and sizes of these zones before the recovery mission is started. This is also captured in our simulation model. The size of the *incident site* highly depends on the affected region (e.g., 20 to 50 meters or 100 to 150 meters perimeter). In our simulation model we use 20 meters around the affected region, for instance, a building with a burning apartment.

Usually, the *treatment zone* is rather small and can be combined with the *transport zone* into one zone under some circumstances. Depending on the situation, the *treatment zone* is at most half in size compared to the *incident site*. The *transport zone*, on the other hand, is rather large and can be up to twice as large as the *incident site*. Usually, the *transport zone* is dependent on qualified locations for transport vehicles to access the zone. These locations should be as close as possible to the *treatment zone*. In our simulation model we use at least 20 meters for the *treatment zone* and at least 50 meters for the *transport zone*.

Additionally, the initial partitioning of the disaster area into zones can change over time when disasters are spreading or are averted. This fact is reflected in our simulation model. Details about the sizes of zones were incorporated after the survey.

V. Conclusion

In this paper, we presented an overview of basic and commonly used methods for mobility simulation in network simulations as well as in disaster scenario simulations. We argued that, especially in disaster simulations, the interaction and communication between the simulated agents are crucial to the resulting simulation. Both aspects require that mobility and network simulators are executed together in parallel so both simulations can influence each other. We presented our approach to simulate movement and communication patterns in disaster recovery missions based on state of the art work and expert knowledge. Our environment and mobility model is based on OpenStreetMap data and software agents, respectively. The event-driven mobility simulation framework is capable of stand alone simulation. But it can also be integrated with network simulators like PeerfactSim.KOM [16] enriching the simulation with realistic network models.

We evaluated our mobility model with questionnaires targeted at experts from police departments, fire fighting departments, and medical facilities. Especially medical personnel and fire fighters participated in the surveys. The responses gathered from the surveys allowed us to further improve our rule sets of the different agent types.

In the future, we plan to evaluate different communication technologies and mechanisms like routing algorithms in the field of disaster relief. We are especially interested in how such technologies influence the works of first responders. Furthermore, we want to enrich our rule set with more detailed rules from handbooks for first responders [11], [15] to further increase the realism of our simulation. Finally, we also plan to implement different agent types for other scenarios like public transportation to analyze communication possibilities between buses and trains in smart cities.

Acknowledgments

This work has been funded by the DFG as part of the CRC 1053 MAKI and the research unit 733 QuaP2P.

References

[1] N. Aschenbruck, R. Ernst, E. Gerhards-Padilla, and M. Schwamborn, “Bonnmotion: A Mobility Scenario Generation and Analysis Tool,” in Proceedings of the 3rd International ICST Conference on Simulation Tools and Techniques , 2010, pp. 1–10.

[2] N. Aschenbruck, E. Gerhards-Padilla, M. Gerharz, M. Frank, and P. Martini, “Modelling Mobility in Disaster Area Scenarios,” in Proceedings of the 10th ACM Symposium on Modeling, Analysis, and Simulation of Wireless and Mobile Systems, 2007, pp. 4–12.

[3] M. Balmer, N. Cetin, K. Nagel, and B. Raney, “Towards Truly Agent-based Traffic and Mobility Simulations,” in Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems , 2004, pp. 60–67.

[4] D. Bradler, B. Schiller, E. Aitenbichler, and N. Liebau, “Towards a Distributed Crisis Response Communication System,” in Proceedings of the 6th International Conference on Information Systems for Crisis Response and Management, 2009, pp. 1–9.

[5] D. Bradler, I. Schweizer, K. Panitzek, and M. Mühlhäuser, “First Response Communication Sandbox,” in Proceedings of the 11th Communications and Networking Simulation Symposium, 2008, pp. 115–122.

[6] T. Camp, J. Boleng, and V. Davies, “A Survey of Mobility Models for Ad hoc Network Research,” Wireless Communications and Mobile Computing, vol. 2, no. 5, 2002, pp. 483–502.

[7] R. A. Guérin, “Channel Occupancy Time Distribution in a Cellular Radio System,” in IEEE Transactions on Vehicular Technology, vol. 36, Issue 3, 1987, pp. 85–97.

[8] M. Haklay and P. Weber, “Openstreetmap: User-generated street maps,” IEEE Pervasive Computing, vol. 7, no. 4, 2008, pp. 12–18.

[9] X. Hong, M. Gerla, G. Pei, and C.-C. Chiang, “A Group Mobility Model for Ad hoc Wireless Networks,” in Proceedings of the ACM International Workshop on Modeling and Simulation of Wireless and Mobile Systems, 1999, pp. 53–60.

[10] D. B. Johnson and D. A. Maltz, “Dynamic Source Routing in Ad hoc Wireless Networks,” in Mobile Computing, T. Imielinski and H. Korth, Eds. Kluwer Publishing Company, 1996, vol. 353, ch. 5, pp. 153–181.

[11] T. Mitschke, “Handbook for First Responder Groups (original: Handbuch für Schnell-Einsatz-Gruppen).” Stumpf & Kossendey, 1994.

[12] S. C. Nelson, A. F. H. III, and R. Kravets, “Event-driven, Role-based Mobility in Disaster Recovery Networks,” in Proceedings of the second ACM Workshop on Challenged Networks, 2007, pp. 27–34.

[13] E. M. Royer, P. M. Melliar-Smith, and L. E. Moser, “An Analysis of the Optimum Node Density for Ad hoc Mobile Networks,” in Proceedings of the IEEE International Conference on Communications, 2001, pp. 857–861.

[14] M. Sánchez and P. Manzoni, “ANEJOS: A Java-based Simulator for Ad hoc Networks,” Future Generation Computer Systems, vol. 17, no. 5, 2001, pp. 573–583.

[15] J. Schreiber, “Work Instructions for First Responder Groups of Medical and Rescue Services (original: Arbeitsanweisungen für SEGen im Sanitäts- und Rettungsdienst).” Stumpf & Kossendey, 2000.

[16] D. Stingl, C. Gross, J. Rückert, L. Nobach, A. Kovacevic, and R. Steinmetz, “PeerfactSim.KOM: A Simulation Framework for Peer-to-peer Systems,” in International Conference on High Performance Computing and Simulation , 2011, pp. 577–584.

[17] J. Yoon, M. Liu, and B. Noble, “Random Waypoint Considered Harmful,” in Proceedings of The 22nd Annual Joint Conference of the IEEE Computer and Communications Societies, 2003, pp. 1312–1321.

KeyGait: Framework for Continuously Biometric Authentication during Usage of a Smartphone

Matthias Trojahn
Volkswagen AG
Wolfsburg, Germany
matthias.trojahn@volkswagen.de

Frank Ortmeier
Otto-von-Guericke University
Magdeburg, Germany
frank.ortmeier@ovgu.de

***Abstract*—The ability of having a secure mobile device is determined by different aspects (e.g., hardened system, authentication or anti-virus). Normal authentication methods are only requesting authentication characteristics at the beginning of the usage. The aim of this paper is to create a framework which can continuously analyze which user is using the device at each moment. While mobile devices are easy to lose or can be stolen, it is important to do an authentication process during usage. We propose a continuous trust model using keystroke dynamics and movements of the device as biometrical modalities to have a certainty of the usage at each time.**

***Keywords*—security framework; mobile devices; biometric authentication; continuously authentication; usability**

I. Introduction

The need of a continuous authentication process exists for mobile devices. These devices can be stolen or lost easily because they are so small. For example, a survey from Credant Technologies reported in 2008 that in six months 55,000 cellular phones were left in London taxis [1]. The first challenge is that these devices are only secured by a password. In addition, only if the device is unlocked properly the password is asked when accessing the device. This means a continuous authentication system with a properly initial authentication combined with a re-authentication during usage is needed. Re-authentication means an additional authentication during usage, which happens in the background.

In this paper, we present a framework which uses inertial sensors and a capacitive display to fulfill the need for the continuous authentication system.

For this, we describe the related work and the contribution of our work in this section. In Section 2, we present our continuous authentication model with the concept and the trust model which handles the certainty of the device which user is using the device. Then, we will present hypothetical test cases for using this model. Finally, we discuss the framework in Section 4 and conclude our research in Section 5.

A. Related work

Two related researchs were identified. First, we will focus on the biometrical authentication via keystroke on smartphones. This can be separated into text dependent or independent analysis. The second approach is concerning the gait recognition of a person.

Prior work for keystroke dynamics was mainly focused on keyboard for a PC [2], [3] or on the mobile phone with 12 keys [4], [5]. Most of the experiments are using the features “duration of pressing one key” or the “time between pressing two / three keys”. In general, error rates like *false acceptance rate* (FAR) or *false rejection rate* (FRR) are used to compare different results. The FAR describes how intruders can access a system. In addition, the percentage of rejections of an authorized person divided by all attempts of authorized person is called the FRR. Both error rates have to be as small as possible to have a secure system. The point where FAR is equal to FRR is called *equal error rate* (ERR). Good results for text dependent authentication are already shown by Karatzouni et al. [6] with a EER of 12.2 % (experiment with 50 person). The advantage of keystroke is that not all attempts by an intruder are successful compared with a simple password authentication. A FRR of 12 % means that only every ninth attempt of all unauthorized attempts is successful.

Zahid et al. [7] did a text independent keystroke authentication. They used key hold time, error rate and different digraphs (horizontal, vertical, non-adjacent horizontal and non-adjacent vertical) as features. The different digraphs are used because there is no prediction of the combination between different keystrokes. These experiments on a mobile phone with 12 keys had a result of FAR 11 % and FRR 9.22 %.

As in the survey done by Banerjee [8], a lot of different experiments showed better results then the previous mentioned experiments but there the number of subjects was low (under 50 people).

The gyroscope is employed to measure any rotation of the device. Only the uniqueness for single persons turned out to be rather low. In an experiment Derawi et al. [9] had in the study an EER of 20 % (the device was carried on the hip of the test person). This is not enough to be a good single authentication method. A fusion with the keystroke dynamics is necessary. Further work has shown that with a higher sampling rate the EER can be improved [10], [11]. If more than one smartphone could be used the recognition rate could be reduced [12].

B. Contribution of this work

The major shortcoming of all existing approaches is that they do not allow continuous authentication on smartphones. Keystroke dynamics with a fixed text is only possible during the unlock process. After this, the user is not typing the same pass phrase again. Text independent keystroke authentication is not analyzed properly for the new generation of mobile phones with capacitive display. Gait authentication is an approach which can be used as a continuous authentication but the error rates are too high to give a certainty which is needed for a

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-313-1

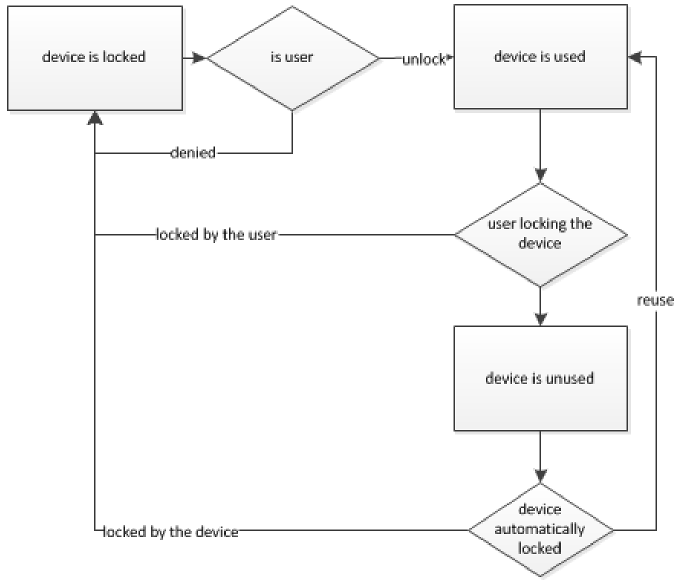


Fig. 1: Activities during usage of a smartphone

secure environment. It only can be used during fusion with another modality.

The goal of this work is to propose a generated framework which can be used to authenticate a person continuously during the whole process of usage. For this, we present a solution using different sensors of smartphones. The solution is based on a trust model where the users are authenticated with a particular certainty.

Fig. 1 describes the whole process of unlocking and locking of a device. If the user wants to use the device, an authentication has to be done. For this task, the keystroke authentication is suitable.

Fig. 1 also shows also which possibilities exist after usage. First, the user locks the device himself. Second, the device is locking itself after a predefined time. The last possibility is that the user wants to use the device again. In this situation it is without a continuous authentication not possible to say whether it is the same user or not. This model which includes a continuous authentication will be described in the next section.

II. Our continuously authentication methodology

The fundament of our model is a fusion of different modalities and a transparent trust measurement. This fusion has to be done continuously while the device is unlocked. First, we will describe which modalities are included, then, we will present the trust model.

A. *Concept for a model*

There are three basic points where a model can be attached: during the unlocking of the device, during the usage of the device and during the time the device is not used but unlocked.

Only if all points are included in an authentication system it can continuously give information about how certain the temporally user is recognized.

Because we focused on the capacity display and the gyroscope of the device, we could do the initial authentication on the device via keystroke or gait. As we already stated, the error rates for gait recognition are too high to give enough certainty about the user. That is why we propose to use the keystroke in addition to the password which we analyzed previously on smartphones [13]. We suggested in different experiments the usage of the capacitive display to extract additional features (e.g., size of finger during typing or the correct coordinates).

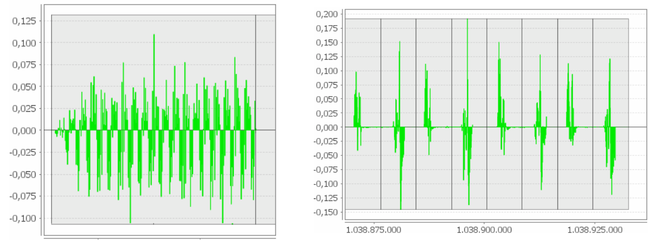


Fig. 2: Data of the gyroscope sensor (Left: Z-axis during walking; Right: X-axis during putting device on table)

In both of the next use cases, it is demonstrated how the gyroscope data can be used. Fig. 2 shows the changes of the one of three vectors (axis x, y and z - z vector was used). On the left, a person was asked to walk with the device in the pocket for 20 seconds. All the recorded steps show consciously similar changes between different steps. The other two axis show the same similarities which means a function can be created which make it possible to recognize these pattern of values as a walking person. The right figure represents how a device is rotated. At first, the user had to put the device form a table to his pocket and then do it the other way around. In this use case it can be seen how sensitive the data are. Small changes are existing even if the device is placed on a table that means a filter has to be used to extract these incorrect data. After cleaning the data streams different scenarios can be extracted by generating models for use cases (unique combination of the gyroscope values).

In general, it is possible to detect whether the device moves or stays at a location. This is important to trace whether the device was unused by the user. In addition, as already stated, the gait recognition could be used for a re-authentication if the user is walking during using the device.

On the other hand, if the user is typing, the capacitive display can be used to record data and authenticates the user by the behaviour during typing (see Subsection I-A).

Gait and the text independent authentication using keystroke can be used for a re-authentication. This could be used to decide whether the user can reuse the device (see Fig. 1). If no decision could be made or the device is already locked, the user has to authenticate himself via his password.

B. *Trust model*

In the previous subsection, we described which modalities we use for our framework. Now, we will define a trust model for continuous authentication.

In Fig. 1, all use cases were shown which have to be represented by the model. The device is locked and the user has to authenticate him by using password and his biometric keystroke behavior. Basically, the trust level depends on the initial authentication. A higher certainty ($auth_{intial}$) results in a higher trust level at the beginning of the usage process.

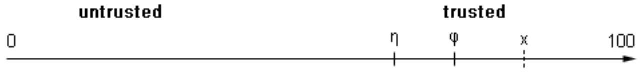


Fig. 3: Scale for the trust model

Furthermore, the time has an important role. With a rising time difference between the initial authentication and the current time the certainty decreases. Only with further re-authentication methods the trust level could rise again. This concept can be represented with the next formulas.

$$trust(t) = auth_{intial} - \alpha \sum_{i=0}^{N} (cert(i)) \tag{1}$$

$$cert(t) = \begin{cases} 0, & if(key(t) \neq null) \ and \\ & (\beta \ move + \chi \ key \geq \delta) \\ \frac{\delta}{2} - \beta \ move(t), & if(key(t) = null) \\ \delta - (\beta \ move(t) & \\ +\chi \ key(t)), & otherwise \end{cases} \tag{2}$$

This means the initial authentication and the decreasing certainty $cert$ during a time box are influencing the trust in which the device know which person is using the device. The time length of the time boxes has to be evaluated in combination with β and χ. During this time box the decreasing certainty is calculated by recognition of the movements of the devices $move(t)$ and the interaction with the capacitive display $key(t)$. Both values are representing the certainty of each sensor whether it is still the same person. They can be between 0 and 100. The value δ describes the threshold which is needed that the trust level does not change. The value $move(t)$ can be walking of the person, text input or a combination. For example, if a user is walking all the time after the unlock it, the trust level should not decrease a lot. The second case represents the case if the user is not using the device (e.g., is in the pocket or is laying somewhere).

Fig. 3 shows the scale for the trust value. The position x represents the initial authentication. The value is influenced by the threshold of the authentication system and the amount in which the authentication value was higher than the threshold. With a bigger difference the $auth_{initial}$ is rising. In Figure 3, additional areas are shown. The trusted area is the range where the device knows who the user is with a specific trust level. If the trust level is under the threshold φ, the device is unsure and the device gets locked (the user cannot be temporally recognized enough). Only with an initial authentication the user could access the device again. Before this could happen, the temporally trust level gets under the value η. Then the certainty is not enough to access all systems. In some companies policies exist where with a one-factor-authentication (password) not every system could be accessed whereas with a two-factor-authentication (public key infrastructure with password) the access is granted. This can be adapted to this model. If the trust is over the trust level η the systems grants access to applications which are normally accessed with a two-factor-authentication. Between point φ and η it could be used as a one-factor-authentication.

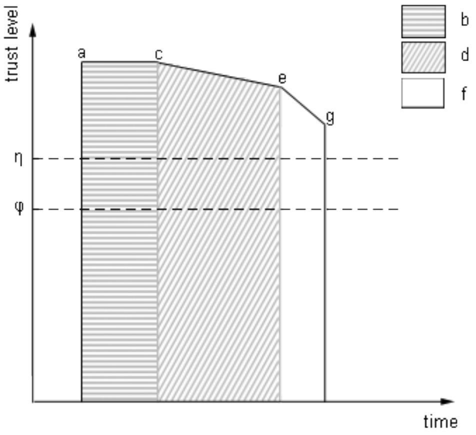


Fig. 4: Trust level for scenario 1

III. Hypothetical test cases

In this section, we will show how the framework works. For this, we present the steps for two different scenarios. The framework was implemented. For this, the variables (χ, β ...) which are described in the formulas of Subsection II-B have been replaced by the number 1 (naive approach). The time for a time box is set to five seconds.

A. *Scenario 1: Trust level is always given*

In the first scenario, a user is writing emails while walking (b) after the initial authentication (a). Then, the user stops writing and puts the device in the pocket of his trousers (c). After this, the user is continuing walking (d) and when an email is incoming, he is taking the device out of the pocket (e) and reads the email (f). The last step is the locking by the user (g). All the time the user was over both trust level lines so the trust was high enough for all applications.

Different context changes can be seen in the Fig. 4. The initial authentication has a very high trust value. This has two reasons, first the error rates for keystroke authentication are low (at least under 5%) and second the user was identified with a high certainty. During writing and walking the sensor collects a lot of data. With this it is possible to recognize a person that is why the trust level does not decrease much. Putting the device in and out of the pocket the system recognizes a context change. This can be used in next time boxes. We know with this context changes that the user is still in the position of the device to a very high level. It is, especially, in the situation that the capacitive display is not used because the error rates of gait are higher than the error rates for keystroke.

On the other side, if walking and reading are compared the trust level decreases more because while reading on the device, not enough input is generated to identify a person.

B. *Scenario 2: Where an automatic lock happens*

The second scenario was generated to show in which situation an automatic lock of a device happens. For this, we

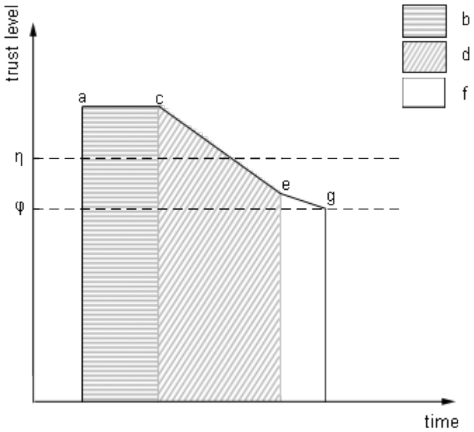


Fig. 5: Trust level for scenario 2

proposed a solution where the device is laying on a table for a non-defined time.

Like the first scenario the start is again an initial authentication (a). After this, the device is unlocked and can be used. Here, the user writes an SMS (b) and after this puts the device on a table (c) and it lays there for a time period (d). Then the user reads a SMS (f) that he just received (e). Then the device is locked (g) because the trust level reached the minimum. In Fig. 5, these steps are shown.

The first two steps are the same like in the previous scenario, one only that a SMS is written. Then, if the user puts the device on a table, both sensors do not get any more data in this case the trust level decreases a lot because the user can be everywhere. During this time the trust decreases under the first trust level line. After this, if the user wants to access a high secure application, a new initial authentication has to be done. While reading the SMS, no input has been done in this case the trust level decreased more until the second line is reached and the device gets locked no matter if the user wants to read the SMS or do other think. In this case the user has to input his initial credentials again to use the device again.

IV. Discussion

This section will discuss this approach and will present some advantages and disadvantages.

With this framework, not only can the trust be established at the beginning of the usage, but even during usage it is possible to recognize the user. For this, no additional interaction has to be done. And no common attack is possible (e.g., shoulder surfing or social engineering) because the additional feature cannot be recorded. That means the security is raised. For comparison, with a password the FAR is 100 %. Together with the knowledge factor this biometric framework could be seen as a two-factor-authentication.

On the other hand, some limitations exist at the moment. The energy consumption for all the required sensors is too high for the general usage during one day. In addition, most of the studies for biometric modalities are not tested in general use cases. All possible movements are not tested how well they can be recognized and the system needs a lot of training to identify the user.

V. Conclusions

In this paper, we first identified the problem of the continuity of an authentication method on smartphones. Therefore, we proposed a continuous authentication method using keystroke dynamics (text dependent and independent) in addition to the movement of the device (e.g., gait recognition). These methods have to be fused and checked during several predefined timed boxes.

We presented a decision model how the trust can be calculated during one and more time boxes. Especially, the introduction of different thresholds is important for using applications with different security level.

We proposed some scenarios, which show how the framework is working, e.g., in which scenario the device locks the device. In addition, advantages and disadvantages are presented.

Overall, the proposed framework is an option for a biometrical authentication on smartphones. It is an important step towards a more effective and continuous authentication.

In the future, more tests have to be done with the modalities, especially, in more general use cases. In addition, the energy consumption has to be reduced.

References

[1] J. Twentyman, “Lost smartphones pose significant corporate risk,” 2009. [Online]. Available: http://www.scmagazineuk.com/lost-smartphones-pose-significant-corporate-risk/article/126759/ [retrieved: 09/2013]

[2] D. Umphress and G. Williams, “Identity verification through keyboard characteristics,” in *Intl. J. of Man-Machine Studies*, 1985, pp. 263–273.

[3] F. Monrose and A. D. Rubin, “Keystroke dynamics as a biometric for authentication,” in *Future Generation Computer Systems*, vol. 16. Elsevier Science Publishers B. V, 2000, pp. 351–359.

[4] S.-s. Hwang, S. Cho, and S. Park, “Keystroke dynamics-based authentication for mobile devices,” in *Computers & Security*, vol. 28, no. 1–2, 2009, pp. 85–93.

[5] A. Buchoux and N. L. Clarke, “Deployment of keystroke analysis on a smartphone,” in *6th Australian Inf. Sec. & Management Conf.*, 2008.

[6] S. Karatzouni and N. L. Clarke, “Keystroke analysis for thumb-based keyboards on mobile devices,” in *Proceedings of the IFIP TC-11 22nd Intl. Inf. Sec. Conf.*, H. S. Venter, M. M. Eloff, L. Labuschagne, J. H. P. Eloff, and R. v. Solms, Eds., 2007, pp. 253–263.

[7] S. Zahid, M. Shahzad, S. A. Khayam, and M. Farooq, “Keystroke-based user identification on smart phones,” in *Intl. Symposium on Recent Advances in Intrusion Detection*, ser. RAID ’09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 224–243.

[8] S. Banerjee and D. Woodard, “Biometric authentication and identification using keystroke dynamics: A survey,” in *Journal of Pattern Recognition Research*, vol. 7, 2012, pp. 116–139.

[9] M. O. Derawi, C. Nickel, P. Bours, and C. Busch, “Unobtrusive user-authentication on mobile phones using biometric gait recognition,” in *6th Intl. Conf. on Intelligent Inf. Hiding and Multimedia Signal Processing*, Washington, DC, USA, 2010, pp. 306–311.

[10] K. Holien, “Gait recognition under non-standard circumstances,” Ph.D. dissertation, Gjøvik University College - Department of Computer Science and Media Technology, 2008.

[11] J. Mäntyjärvi, M. Lindholm, E. Vildjiounaite, S.-m. Mäkelä, and H. Ailisto, “Identifying users of portable devices from gait pattern with accelerometers,” in *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 2005.

[12] G. Pan, Y. Zhang, and Z. Wu, “Accelerometer-based gait recognition via voting by signature points,” in *Electronics Letters*, vol. 45, no. 22, 2009, pp. 1116–1118.

[13] M. Trojahn and F. Ortmeier, “Biometric authentication through a virtual keyboard for smartphones,” in *International Journal of Computer Science & Information Technology (IJCSIT)*, 2012.

Quality of Experience and Human-computer Interaction: A Relation Overview

Luis Guillermo Martinez Ballesteros, Zary Segall
Mobile Services Lab
KTH Royal Institute of Technology
Kista, Sweden
lgmb@kth.se,segall@kth.se

Abstract—**This paper first traces the historic evolution of the Quality of Experience (QoE) concept, and then connects common points between the study of user experience as practiced within Human Computer Interaction (HCI) disciplines and recent efforts to understand how user perception can be incorporated into the definition and management of resources in the area of (ICT). After an analysis of the history of QoE and an examination of its current role in HCI, some research challenges are proposed that open doors to future research projects.**

Keywords-Quality of Experience; Human Computer Interaction; Challenges in mobile environments

I. INTRODUCTION

The telecommunications industry has been a fertile area for applying user-centred solutions [1] as well as a vital part of the economies of all nations, shaping the quality of life of people around the world. It is this area where new research issues are emerging and changing the way that people interact with networks and content. Some of these efforts are oriented to technical developments, while others are focused on non-technical aspects, and it is in this new environment where concepts such as interaction, quality, content, context, and perception become more and more important to the market through operators, content providers, and handset manufacturers for whom the concept of user satisfaction is becoming a new competitive factor. A representative case is Apple; with the creation and consolidation of devices like the iPad and iPhone, this company opened the door to both a new market conception oriented to satisfy user needs through design and detail, and to a greater use of data networks by the same users through increased interest in applications and content. As a consequence of this phenomenon, recent years have witnessed an increase in the network traffic caused by a high demand for content, with users more and more interested in the quality of the content, not only from a network performance perspective, but also in how this content is distributed and consumed, including devices and interfaces. The new paradigmatic eco system (user-interface-network-content) requires novel and disruptive end-to-end considerations, in order to enable and sustain the next generation of services and user experience. In particular, networks are currently agnostic and have no knowledge about the type and characteristics of the specific mobile services they are providing. Further, there is a knowledge separation between service designers/builders, service providers, terminals, Operating Systems and networks. These facts are producing substantial resource optimization deficiencies and discontinuities that affect the user's level of satisfaction. Although Quality of Experience (QoE) has made rapid gains as a new metric influencing the success or failure of new applications and services by involving the user's perception in the evaluation process, most of the methodologies developed to measure it ([2]-[4]) depend largely on the end-to-end Quality of Service (QoS) metrics, which can be categorized as a techno-centric approach.

In response to this, a user-centric point of view is getting more attention as a new and interesting research topic where areas such as human-computer interaction (HCI) have shown interest in developing standardized assessment methodologies, optimization processes and metrics definitions taking into account concepts like User Experience (UX)([5]-[8]). While from a technical approach, user's satisfaction is a result of the adjustment of some network parameters, with a user centric point of view; QoE has a multidimensional character and can be studied from an interdisciplinary perspective [7]. However, this multidimensional character and pluralism of perspectives have naturally contributed to the existence of several definitions and approaches to the same concept. This has not allowed the emergence of a single definition that encompasses within itself the multiplicity of concepts around QoE, as well as the options for standardization in the methods of evaluation, measurement and improvement of the QoE perceived by users. As mentioned by Moor et al. [8] "It is rather uncommon to integrate concepts from other fields less technical than telecommunications in definitions of QoE. A relevant example is the domain of HCI, in which concepts such as UX and "Usability" closely related to QoE are very important." The goal of this paper is to identify coincident points between techno-centric and user-centric approaches, taking into account a review of their respective historical evolution processes that allows one to establish a basis for the development of scheme that allocates resources in a wireless infrastructure based on the evaluated QoE obtained through the implementation of an assessment methodology.

The rest of the paper is organized as follows: In Section II, we present the QoE and the techno-centric approach

ISBN: 978-1-61208-313-1

description, In Section III, the concept of QoE is analysed from a user-centric approach remarking the potential contribution of HCI to the QoE evaluation. We conclude the paper in Section V proposing some future challenges in the area and future work.

II. QoE and the Techno-centric Approach

From a technical point of view, different QoE definitions have been proposed ([9] - [11]). A review of some of these definitions allows one to see a gradual evolution since an initial idea of Quality of Service (QoS), with a ”rich tradition in engineering and developing environments” [12] basically oriented to the evaluation and adjustment of some network parameters, to a ’semantic variant’ and user-centric approach denominated QoE, emerging in the late 90’s, where user’s interests and experiences became more important [12].

One of the first mentions of the QoS concept can be found in [13]. Here, the discussion is focused on describing and determining the relationship between telephone circuit loads and corresponding delays to traffic, and how these delays directly relate to the quality of service. Even though QoS is mentioned, there is no definition of the concept. After three decades, a new mention of the term QoS appears in [14], where the financial influence of the quality of service provided by telephone operators is remarked upon. Even if the work of Pocock also attempts to open the discussion about the importance of the overall quality that user can experience, he just focused the results of the research in the necessary adjustments of the quality transmission. Pocock also tries to show the relation between the user’s appreciation and the speed, availability and reliability of a service, making explicit, for the first time, the relationship between the user’s opinion about a service and the technical factors behind its provision. However, there is no mention of external factors (i.e., economic, social, etc.) that could affect a user’s perception. In the same line of Pocock, different research efforts ([15]-[18]) proposed mechanisms to increase the reliability, and consequently the QoS, through technical modifications in both wired and wireless networks, without mentioning mechanisms for QoS evaluation.

In 1986, Gruber [19] discusses the creation of a QoS framework according to the competitive environment that appeared on the horizon of the telecommunications sector at that time. To Gruber, the possibility of unifying network infrastructures to provide a multiplicity of services required the implementation of monitoring and surveillance systems to manage and automatically control network resources and resulting QoS. While the article sees the prospect of a competition based on the provision of high QoE, neither the assessment methods nor the role that the external factors can play were considered. Only with the research results shown in [20] is presented the option to involve the user directly in the QoS evaluation process. This might be called the first attempt to incorporate subjectivity in the QoS assessment.

With the advent of packet-switching based networks and the opportunities given by this technology to provide multiple services such as telephony and television within the same infrastructure, the QoS concept gained more importance due to the need for emulating the performance of the classic and reliable telephone and television networks using, in the early years, technologies such as the Asynchronous Transfer Mode (ATM). In that sense, one of the first research efforts oriented to work with the concept of QoS focused on broadband networks was developed in the NETMAN Project [21]. According to the Brander et.al, QoS can be expressed as ”the collective effect of service performances, which determine the degree of satisfaction of a user of the service.” Here the term ”satisfaction of a user” appears in the context of this, until now, technical world. As in other cases where the role of the user is considered, until that moment there was no specific methodology to obtain information about user opinion regarding the QoS level of a network. Another important milestone in terms of QoS is expressed in [22]. Authors expressed the importance of a good end-to-end performance within the networks; first considering a clear identification of the QoS parameters to be guaranteed in real-time communications, and at the same time presented a proposal of a performance reference model for real-time packet network analysis and a real-time estimation.

Throughout the 90’s and early twenty-first century, various studies and proposals for the evaluation, improvement and implementation of QoS-based methods were made. The role of the user in the evaluation of QoS was incorporated with the development of assessment schemes such as Mean Opinion Score (MOS), Perceptual Evaluation of Speech Quality (PESQ) and Video Quality Measurement (VQM), which attempted to quantify the subjective opinion of people, giving greater weight to the evaluation from the user. The gradual process of separation between QoS and QoE was revealed in the early twentieth century by authors such as Anna Bouch, Allan Kuchinsky and Nina Bhatti in [23]. According to them, at that moment ”the majority of research on QoS is systems oriented, focusing on traffic analysis, scheduling, and routing. Relatively minor attention has been paid to user-level QoS issues.” With the development of the Internet and the growing usage of applications and different services there is a need for a new approach, where users and their perceptions can get even more involved in the final result of a quality evaluation. Even though this paper is a first approach to establish a mapping between objective and perceived QoS in the context of Internet commerce, and the term QoE is not used, it can be considered as the first attempt to incorporate a new way to evaluate the set of user perceptions regarding new services offered by the Internet. After this article, in [24], the QoE is defined as ”the totality of the Quality of Service mechanisms, provided to ensure smooth transmission of audio and video over IP networks,” which highlights the interest of the telecommunications sector in multimedia content and its effect in a world based on IP networks.

In the same line of thought, Heddaya [25] presents the Internet and its penetration level as the key factors to evolve from the old conception of QoS to a new concept, where there is a clearer need to separate the internal aspects of the network, beyond the control of the user, from the perceptible results delivered to the user by the network and its content. However, there is no mention of the effect of interfaces and presentation formats on the user's perception. In 2003, different researchers, such as Siller and Woods [26], proposed frameworks to evaluate QoE using QoS metrics, network feedback and user requirements. At the same time, Siller and Wood proposed a definition for QoE where the effect of the application/interface layer over the user's perception is remarked: "QoE is the user's perceived experience of what is being presented by the Application Layer, where the application layer acts as a user interface front-end that presents the overall result of the individual Quality of Services." In fact, this article states that QoS required and perceived by the user can be specified as a single parameter: low, fair, good and excellent, while, by contrast, the user requirements can also be specified by several parameters such as resolution, height, width, colour, etc., directly linked to the application layer and the QoE evaluation. With this new idea in the air, different researchers have tried to establish mechanisms to deepen the understanding and evaluation of user's perception [27], [28]. Some others have attempted to adjust technical parameters, related to QoS, considering the results generated by assessment tests [29]. The impact of the QoE over wireless infrastructures has been evaluated ([29]-[31]) while other researchers have talked about the growing commercial and economic importance of QoE applied in the distribution of different types of content.

As a result of the different research efforts focused on QoE, ITU decided, in 2007, to incorporate within the recommendation P.10/G.100 [9] a standard definition for QoE "The overall acceptability of an application or service, as perceived subjectively by the end-user." In the same recommendation, ITU considers that the overall acceptability may be influenced by user expectations and context, and includes complete end-to-end system effects (client, terminal, network, services infrastructure, etc.). On the other hand, when ITU defines QoS in the recommendation E.800 [10] as the "Totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service," claims that QoE measures the effect that a service or application has in the user, considering the external factors, as well. In contrast, QoS, with its point of view focused on the network performance is seen as one of the factors, together with the Grade of Service (GoS), the environmental aspects, the user profile and the Quality of Resilience (QoR), which affects the user's perception assessed in terms of QoE (Figure 1). Based on this separation, current research looks for the development of a well-defined methodology that allows establishing a clearer measurement of the user's perception in order to incorporate these results into the technical adjustments related to the network performance, to achieve the desired level of user's satisfaction based on the type of infrastructure and the applications running over it.

Nevertheless, and as mentioned by Moor et al. [7] Thakolsri et al. [4] and by Stankiewicz et al. [12], "literature on QoE and its related concepts (such as Quality of Service, User Experience), is rather fragmented. As a result, it is still largely unknown which factors affect the mobile QoE and how users' subjective experiences of such applications and services could be adequately identified and optimized."

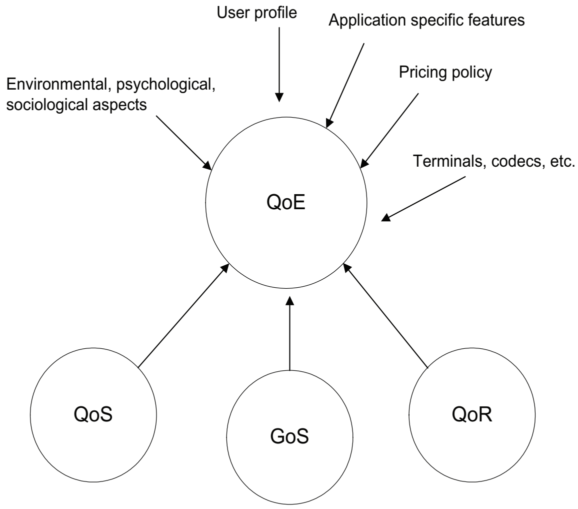

User profile
Application specific features
Pricing policy
Environmental, psychological, sociological aspects
Terminals, codecs, etc.
QoE
QoS
GoS
QoR

Figure 1. Elements influencing QoE [11]

III. QOE AND THE USER-CENTRIC APPROACH

An attempt to trace out the historical evolution of HCI and its relation with the concept of QoE might start by mentioning the research done by Card, Moran and Newell's, who in the book "The Psychology of Human-Computer Interaction"(PHCI) [32] proposed an empirically based cognitive theory of a skilled HCI and applied it to the specific problem of text editing. They discussed the processes involved, the techniques to use and the methods to follow when human factors research is performed. Finally, the Goals, Operators, Methods, and Selection rules model (GOMS) is proposed, while exploring ways in which this proposal can be extended and used to predict performance in other task-related areas. In [33], Newel and Card extended the vision proposed in [32], restating the importance of the psychological science in the design and development of interfaces, but remarking, as well, the need to provide engineering tools for this science in order to make less marginal its influence in the HCI area. While it is claimed that there is a chance to incorporate user's perception into the daily work of HCI, the need for the development of mental and cognition models that help to adjust the interfaces design is also shown. As a pioneering work on cognitive engineering models, Newel and Card's arguments were not without critics. Some authors, such as Newell et al. [34] and Carroll et al. [35], remarked on the gaps in the understanding of the whole process of interacting with computers. Even if the utility of the GOMS model in both initial design, evaluation and training is recognized, the lack of a deeper description of the user's context and its effect on the implementation of the model in a real design scenario was pointed out as its main weakness [36]. Another point of discussion was the hard science interpretation of the PHCI framework. To Caroll and Campbell, rather than reducing everything to a ”monolithic view of conflict between hard computer science and soft psychology” [34], an interdisciplinary field of HCI was taking form. According to their point of view, only a joint effort between science and psychology would allow the development of research areas like artificial intelligence and rapid prototyping. One of the proposed areas of cooperation was the design of interfaces, where psychologists would play an effective role addressing the questions that designers need answered, and contributing to the analysis involved in the design process of new products, but without touching the level of numerical quantification proposed by Card et.al. Meanwhile, in [37], the importance of the a deeper comprehension of the user’s understanding process based on the application of cognitive theories was remarked by Booth, with the consideration of making this new framework more accessible to designers. At this moment, while a continuous analysis of the role of HCI and the future of this area is under discussion, the potential effect of some elements of this area of study in infrastructure issues, or content management is not considered or studied.

Despite the discussed weaknesses, over the years different studies ([38]-[48]) have developed and extended the use of GOMS as a cost-effective way of evaluating designs without the participation of end users in human-computer interaction fields. On the other hand, this evolution led to the gradual consolidation of the soft science perspective in the PHCI, as mentioned by Holleis et al. [49], where most of the papers published in the HCI area from 1990s became more oriented to show case study, field experiment or field study than research work based on lab experiments. Within this new panorama, the input of Norman and Card work is reflected in the development of different cognitive modelling approaches and the consolidation of a multidisciplinary HCI, but without the idea of building a monolithic science with the integration of PHCI within an engineering conception. So far, only the importance of cognitive processes in the growth of HCI has been treated. From here, some similarities between the evolution of the QoE techno-centric approach and the process described above can be identified. However, the lack of multidisciplinary work enhancing the development of more complete models to link user experience and resource allocation is still evident. Up to now, most of the communication network deployments have been done taking into account economic, technical and ecological considerations with the user’s satisfaction regarding content provided through these infrastructures will be reached only by having better technology and higher bandwidth. Before this fact, HCI and the research about user’s comprehension might become a new tool to develop infrastructures providing, in a smart and efficient way, content and information to a user fulfilling expected levels of quality.

In tune with the evolution of the soft science concept, the gradual consolidation of a user-centric approach has allowed the growth of new and complimentary areas within HCI such as Experience Design (XD) and User Experience (UX), which can be considered in the future as providers of judgement elements to clarify the QoE concept in the techno-centric approach. In recent years, users have had more chances to choose among multiple options with different levels of design, complexity and innovation. This has empowered users, who have become more demanding and critical, and the HCI field has not been immune to this phenomenon. According to Stankiewicz et al. [12], during the ’70-’80s, people involved in HCI was focused on understanding the way that people thought and processed information in order to increase the efficiency and provide more functionality in their solutions. However, at this moment, the users’ expectations and experiences were not considered. Some efforts to involve people in the development and design of HCI solutions were done from the late 80’s and early 90’s with the origin of participatory design and contextual design [12]. But since the late 90’s, with the wider presence of computers and technology, more importance is given to evaluate how this success of technology adoption and diffusion is explained. There were efforts to incorporate aspects such as beauty, enjoyment, or fun, into HCI in general and usability engineering specifically ([50][51]). These approaches have three aspects in common: ”a focus on the subjective side of usability, namely user perceptions and experiences; on the positive sides of using products (instead of simply avoiding usability problems), and on human needs as a whole [52]”. At the same time, there was an identical shift from a more R&D-driven ’push’-oriented mentality towards a more (marketing-driven) ’pull’-oriented stance in which the user became the starting point of the technology development [12]. User is now the king, and this consumer-oriented mentality as is defined by Edwards [7], is orienting the efforts to measure the user experience and reflect it in the provision of high quality. Under these conditions, the experience of a user regarding a device, product or interface gains more importance, despite the differences expressed by UX and XD experts. In words of Marc Hassenzahl [53], UX is focused on usage and only rooted in action, while XD is a way to create experiences considering with more interest the history behind what the user experience. In certain way, UX has seen focused on how a person feels about using a system, considering the external things that can affect this experience (i.e., brand, cost of the system, image, ease of use, etc) [54], but with the introduction of usability in the design process. Most of the efforts in the area seems oriented to the design of ways to interact with computers, but with a failure to understand how information is communicated to a person and how they interact with and interpret that information to accomplish their goal [55].

IV. CHALLENGES AND FUTURE WORK

An attempt to trace out the historical evolution of HCI and its relation with the concept of QoE might start by mentioning the research done by Card, Moran and Newell’s, who, in the book “The Psychology of Human-Computer Interaction” (PHCI) [32], proposed an empirically based cognitive theory of a skilled HCI and applied it to the specific problem of text editing. They discussed the processes involved, the techniques to use and the methods to follow when human factors research is performed. Finally, the Goals, Operators, Methods, and Selection rules model

(GOMS) is proposed, while exploring ways in which this proposal can be extended and used to predict performance in other task-related areas. Technology-centric interpretations of QoE go hand-in-hand with the assumption that by optimizing the QoS, the end user's QoE will also increase. However, this is not always the case: Even with excellent QoS, QoE can be really poor [8]. These gaps are usually caused by a lack of insight in the totality of dimensions of a customer's experience, and here is where HCI can offer the tools to complete the development of a structured QoE system of assessment and implementation where users are really involved. Some authors, like Stankiewicz et al. [12], claim that it is necessary to involve users in certain stages of the development process of a new technology or application, but there is no complete clarity about issues like the right stage of the process to involve their opinions, and the type of users that should be involved, etc. [12]. Another fact to mention is pointed out by Stankiewicz et al. [12], when they say that QoE ”is usually measured in terms of technical metrics (QoS), ignoring the fact that the ultimate goal should not be to deliver applications with the most advanced features, but to deliver products that will ensure a good Quality of Experience.” On the other hand, a challenge for HCI is to understand how layers of underlying technological infrastructure that may not be designed with the full range of human-centred concerns in mind work, and, based on this knowledge, adjust to these constraints to maintain the user experience at the highest level. Being UX subjective, the user is the centre of the whole system, and his/her opinions and concepts will determine the adjustment of technical features inside the network. However, there is also a need for an understanding of those external factors that can influence the final perception and experience of the user. Here, is where the QoE concept becomes “the picture to measure the perceived connection quality in the current context” [54]. Considering the multiple aspects that affect the overall experience, evaluators need to understand the whole picture and identify the reasons behind each good or bad experience. About this last point, and how Edwards [7] expresses “experience has a multi-dimensional character, where some authors highlight the importance of emotions, expectations, and the relationship to other people and the context of use, while others remarks the importance of the broader context.” The challenge with respect to the QoE area is to look for the way to combine both points of view (technical and user-centric) so that, when seeking the satisfaction of a user with a specific content or application, we have a broad understanding of how complex human beings are. As mentioned by Kellerer et al. [5] “we need to develop practices that allow infrastructure and interaction features to be co-designed.” But, a challenge for the HCI community is to create communication bridges with other disciplines to become real a scenario where, effectively, the use of technical resources and user satisfaction work hand-in-hand.

Regarding deployment of mobile networks, recent introduction of new generation of wireless infrastructures is being accompanied by an increase in both the number of users and their interest in multimedia content. This growth has been driven in the last decade by the popularity of multimedia content (e.g., video-sharing websites, social networks, video on demand sites, mobile IPTV, etc.), that according to the tendency will generate much of the mobile traffic growth through 2016, showing, at the same time, the highest growth rate of any mobile application. From this point of view, mobile operators have to tackle increasing operational costs given by energy consumption due to the traffic growth. From the users perspective, it represents the need for the development of mechanisms to extend mobile terminals life to enjoy during more time multimedia content with higher quality level. In that sense, classical approaches like deploying additional infrastructure are not likely to be economically viable for this challenge. On the other side, severe resource limitations in mobile networks can lead to dramatic levels of delays and interruptions, which can significantly affect user perceived experience (QoE). In this scenario, the need for obtaining improvements in terms of the quality perceived by users is more and more important in the networks evolution scenario. An alternative way to improve the QoE is having networks capable of identifying users expectations and using this information to dynamically allocate resources adjusted to a semantic model of the mobile service requirements while the content is being processed in the user terminal. An it is here where a better understanding of user's perceptions might contribute to the creation of network infrastructures with better performance based on the evaluation of predefined QoE model.

REFERENCES

[1] E. Israelski and A. M. Lund, “The human-computer interaction handbook,” J. A. Jacko and A. Sears, Eds. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 2003, ch. The evolution of human-computer interaction during the telecommunications revolution, pp. 772–789. [Online]. Available: http://dl.acm.org/citation.cfm?id=772072.772121

[2] G. Aristomenopoulos, T. Kastrinogiannis, V. Kaldanis, G. Karantonis, and S. Papavassiliou, “A novel framework for dynamic utility-based qoe provisioning in wireless networks,” GLOBECOM 2010, 2010 IEEE Global Telecommunications Conference, pp. 1–6, 2010.

[3] E. Cerqueira, L. Veloso, M. Curado, and P. M. E. Monteiro, “Quality level control for multi-user sessions in future generation networks,” Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008, pp. 1–6, 2008.

[4] S. Thakolsri

[5] , W. Kellerer, and E. Steinbach, “Qoe-based cross-layer optimization of wireless video with unperceivable temporal video quality fluctuation,” in ”IEEE International Conference on Communications (ICC 2011)”, Kyoto, Japan, 2011.

[6] W. K. Edwards, HCI Remixed: Essays on Works That Have influenced the HCI Community. The MIT Press, 2008, ch. Infrastructure and Its Effect on the Interface, pp. 119–122.

[7] W. K. Edwards, M. Newman, and E. Poole, “The infrastructure problem in HCI,” in Proceedings of the SIGCHI Conference on Human Factors in Computer Systems (CHI 2010), 2010.

[8] K. De Moor, I. Ketyko, W. Joseph, T. Deryckere, L. De Marez, L. Martens, and G. Verleye, “Proposed framework for evaluating quality of experience in a mobile, testbed-oriented living lab setting,” Mobile Networks and Applications, vol. 15, pp. 378–391, 2010, 10.1007/s11036-010-0223-0. [Online]. Available: http://dx.doi.org/10.1007/s11036-010-0223-0

[9] K. De Moor, W. Joseph, I. Ketyk´o, E. Tanghe, T. Deryckere, L. Martens, and L. De Marez, “Linking users' subjective qoe evaluation to signal strength in an ieee 802.11b/g wireless lan environment,”

EURASIP J. Wirel. Commun. Netw., vol. 2010, pp. 6:1–6:10, Apr. 2010. [Online]. Available: http://dx.doi.org/10.1155/2010/541568

[10] "ITU-T recommendation P.10/G.100. Vocabulary and effects of transmission parameters on customer opinion of transmission quality,"Jul.2008.[Online].Available:http://www.itu.int/itut/recommendations/index.aspx?ser=P

[11] "ITU-T recommendation E.800. Quality of telecommunication services: concepts, models, objectives and dependability planning. Terms and definitions related to the quality of telecommunication services," Sep. 2008.

[12] R. Stankiewicz, P. Cholda, and A. Jajszczyk, "Qox: What is it really?" Communications Magazine, IEEE, vol. 49, no. 4, pp. 148 – 158, april 2011.

[13] L. De Marez and K. De Moor, "The Challenge of User-And QoE-Centric Research and Product Development in Today's ICT-Environment," Observatorio (OBS*), vol. 1, no. 3, pp. 1–22, 2007. [Online].Available:http://www.obs.obercom.pt/index.php/obs/article/view/141/101

[14] F. F. Fowle, "Toll telephone traffic," American Institute of Electrical Engineers, Transactions of the, vol. XXXIII, no. 2, pp. 1263 –1272, june 1914.

[15] L. Pocock, "A survey of the telephone transmission-rating problem," Electrical Engineers - Part III: Radio and Communication Engineering, Journal of the Institution of, vol. 95, no. 36, pp. 253 – 265, july 1948.

[16] J. Das, "Some effects of bandwidth limitation on p.s.m., p.l.m., and p.p.m. signals," Proceedings of the IEE - Part C: Monographs, vol.109, no. 16, pp. 646 –655, september 1962.

[17] M. Eisen, "On switching problems requiring queuing theory in computer-based systems," Communications Systems, IRE Transactions on, vol. 10, no. 3, pp. 299 –303, september 1962.

[18] M. Pennotti and M. Schwartz, "Congestion control in store and forward tandem links," Communications, IEEE Transactions on, vol. 23, no. 12, pp. 1434 – 1443, dec 1975.

[19] D. Goodman and C.-E. Sundberg, "Quality of service and bandwidth efficiency of cellular mobile radio with variable-bit-rate speech transmission," in Vehicular Technology Conference, 1983. 33rd IEEE, vol. 33, may 1983, pp. 316 – 321.

[20] J. Gruber, E. Abdou, P. Richards, and G. Williams, "Quality-of-service in evolving telecommunications networks," Selected Areas in Communications, IEEE Journal on, vol. 4, no. 7, pp. 1084 – 1089, oct 1986.

[21] J. Richters and C. Dvorak, "A framework for defining the quality of communications services," Communications Magazine, IEEE, vol. 26, no. 10, pp. 17 –23, oct. 1988.

[22] H. Brander, J. Howard, G. Menicou, S. Plagemann, E. Protonotarios, and M. Theologou, "Quality of service in broadband communications," in Integrated Broadband Services and Networks, 1990., International Conference on, oct 1990, pp. 166 –171.

[23] Y. B. Kim and A. Vacroux, "Real-time packet network analysis for iso/osi performance management," in Global Telecommunications Conference, 1990, and Exhibition. 'Communications: Connecting the Future', GLOBECOM '90., IEEE, dec 1990, pp. 397 –401 vol.1.

[24] A. Bouch, A. Kuchinsky, and N. Bhatti, "Quality is in the eye of the beholder: meeting users' requirements for internet quality of service," in Proceedings of the SIGCHI conference on Human factors in computing systems, ser. CHI '00. New York, NY, USA: ACM, 2000, pp.297–304.[Online].Available: http://doi.acm.org/10.1145/332040.332447

[25] T. O'Neil, "Quality of experience and quality of service for IP video conferencing." Polycom, Tech. Rep., 2002. [Online]. Available: http://hive2.hive.packetizer.com/users/h323forum/papers/polycom/QualityOfExperience+ServiceForIPVideo.pdf

[26] A. Heddaya, "An economically scalable internet," IEEE Computer, vol. 35, no. 9, pp. 93–95, 2002.

[27] M. Siller and J. Woods, "Qos arbitration for improving the qoe in multimedia transmission," in Visual Information Engineering, 2003. VIE 2003. International Conference on, july 2003, pp. 238 – 241.

[28] A. Perkis, S. Munkeby, and O. Hillestad, "A model for measuring quality of experience," in Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic, june 2006, pp. 198 – 201.

[29] E. Gallo, M. Siller, and J. Woods, "An ontology for the quality of experience framework," in Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on, oct. 2007, pp. 1540 –1544.

[30] D. Soldani, "Means and methods for collecting and analyzing qoe measurements in wireless networks," in World of Wireless, Mobile and Multimedia Networks, 2006. WoWMoM 2006. International Symposium on a, 0-0 2006, pp. 5 pp. –535.

[31] O. Bradeanu, D. Munteanu, I. Rincu, and F. Geanta, "Mobile multimedia end-user quality of experience modeling," in Digital Telecommunications, , 2006. ICDT '06. International Conference on, aug. 2006, p. 49.

[32] S. Qiu, H. Rui, and L. Zhang, "No-reference perceptual quality assessment for streaming video based on simple end-to-end network measures," in Networking and Services, 2006. ICNS '06. International conference on, july 2006, p. 53.

[33] S. K. Card, A. Newell, and T. P. Moran, The Psychology of Human-Computer Interaction. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1983.

[34] A. Newell and S. K. Card, "The prospects for psychological science in human-computer interaction," Hum.-Comput. Interact., vol. 1, no. 3, pp. 209–242, Sep. 1985.

[35] J. M. Carroll and R. L. Campbell, "Softening up hard science: Reply to newell and card," SIGCHI Bull., vol. 19, no. 1, pp. 74–, Jul. 1987. [Online]. Available: http://dl.acm.org/citation.cfm?id=28189.1044812

[36] J. R. Olson and G. M. Olson, "The growth of cognitive modeling in human-computer interaction since goms," Human-Computer Interaction, vol. 5, no. 2, pp. 221–265, Jun. 1990.

[37] D. E. Kieras and T. P. Santoro, "Computational goms modeling of a complex team task: lessons learned," in Proceedings of the SIGCHI conference on Human factors in computing systems, ser. CHI '04. New York, NY, USA: ACM, 2004, pp. 97–104.

[38] P. A. Booth, "Errors and theory in human-computer interaction," Acta Psychologica, vol. 78, no. 13, pp. 69 – 96, 1991.

[39] J. L. Drury, J. Scholtz, and D. Kieras, "Adapting goms to model human-robot interaction," in Proceedings of the ACM/IEEE international conference on Human-robot interaction, ser. HRI '07. New York, NY, USA: ACM, 2007, pp. 41–48.

[40] B. E. John and D. E. Kieras, "Using goms for user interface design and evaluation: which technique?" ACM Trans. Comput.-Hum. Interact., vol. 3, no. 4, pp. 287–319, Dec. 1996.

[41] B. John, A. Vera, M. Matessa, M. Freed, and R. Remington, "Automating cpm-goms," in Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves, ser. CHI '02. New York, NY, USA: ACM, 2002, pp. 147–154.

[42] S. Pronovost and R. L. West, "A goms model of virtual sociotechnical systems: using video games to build cognitive models," in Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction, ser. ECCE '08. New York, NY, USA: ACM, 2008, pp. 15:1–15:2.

[43] D. E. Kieras and T. P. Santoro, "Computational goms modeling of a complex team task: lessons learned," in Proceedings of the SIGCHI conference on Human factors in computing systems, ser. CHI '04. New York, NY, USA: ACM, 2004, pp. 97–104.

[44] T. Jokela, J. Koivumaa, J. Pirkola, P. Salminen, and N. Kantola, "Methods for quantitative usability requirements: a case study on the development of the user interface of a mobile phone," Personal Ubiquitous Comput., vol. 10, no. 6, pp. 345–355, Sep. 2006.

[45] F. Magrabi, "Using cognitive models to evaluate safety-critical interfaces in healthcare," in CHI '08 extended abstracts on Human

factors in computing systems, ser. CHI EA ’08. New York, NY, USA: ACM, 2008, pp. 3567–3572.

[46] C. Haimson and J. Grossman, “A gomsl analysis of semi-automated data entry,” in Proceedings of the 1st ACM SIGCHI symposium on Engineering interactive computing systems, ser. EICS ’09. New York, NY, USA: ACM, 2009, pp. 61–66.

[47] H. Li, Y. Liu, J. Liu, X. Wang, Y. Li, and P.-L. P. Rau, “Extended klm for mobile phone interaction: a user study result,” in Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems, ser. CHI EA ’10. New York, NY, USA:ACM, 2010, pp. 3517–3522.

[48] E. Abdulin, “Using the keystroke-level model for designing user interface on middle-sized touch screens,” in Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems, ser. CHI EA ’11. New York, NY, USA: ACM, 2011, pp. 673–686.

[49] P. Holleis, M. Scherr, and G. Broll, “A revised mobile klm for interaction with multiple nfc-tags,” in Proceedings of the 13th IFIP TC 13 international conference on Human-computer interaction - Volume Part IV, ser. INTERACT’11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 204–221.

[50] T. Clemmensen, “Whatever happened to the psychology of humancomputer interaction?: A biography of the life of a psychological framework within a hci journal,” Information Technology & People, vol. 19, pp. 121 – 151, 2006.

[51] S. W. Draper, “Analysing fun as a candidate software requirement, ”Personal Technology, vol. 3, pp. 117–122, 1999.

[52] M. Hassenzahl, “The effect of perceived hedonic quality on product appealingness,” Int. J. Hum. Comput. Interaction, vol. 13, no. 4, pp. 481–499, 2001.

[53] ——, “The interplay of beauty, goodness, and usability in interactive

[54] products,” Human-Computer Interaction, vol. 19, no. 4, Dec. 2008.

——, User Experience and Experience Design. Aarhus, Denmark: The Interaction-Design.org Foundation, 2011.

[55] M. Fiedler, K. Kilkki, and P. Reichl, “09192 abstracts collection – from quality of service to quality of experience,” in From Quality of Service to Quality of Experience, ser. Dagstuhl Seminar Proceedings,

[56] M. M. J. Albers, “Human-information interaction,” in Proceedings of the 26th annual ACM international conference on Design of Communication, ser. SIGDOC ’08. New York, NY, USA: ACM, 2008, pp. 117–124.

A Helpful Positioning Method with Two GNSS Satellites in Urban Area

Hiroyuki Hatano
Graduate School of Engineering,
Utsunomiya University
Tochigi, Japan
e-mail: hatano@is.utsunomiya-u.ac.jp

Tomoya Kitani
Graduate School of Informatics,
Shizuoka University.
Shizuoka, Japan
e-mail: t-kitani@ieee.org

Masahiro Fujii, Yu Watanabe
Graduate School of Engineering,
Utsunomiya University
Tochigi, Japan
e-mail: {fujii, yu}@is.utsunomiya-u.ac.jp

Hironobu Onishi
Graduate School of Engineering,
Shizuoka University
Shizuoka, Japan
e-mail:onishi@hatanolab.eng.shizuoka.ac.jp

***Abstract*—For estimating user's location, Global Positioning System (GPS) is very popular and important technique. The GPS estimation uses satellites which fly in the air. The estimator needs the open sky and multiple satellites (usually 4 and more satellites). However, there are many buildings in urban area. The receiver tends to receive bad signals, called as non-line-of-sight (NLOS) or multipath signals. The problem is that the receiver cannot get direct path signals from adequate number of the satellites coinstantaneously. This case leads to degrade estimation quality. So, we introduce the novel estimation algorithm, which can estimate own position with as low number of satellites as possible.**

Keywords-GPS; Global Positioning System; GNSS; Localization; Multipath; NLOS; Shielded signals

I. Introduction

For next generation, a plan of Intelligent Transportation System (ITS) is very attractive. ITS can resolve problems such as traffic accidents and jam. By realizing ITS, convenient and safety life will be turn up. As a hot topic, electrical vehicles and hybrid vehicles are famous. Compared to fossil fuel, the electrical vehicles can miniaturize vehicle's size. Then, personal mobility, such as "Segway" and "Winglet", can be provided [1].

In development of high performance vehicles, own location is one of important information. The most popular technique is Global Positioning System (GPS). GPS can estimate user's location by receiving signals from GPS satellites. In our life, we have been getting benefits of GPS already. Route guidance by car navigation system or smart phones is very useful. However, the estimation by GPS still has problems [2], [3]. Especially, we cannot get own position precisely in urban area. In urban area, there are a lot of structures such as buildings. So, GPS receivers receive the signals from the satellites via multipath propagation. Also, the receivers cannot watch the satellites in line-of-sight (LOS) when the satellites are shield by the structures. That is the case called as non-line-of-sight (NLOS). In case of riding the personal mobility and walking on street, users exist near building. The harmful effect such as receiving multipath or shielded signals often occurs. The quality of positioning becomes worse. So, it is important to realize robust and accurate positioning even if the above bad environment. Ideally, it is important to be able to estimate own position everywhere. The above robust and accurate positioning leads to high functional drivers assistant systems, such as supporting every each of road lanes or automatic drivng [4]–[6]. Moreover, the application to probe car systems for collection of various data, road-to-vehicle / vehicle-to-vehicle communication will be realized effectively [7]–[9]. In our research, we focus on the positioning algorithm under such a bad environment of receiving multipath or NLOS signals.

In GPS positioning, the receiver estimates the own position from the propagated distance between the LOS satellites and the receiver [10]. The distance can be derived from both a transmitted time at the satellite and a received time at the receiver. So, the estimated parameters are four parameters, that is 3 dimension position and time clock error between the satellite and receiver. The estimator needs four and more satellites. In urban area, the number of satellites which can observe in LOS fluctuates. Unfortunately, there are the places where we cannot observe four satellites frequently. In this paper, we show the novel positioning method which can estimate own position with as low number of satellites as possible. Our estimation uses two LOS satellites. Also, we assume a moving vehicle. So, we use a travelling distance as sensor data too. In our method, we can estimate user's position even if the number of LOS satellites is reducing.

This paper is organized as follows. In Section II, we introduce the related works. And, we confirm the problem of the urban positioning. In Section III, we introduce the coordinate systems which is required in our positioning algorithm. Then, in Section IV, we introduce the our positioning algorithm and performance. Finally, Section V summarizes the paper.

II. Related works and current problems

In this section, we will introduce the related works. There are a lot of works which can improve the positioning performance in urban area. After the introduction, the simple experiment, which we confirm the problem of the urban positioning, will be shown.

A. Related works

The propagation environment in urban area is multipath propagation. There are a lot of works which can mitigate the multipath interference. For example, techniques which focus on signal tracking on a receiver or channel estimation by multiple correlators have been shown [11], [12]. Moreover, the mitigation methods of Direction-of-Arrival (DoA) estimation by array antenna have also been shown [13], [14]. The purpose of these methods is to prevent the degradation of the positioning performance in case of the multipath propagation with LOS satellites.

In our research, we focus on the multipath environment with the NLOS satellites. That is, the direct signal is shielded by some structures such as buildings. This situation may often occur in urban area. One of the researches under assumption of NLOS is [15]. Meguro et al. [15] uses a camera which watches the sky. The camera decides if there is the structure or not.

If the receiver exists at static position, observed satellites are always the same satellites. However, in case of a moving receiver in urban area, the observed satellites are changed frequently. The changing the satellites leads to the fluctuation of the positioning accuracy. Irie's work [16] is to prevent such a problem of the accuracy degradation. Kawamura and Tanaka's work [17] is also to keep the accuracy good by calculating weightings to the satellites.

When the number of satellites which can be observed is reduced, the positioning accuracy becomes worse. In the worst case, the receiver cannot estimate own position at all. This is because the number of observed satellites is few. In Fan et al.'s work [18], the receiver adds the virtual satellites by using the reflected signals at the ground.

As other approaches, the estimator can keep the positioning accuracy fine by using other information and devices. Map matching techniques may be most famous. The devises such as camera, direction, gyro or travelling distance sensors can improve the positioning performance. The vehicle-to-vehicle or vehicle-to-road communication can be also used for improvement. The above researches are shown in [4], [8], [9], [19].

In Japan, in order to improve the positioning accuracy in urban area and the environment decreasing the number of observable LOS satellites, the project of Quasi-Zenith Satellite System (QZSS) is under way [20]–[22]. In September 2010, the first QZSS "MICHIBIKI" was launched. The satellite of QZSS moves on quasi-zenith orbit of Japan. So, the receiver can observe QZS at high angle of elevation. Even if there are a lot of buildings, the QZS can tend to become the LOS satellite.

B. Checking current probrem and our approach

In urban area, there are a lot of structures such as buildings. It leads to the bad effect of both reduction of the observable LOS satellites and generation of multipath propagation. This bad effect generates large positioning error. In order to confirm the positioning performance and its problem, we tested simple experiments of the positioning in urban area. We used GPS logger (model: m-241, made by HOLUX), which is popular and we can buy at stores easily. The frequency of records is 1 Hz. We tried estimation a static position in 10 minutes. The receiver's position and estimated positions are summarized in Figure 1. The positions are plotted to Google Map by a useful plotting tool "Wadachi Ver.3.44".

The location in the experiment is near Hamamatsu Station in Shizuoka, Japan. There are a lot of buildings. We set the logger on the edge of the street, that is by the side of the building wall. From Figure 1, the estimated positions have about 20-30m error compared to the true position of the logger. We can confirm that the positioning in urban area has large errors. The reason may be NLOS or multipath environment.

The bad satellites, which have NLOS path or multipath should not be used to estimate the receiver's position. However, the estimate algorithm needs four or more satellites. So, conventionally, in order to get adequate number of the satellites, the estimator needs to use the bad satellites too.

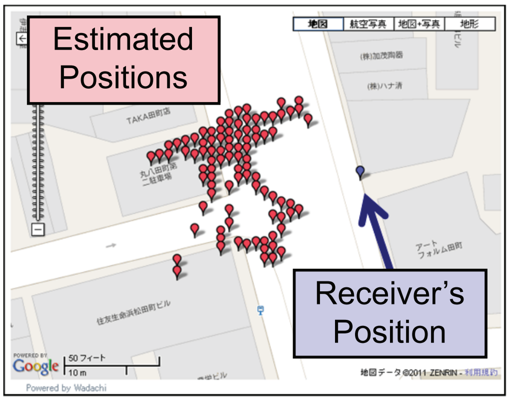


Figure 1. Positioning results in urban area (front of building)

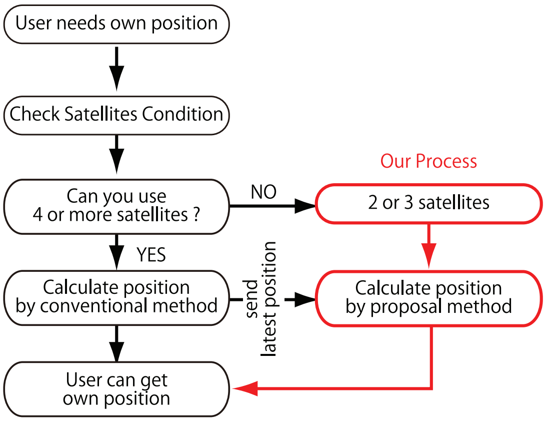

User needs own position
Check Satellites Condition
Our Process
Can you use 4 or more satellites ?
NO
2 or 3 satellites
YES
send latest position
Calculate position by conventional method
Calculate position by proposal method
User can get own position

Figure 2. Process flow in positioing

In our approach, we avoid using these bad satellites. So, we consider another algorithm which can estimate own position with less than four satellites. In this paper, we introduce the estimation algorithm which uses two satellites and the information of the receiver's movement. We assume the positioning of a moving vehicle. So, the receiver can get the information of the travelling distance.

The flow of the usage is illustrated in Figure 2. Usually, the receiver estimates own positions by conventional algorithm, that is with four and more satellites. In case of bad environment, we continue the positioning by selecting just two satellites which have high elevation angle. Our approach may help in the robust positioning everywhere.

III. Coordinate system (ECEF, ENU coordinate)

In this section, before the introduction to our positioning algorithm, the coordinate systems which are used in the paper will be shown. The conversion equation will be also shown. The coordinate systems are both ECEF and ENU. Both coordinates illustrate as Figure 3. Our algorithm, which is presented in Section IV, use the assumption "the altitude of the receiver do not change in short time". For applying the assumption, the conversion of the coordinate is needed.

ECEF stands for Earth-Centered Earth-Fixed. The origin of ECEF is the center of the earth, that is the point O_{enu} in

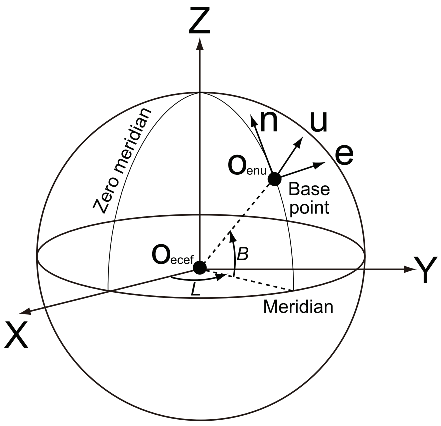


Figure 3. ECEF and ENU coordinate system

Figure 3. ECEF is $x-y-z$ orthogonal coordinates on the fixed earth. The unit of x, y, z is meter.

ENU means East-North-Up. As pointing out in Figure 3, each of coordinates e, n, u means east-direction, north-direction, vertical(up)-direction respectively. The origin of ENU is defined as a arbitrary base position, such as the position O$_{\text{enu}}$ in Figure 3. The unit of x, y, z is also meter.

Next, we introduce the conversion equation. We define a ECEF position (x, y, z) as a vector $\vec{x}_{\text{ecef}}$, and also define a ENU position (e, n, u) as $\vec{x}_{\text{enu}}$. The conversion equation from the ECEF position $\vec{x}_{\text{ecef}}$ to the ENU position $\vec{x}_{\text{enu}}$ is the following:

$$\vec{x}_{\text{enu}} = \mathbf{R}(B, L) \cdot [\vec{x}_{\text{ecef}} - \vec{x}_{0,\text{ecef}}] \tag{1}$$

where the vector $\vec{x}_{0,\text{ecef}}$ means the base position which is expressed in the ECEF coordinate. Generally, the matrix $\mathbf{R}(B, L)$ is called as rotation matrix. In (1), in order to rotate the coordinate, the matrix $\mathbf{R}(B, L)$ is used. So the matrix is:

$$\mathbf{R}(B, L) = \begin{pmatrix} -\sin L & \cos L & 0 \\ -\cos L \sin B & -\sin L \sin B & \cos B \\ -\cos L \cos B & \sin L \cos B & \sin B \end{pmatrix} \tag{2}$$

where the parameter B is the degree of latitude and the parameter L is the degree of longitude at the base position $\vec{x}_{0,\text{ecef}}$ (Figure 3). By using (1), we can convert positions from ECEF to ENU.

In this paper, we apply World Geodetic System 84 (WGS-84) as a geographical coordinate system. Then, the semi-major axis R_e of the earth and the oblateness f_e are:

$$R_e = 6378137.0[m] \tag{3}$$

$$f_e = \frac{1.0}{298.257223563}. \tag{4}$$

When we want to convert from ENU to ECEF, the conversion equation is the following:

$$\vec{x}_{\text{ecef}} = \mathbf{R}^{-1}(B, L) \cdot \vec{x}_{\text{enu}} + \vec{x}_{0,\text{ecef}} \tag{5}$$

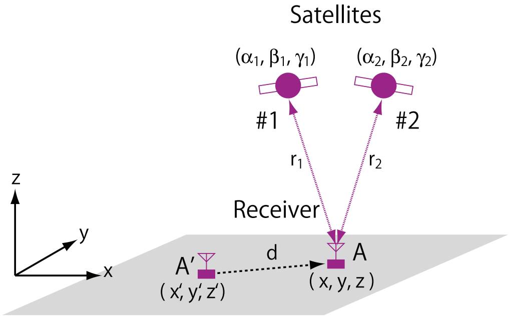


Figure 4. System model (positioning algorithm with two satellites)

where the inverse matrix $\mathbf{R}^{-1}(B, L)$ is negative rotation to (1) and can be expressed as

$$\mathbf{R}^{-1}(B, L) = \begin{pmatrix} -\sin L & -\cos L \sin B & \cos L \cos B \\ \cos L & -\sin L \sin B & \sin L \cos B \\ 0 & \cos B & \sin B \end{pmatrix}. \tag{6}$$

IV. Positioning algorithm using two satellites

In this section, we present the positioning algorithm which can estimate with the two satellites. Usually, the conventional estimator needs four and more satellites. However, in urban area, the number of satellites which can observe in LOS fluctuates. Moreover, the place where we cannot observe four and more satellites often occurs. In these bad situations, our positioning method can estimate own position even if the number of LOS satellites is reducing.

A. System model

Our system model is shown in Figure 4. There are the two LOS satellites in the sky. Their ECEF coordinates are $(\alpha_1, \beta_1, \gamma_1), (\alpha_2, \beta_2, \gamma_2)$[m] respectively. The position of the receiver which should be estimated is (x, y, z)[m] and the position is shown as point A. It is ECEF coordinate. The parameter t [sec] means time at the receiver. The previous time is also denoted as the parameter t' and the relation is $t = t' + \Delta t$. The past receiver's position A' is (x', y', z')[m], that is at the time t'. The travelling distance between the time t' and t is d[m]. The time interval Δt is assumed as short time. So, we assume that the altitude of the receiver at t and t' is the same. The distance d can be gotten from the sensor which measures car speed. The parameter r_1[m] means the range between the satellite #1 and the receiver. The parameter r_2[m] also means the range to the satellite #2. These ranges r_1, r_2 are called as pseudorange because the clock error between the satellites and the receiver is added to the true range.

B. Algorithm explication

In order to estimate the receiver's position A, we enumerate the related equations as follows.

$$r_1 = \sqrt{(\alpha_1 - x)^2 + (\beta_1 - y)^2 + (\gamma_1 - z)^2} + s \tag{7}$$

$$r_2 = \sqrt{(\alpha_2 - x)^2 + (\beta_2 - y)^2 + (\gamma_2 - z)^2} + s \tag{8}$$

$$d = \sqrt{(x' - x)^2 + (y' - y)^2 + (z' - z)^2} \tag{9}$$

Equations (7), (8) are the pseudorange to the satellite #1 and #2, respectively. The parameter s means the error [m] of the clock difference between the satellites and the receiver. Equation (9) means the travelling distance of the movement of the receiver in the time interval Δt.

We must estimate both the receiver's position (x, y, z) and the clock error s from (7)$\sim$(9). Equations (7)$\sim$(9) are non-linear simultaneous equations. So, we try estimating the parameters by sequential approach as follows.

[Step 1] We set initial values x^0, y^0, z^0, s^0 of the unknown parameters x, y, z, s.

[Step 2] When the receiver's position and the clock error are x^0, y^0, z^0, s^0, the pseudoranges and the travelling distance can be denoted as the following:

$$r_1^0 = \sqrt{(\alpha_1 - x^0)^2 + (\beta_1 - y^0)^2 + (\gamma_1 - z)^2} + s^0 \tag{10}$$

$$r_2^0 = \sqrt{(\alpha_2 - x^0)^2 + (\beta_2 - y^0)^2 + (\gamma_2 - z')^2} + s^0 \tag{11}$$

$$d^0 = \sqrt{(x' - x^0)^2 + (y' - y^0)^2 + (z' - z)^2} \tag{12}$$

[Step 3] The residual errors between the observed value and the above r_1^0, r_2^0, d^0 are:

$$\Delta r_1^0 = r_1 - r_1^0 \tag{13}$$

$$\Delta r_2^0 = r_2 - r_2^0 \tag{14}$$

$$\Delta d^0 = d - d^0 \tag{15}$$

Starting from the initial values x^0, y^0, z^0, s^0, we update the estimating unknown parameters x, y, z, s as the above residual errors become zeros.

[Step 4] In order to derive the updating values $\Delta x, \Delta y, \Delta z, \Delta s$, we approximate (7), (8), (9) to the followings. That is, the non-linear simultaneous equations are transformed to linear simultaneous equations.

$$\Delta r_1 = \frac{\partial r_1}{\partial x}\Delta x + \frac{\partial r_1}{\partial y}\Delta y + \frac{\partial r_1}{\partial z}\Delta z + \frac{\partial r_1}{\partial s}\Delta s \tag{16}$$

$$\Delta r_2 = \frac{\partial r_2}{\partial x}\Delta x + \frac{\partial r_2}{\partial y}\Delta y + \frac{\partial r_2}{\partial z}\Delta z + \frac{\partial r_2}{\partial s}\Delta s \tag{17}$$

$$\Delta d = \frac{\partial d}{\partial x}\Delta x + \frac{\partial d}{\partial y}\Delta y + \frac{\partial d}{\partial z}\Delta z + \frac{\partial d}{\partial s}\Delta s \tag{18}$$

[Step 5] The above simultaneous equations can be represented as matrix forms as below:

$$\Delta\vec{r} = \mathbf{G}\Delta\vec{x} \tag{19}$$

, where

$$\Delta\vec{r} = [\Delta r_1, \Delta r_2, \Delta d]^{\mathrm{T}} \tag{20}$$

$$\Delta\vec{x} = [\Delta x, \Delta y, \Delta z, \Delta s]^{\mathrm{T}} \tag{21}$$

$$\mathbf{G} = \begin{pmatrix} \frac{\partial r_1}{\partial x} & \frac{\partial r_1}{\partial y} & \frac{\partial r_1}{\partial z} & \frac{\partial r_1}{\partial s} \\ \frac{\partial r_2}{\partial x} & \frac{\partial r_2}{\partial y} & \frac{\partial r_2}{\partial z} & \frac{\partial r_2}{\partial s} \\ \frac{\partial d}{\partial x} & \frac{\partial d}{\partial y} & \frac{\partial d}{\partial z} & \frac{\partial d}{\partial s} \end{pmatrix} \tag{22}$$

$$= \begin{pmatrix} \frac{-(\alpha_1 - x)}{r_1 - s} & \frac{-(\beta_1 - y)}{r_1 - s} & \frac{-(\gamma_1 - z)}{r_1 - s} & 1 \\ \frac{-(\alpha_2 - x)}{r_2 - s} & \frac{-(\beta_2 - y)}{r_2 - s} & \frac{-(\gamma_2 - z)}{r_2 - s} & 1 \\ \frac{-(x' - x)}{d} & \frac{-(y' - y)}{d} & \frac{-(z' - z)}{d} & 0 \end{pmatrix}. \tag{23}$$

[Step 6] Equation (19) has four unknown parameters. However, the row of the matrix $\mathbf{G}$ is three. So, we cannot solve the each of the unknown parameters. Then, we assume that the altitude values of the receiver A and A' are the same. In order to use the assumption, we try converting the coordinate from ECEF to ENU.

[Step 7] The matrix $\mathbf{G}$ is presented as ECEF coordinate system. We convert the matrix $\mathbf{G}$ to ENU coordinate system. The matrix $\mathbf{G}$ means the amount of change in terms of each direction x, y, z. For example, in the satellite #1, the value $\frac{\partial r_1}{\partial x}$ means the change of r_1 in terms of the direction x. The value $\frac{\partial r_1}{\partial y}$ also means the change of r_1 in terms of the direction y. The value $\frac{\partial r_1}{\partial z}$ of course means the change of r_1 in terms of the direction z. Then, we just have to convert each of directions to the east, north and vertical (altitude), respectively. So, we represent the matrix $\mathbf{G}$ in ENU as follows:

$$\mathbf{G}_{\text{enu}} = \begin{pmatrix} g_{11} & g_{12} & g_{13} & 1 \\ g_{21} & g_{22} & g_{23} & 1 \\ g_{31} & g_{32} & g_{33} & 0 \end{pmatrix} \tag{24}$$

By using (1), the each of components means as follows:

$$\begin{pmatrix} g_{k1} \\ g_{k2} \\ g_{k3} \end{pmatrix} = \mathbf{R}(B, L) \begin{pmatrix} \frac{\partial r_k}{\partial x} \\ \frac{\partial r_k}{\partial y} \\ \frac{\partial r_k}{\partial z} \end{pmatrix}, k = 1, 2 \tag{25}$$

$$\begin{pmatrix} g_{31} \\ g_{32} \\ g_{33} \end{pmatrix} = \mathbf{R}(B, L) \begin{pmatrix} \frac{\partial d}{\partial x} \\ \frac{\partial d}{\partial y} \\ \frac{\partial d}{\partial z} \end{pmatrix} \tag{26}$$

The values B, L are the degrees of latitude and longitude at the position (x, y, z).

[Step 8] Because of the matrix $\mathbf{G}_{\text{enu}}$ is ENU coordinate, the vector $\Delta\vec{x}$ have to be ENU coordinate. Then, (19) have to be converted to ENU coordinate system.

$$\Delta\vec{r} = \mathbf{G}_{\text{enu}}\Delta\vec{x}_{\text{enu}} \tag{27}$$

$$= \begin{pmatrix} g_{11} & g_{12} & g_{13} & 1 \\ g_{21} & g_{22} & g_{23} & 1 \\ g_{31} & g_{32} & g_{33} & 0 \end{pmatrix} \begin{pmatrix} \Delta x_{\text{enu}} \\ \Delta y_{\text{enu}} \\ \Delta z_{\text{enu}} \\ \Delta s \end{pmatrix} \tag{28}$$

In this paper, the movement of the altitude is assumed as zero. This means that $\Delta z_{\text{enu}} = 0$. Equation (28) can be represent as the following.

$$\Delta\vec{r} = \begin{pmatrix} g_{11} & g_{12} & 1 \\ g_{21} & g_{22} & 1 \\ g_{31} & g_{32} & 0 \end{pmatrix} \begin{pmatrix} \Delta x_{\text{enu}} \\ \Delta y_{\text{enu}} \\ \Delta s \end{pmatrix}$$

$$\equiv \mathbf{G}_{\text{enu2}}\Delta\vec{x}_{\text{enu2}} \tag{29}$$

The row of the above matrix is three. The unknown parameters is also three. Then, we can solve these parameters.

[Step 9] We transform (29) to the following, and calculate the updating value $\Delta\vec{x}_{\text{enu2}}$.

$$\Delta\vec{x}_{\text{enu2}} = \mathbf{G}_{\text{enu2}}^{-1}\Delta\vec{r} \tag{30}$$

Table I
SIMULATION PARAMETERS (FOR OUR ALGORITHM)

Satellite Position #1 [m]	α_1	-26309844.5749
	β_1	3237477.5201
	γ_1	2627019.5575
Satellite Position #2 [m]	α_2	5096038.9206
	β_2	15688974.7606
	γ_2	21240453.8041
Receiver Position A [m]	x	-3460143.2936
	y	3657442.8374
	z	3616321.2928
Past Receiver Position A' [m]	x'	-3760139.5967
	y'	3657446.1923
	z'	3616321.0131
Pseudorange Error [m]	4σ	1.0/5.0/10
Distance Sensor Error[m]	$\epsilon_{\max}$	1.25

Table II
SIMULATION PARAMETERS (ADDING FOR CONVENTIONAL)

Satellite Position #3 [m]	α_3	-15787358.0601
	β_3	20079908.0255
	γ_3	7249534.1559
Satellite Position #4 [m]	α_4	-14284027.2837
	β_4	-12788411.1086
	γ_4	19047366.2733

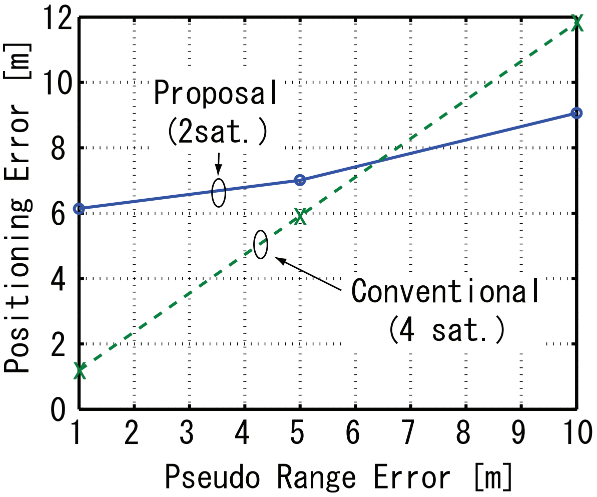


Figure 5. Positioning error (Our algorithm and Conventional)

[Step 10] The calculated updating values $\Delta x_{\text{enu}}, \Delta y_{\text{enu}}$ are converted to ECEF coordinate.

$$\begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} = \mathbf{R}^{-1}(B, L) \begin{pmatrix} \Delta x_{\text{enu}} \\ \Delta y_{\text{enu}} \\ \Delta z_{\text{enu}} \end{pmatrix} \tag{31}$$

, where $\Delta z_{\text{enu}} = 0$.

[Step 11] Based on the updating values in ECEF, we update the estimating parameters as follows.

$$\begin{aligned} x^1 &= x^0 + \Delta x, & y^1 &= y^0 + \Delta y, \\ z^1 &= z^0 + \Delta z, & s^1 &= s^0 + \Delta s \end{aligned} \tag{32}$$

After that, we return back to the step 2 with the updated values x^1, y^1, z^1, s^1 instead of the initial value x^0, y^0, z^0, s^0. We update the estimating unknown parameters x, y, z, s again and again as the residual errors $\Delta r_1, \Delta r_2, \Delta d$ become small adequately. Finally we get the solution of x, y, z, s.

By the above steps, we can estimate own position with only two satellites. Our algorithm uses also the travelling distance. In case of vehicles, we should apply the distance to the positioning effectively because we can get the distance easily. As mentioned before, in urban area, the number of satellites which can observe in LOS fluctuates. Moreover, the place where we cannot observe four and more satellites often occurs. Our algorithm can continue the positioning by selecting just two satellites which have high elevation angle. Our approach may achieve the robust positioning everywhere.

C. Performance evaluation

In this section, we show the examples of our estimation performance. The simulation parameters are summarized in Table I. The positions of the two satellites and the receiver are decided from the real data which was recorded at the forth order triangulation point (reference code: TR45235161301, Housono, Kyoto Japan). Especially, we set the position of the satellites as the decoded position at 17:50:4.801, December 27, 2012. The simulated satellite #1, #2 is set as the real GPS satellites whose PRN is #5, #18, respectively. The distance d is 5 [m].

The pseudorange includes some errors such as troposphere and ionosphere delay error, clock error, multipath error and some noise. The experiment was done under the open sky. So, we can ignore the multipath error. The troposphere and ionosphere delay error can be modeled and the satellites sends the modeled delay in their messages. So, we can subtract the above modeled delay from the pseudoranges. The remained error is sum of model error and clock error and some noise. In this simulation, in order to confirm robust convergence, we assume the remained error as a normal distribution with standard variation σ. By using the positions of both the satellites and receiver in Table I, we calculated the true range. And we prepared the pseudorange r_1, r_2 by adding the above remained error to the true range. In the simulation, the standard variation of the error is set as 3 cases, 1.0m, 5.0m, 10.0m. These values are affected in terms of the elevation angle to the satellites. We also modeled the error of the travelling sensor as uniform distribution whose range is $\epsilon_{\max} = \pm 1.25$[m]. The range is decided as a outer perimeter of a tire of a personal vehicle.

For comparison, we also estimated the receiver's position by using the conventional algorithm with four satellites. We added more two satellites #3, #4 in Table II to the satellites #1, #2 in Table I. The PRN of the added satellites #3, #4 is #24 and #28 respectively.

As performance measure, we evaluate the positioning error which means the Euclidean distance between the true position and the estimated position. The trial times is 100,000. The final positioning error is defined as the mean value of each of trials. The results are summarized in Figure 5. The horizontal axis is the pseudorange error and the vertical axis is the final positioning error.

From Figure 5, our algorithm has large positioning error when the pseudorange error is small. This is because the error of sensor affects the positioning error as a dominant factor. On the other hand, our algorithm becomes better when the pseudorange error is large. This is because the error of sensor is small compared to the pseudorange error, can suppress the positioning error. Our algorithm has some errors, but noteworthy big advantage. We note our algorithm can estimate the position though the conventional cannot estimate with only two satellites.

V. Conclusion

In this paper, we presented the novel positioning algorithm in urban area where propagation environment of multipath or NLOS occurs. Especially, our algorithm can estimate own positions even if the number of LOS satellites is reduced because of the structures such as buildings. The conventional algorithm needs four and more LOS satellites. Our algorithm can estimate with just two LOS satellites by using the sensor data of the travelling distance. In the computer simulations, our algorithm achieved better positioning performance than that of the conventional in case that the pseudorange errors of the satellites were large.

Acknowledgment

A part of this work was supported by Grants-in-Aid for Scientific Research, Japan Society for the Promotion of Science (JSPS).

References

[1] T. Goto and M. Yamaoka, “Personal mobility robot,” Journal of Siciety of Automotive Engineers of Japan, vol. 64, no. 5, May 2002, pp. 75–78.

[2] J. Soubielle, I. Fijalkow, and A. Bibaut, “GPS positioning in a multipath environment,” IEEE Trans. Signal Processing, vol. 50, no. 1, Jan. 2002, pp. 141–150.

[3] E. Costa, “Simulation of the effects of different urban environments on GPS performance using digital elevation models and building databases,” IEEE Transactions on Intelligent Transportation Systems, vol. 12, no. 3, Sep. 2011, pp. 819–829.

[4] T.-S. Dao, K. Leung, C. Clark, and J. Huissoon, “Markov-based lane positioning using intervehicle communication,” IEEE Transactions on Intelligent Transportation Systems, vol. 8, no. 4, Dec. 2007, pp. 641–650.

[5] A. Vu, A. Ramanandan, A. Chen, J. Farrell, and M. Barth, “Real-time computer vision/DGPS-aided inertial navigation system for lane-level vehicle navigation,” IEEE Transactions on Intelligent Transportation Systems, vol. 13, no. 2, Jun. 2012, pp. 899–913.

[6] J. Naranjo, C. Gonzalez, R. Garcia, and T. de Pedro, “ACC+stop go maneuvers with throttle and brake fuzzy control,” IEEE Transactions on Intelligent Transportation Systems, vol. 7, no. 2, Jun. 2006, pp. 213–225.

[7] T. Yamasaki, T. Ishikawa, and K. Aizawa, “Retrieval of images captured by car cameras using its front and side views and GPS data,” IEICE transactions on information and systems, vol. 90, no. 1, Jan. 2007, pp. 217–223.

[8] I. Parra Alonso, D. Fernández Llorca, M. Gavilan, S. Álvarez Pardo, M. Garcia-Garrido, L. Vlacic, and M. Sotelo, “Accurate global localization using visual odometry and digital maps on urban environments,” IEEE Transactions on Intelligent Transportation Systems, vol. 13, no. 4, Dec. 2012, pp. 1535–1545.

[9] N. Alam, A. Tabatabaei Balaei, and A. Dempster, “Relative positioning enhancement in VANETs: A tight integration approach,” IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 1, Mar. 2013, pp. 47–55.

[10] P. Misra and P. Enge, Global Positioning System: Signals, Measurements, and Performance Ganga-Jamuna Press, 2001.

[11] J. Byeong-Chan and S. Kim, “Multipath interference cancellation technique for high precision tracking in GNSS receiver,” IEICE transactions on communications, vol. 93, no. 7, Jul. 2010, pp. 1961–1964.

[12] N. Kubo, S. Kondo, and A. Yasuda, “Evaluation of code multipath mitigation using a software GPS receiver,” IEICE transactions on communications, vol. 88, no. 11, Nov. 2005, pp. 4204–4211.

[13] S. Kim, J. Byeong-Chan, and S. Lee, “DoA estimation of line of sight signal in multipath channel for GNSS receiver,” IEICE transactions on communications, vol. 92, no. 11, Nov. 2009, pp. 3397–3400.

[14] S. J. Hwan, J. Heo, S. Yoon, and K. S. Young, “Interference cancellation and multipath mitigation algorithm for GPS using subspace projection algorithms,” IEICE transactions on fundamentals of electronics, communications and computer sciences, vol. 91, no. 3, Mar. 2008, pp. 905–908.

[15] J. Meguro, T. Murata, J. Takiguchi, Y. Amano, and T. Hashizume, “GPS multipath mitigation for urban area using omnidirectional infrared camera,” IEEE Transactions on Intelligent Transportation Systems, vol. 10, no. 1, Mar. 2009, pp. 22–30.

[16] H. Irie, “Accuracy in changing a number of GPS satellites,” Technical Report of IEICE, vol. 99, no. 248, Jul. 1999, pp. 63–68.

[17] K. Kawamura and T. Tanaka, “Improvement of accuracy in changing the number of GPS satellites(measurement technology),” IEICE transactions on fundamentals of electronics, communications and computer sciences, vol. 89, no. 7, Jul. 2006, pp. 2092–2095.

[18] T. Fan, T. Sato, T. Sakamoto, and X. Mao, “An approach to improving GPS positioning accuracy using reflected signals,” IEICE General Conference, vol. 2010, no. 1, Mar. 2010, p. 266.

[19] H. Onishi, H. Hatano, and Y. Kuwahara, “Novel positioning algorithm using a GNSS satellite and two ground receivers,” International Journal of Automotive Engineering, Society of Automotive Engineers of Japan, vol. 4, no. 2, Apr. 2013, pp. 25–32.

[20] K. Ito, S. Fukushima, N. Arai, and T. Sakai, “Highly-accurate positioning experiment using a quasi-zenith satellite system at ENRI,” Technical Report of IEICE, vol. 104, no. 697, Feb. 2005, pp. 59–63.

[21] S. Hama, Y. Takahashi, J. Amagai, H. Ito, T. Morikawa, S. Yokota, M. Fujieda, and K. Kimura, “Quasi-zenith satellite system (QZSS) : Outline and its time related mission,” Technical Report of IEICE, vol. 105, no. 322, Oct. 2005, pp. 13–17.

[22] F. Wu, N. Kubo, and A. Yasuda, “Performance evaluation of GPS augmentation using quasi-zenith satellite system,” IEEE Transactions on Aerospace and Electronic Systems, vol. 40, no. 4, Oct. 2004, pp. 1249–1260.

Integrated Technologies for Communication Security on Mobile Devices

Alexandre Melo Braga
Fundação CPqD – Centro de Pesquisa e Desenvolvimento em Telecomunicações
Campinas, Brazil
ambraga@cpqd.com.br

***Abstract*—Communication security, information disclosure and vulnerability exploitation are always a concern and a challenge, especially these days, when everything goes mobile. This short paper describes preliminary results on the construction of an integrated framework of applications for secure communication via mobile devices. Particularly, the paper discusses major design decisions on three topics, namely, framework architecture, mobile applications for communication security, and cryptographic service providers.**

***Keywords*—mobile security; commuication security; SMS security; Instant Message security**

I. INTRODUCTION

A recently reported incident on Android [5][9] brought to light, once again, the issue of blindly relying on a single vendor's defenses. Also, a recent disclosure of top-secret NSA (U.S. National Security Agency) documents to The Guardian [19] exposed U.S. government's surveillance activities on phone and Internet communication. These two incidents suggest more than ever that there is a need, on mobile devices, for security solutions suitable to the regular people and that go beyond the ordinary software for antivirus and e-mail security.

On the other hand, NSA recently started to encourage the use of Commercial-Off-The-Shelf (COTS) mobile devices, in particular smartphones running Android, for communication of classified information [14], and fostering a worldwide improvement of mobile security products.

This short paper presents preliminary results of a Brazilian project that fosters security technologies on mobile environments. The main motivation for the project is to offer technological solutions for the growing demand for security in mobile environments. This demand was caused not only by the significant increase in the use of smart mobile devices (smartphones and tablets), but also by the growing interest of cyber criminals in mobile environments. It is important to note the existence of malicious software specific to the Brazilian context, according to the Computer Security Incident Response Team (CSIRT) for the Brazilian Internet [7]. Furthermore, according to the Brazilian Telecommunications Agency, in 2010 the number of mobile accesses has exceeded the number of people in Brazil [2], and a large portion of it is for data.

The remainder of this text is organized according to the results obtained so far. Section II describes the overall architecture of the proposed framework. Section III outlines related work. Section IV details the set of applications for

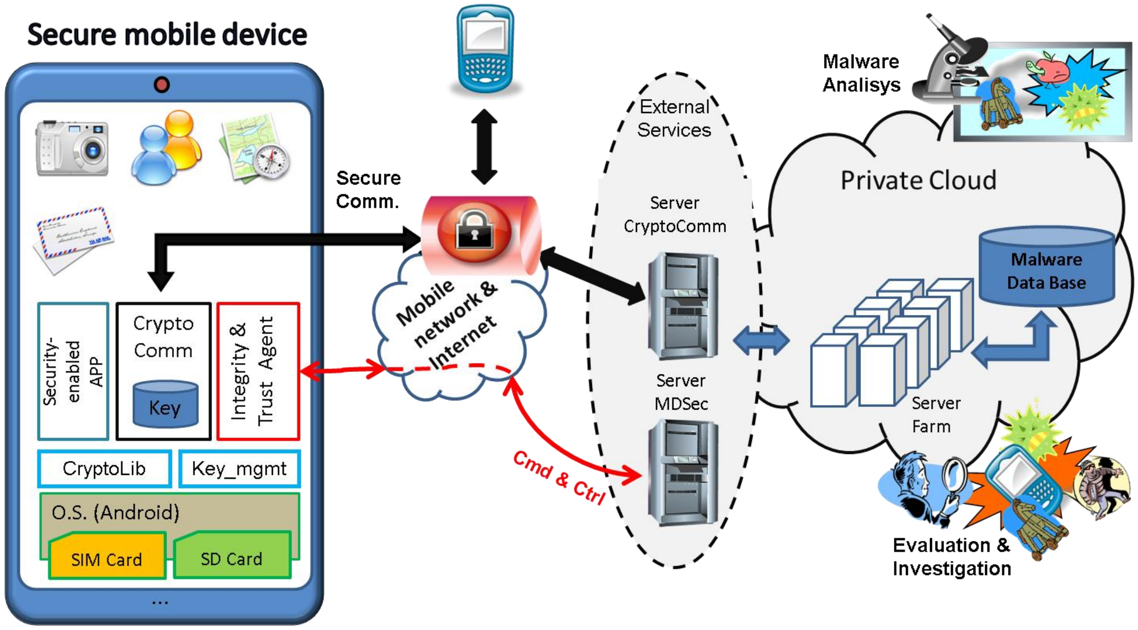


Figure 1. Framework architecture, secure communication, trust management and back office services.

secure communication. Section V explains the main aspects of a cryptographic service provider built for the prototypes. Section VI concludes the paper and points to future directions. For clarification purposes, Table I, at the end, shows a small glossary of cryptographic terms used in this paper.

II. INTEGRATED VIEW AND ARCHITECTURE

There are three main objectives that drive the proposed architecture shown in Figure 1. The first is to build prototypes of secure data communication as well as secure voice over data packets (or over IP), both of them through smartphones on public networks (e.g. 3G, 4G, WiFi). The second is to develop tools for integrity checking on smartphones, as well as techniques for active investigation of security incidents and penetration tests on mobile platforms. Finally, the third objective is to build an environment for experimentation, observation and analysis of mobile malware.

The everyday work is supported by a laboratory for mobile security, which is able to carry out assessments on mobile environments, including platforms, applications and communications, as well as security analysis of mobile malware. The knowledge acquired by the lab team feeds the development team with security controls and counter measures. A private cloud provides services to the development team. Not only security services are provided, but also hosting for servers.

On the client side, Android was chosen as the preferred platform for development of prototypes. The preliminary results described in the next sections addresses three main points of these prototypes: (i) design decisions for secure communication, (ii) secure instant messages and secure SMS, and (iii) a cryptographic library for Android.

III. RELATED WORK

This section outlines recent publications related to the work shown in this paper. Enck et al. [22] show a comprehensive study on the general aspects of Android security. Recently, Braga and Nascimento [1] assessed the feasibility of sophisticated cryptographic services on modern smartphones running Android.

Concerning Short Message Service (SMS) encryption, Pereira et al. [6] show an experimental framework for securing SMS-based communications in mobile phones, which encloses a tailored selection of lightweight cryptographic algorithms and protocols, providing encryption, authentication and signature services. Saxena and Chaudhari [12] researched an approach for securing of SMS which is based upon a variant of ECDSA algorithm. Also, Chavan and Sabnees [16] proposed a technique that combines encryption and compression of SMS messages.

The work of Xuefu and Ming [4] shows the use of eXtensible Messaging and Presence Protocol (XMPP) for Instant Messaging on web and smartphones. Massandy and Munir [8] have done experiments on security aspects of communication, but there are unsolved issues, such as strong authentication, secure storage, and implementation of good cryptography, as shown by Schrittwieser et al. [18].

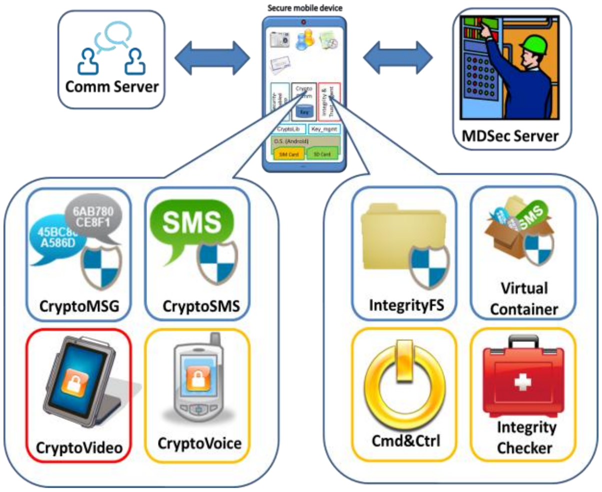


Figure 2. Secure communication and trust management.

Concerning the secure storage of data on mobile devices, a survey by Diesburg and Wang [17] summarizes and compares existing methods of providing confidential storage and deletion of data in personal computing environments, and points that secure deletion of files from flash memory devices is a goal hard to achieve. Wang et al. [23] presented an encrypted file system in user space to protect the removable and persistent storage on smart devices running Android. Reardon et al. [10] addressed the secure deletion of user-space files on Android devices, but with a slow solution.

IV. SECURE COMMUNICATION

Nowadays, secure phone communication does not mean only voice encryption, but encompasses a plethora of security services built over the ordinary smartphone capabilities. To name just a few of them, these are SMS encryption, Instant Message (IM) encryption, voice and video chat encryption, secure conferencing, secure file transfer, secure data storage, secure application containment, and remote security management on the device, including management of cryptographic keys.

Figure 2 illustrates four services of secure communication in scope of this work: CryptoMsg for Instant Messages, CryptoSMS for secure SMS, CryptoVoice and CryptoVideo for secure voice or face-to-face communication, along with a communication server. Figure 2 also shows four secure management services in scope: IntegrityFS for file encryption and integrity, VirtualContainer for application aggregation and containment, Cmd&Ctrl for remote management, and IntegrityChecker for assurance of device's integrity.

By the time of this writing, not all of the services were implemented. In fact, this section describes only two of those cryptographically secure services, namely, encrypted instant messages (CryptoMsg) and encrypted SMS (CryptoSMS).

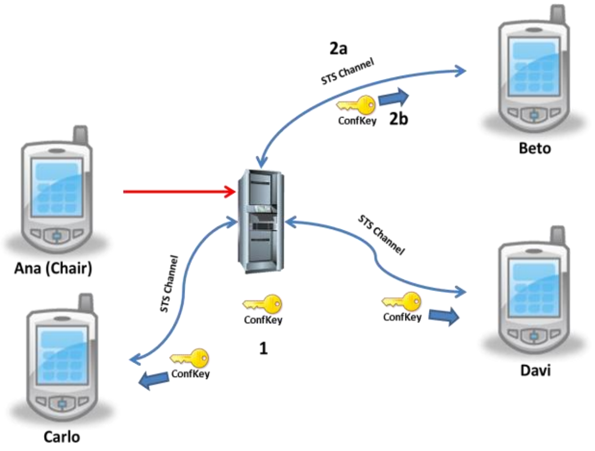


Figure 4. Key agreement for secure conference.

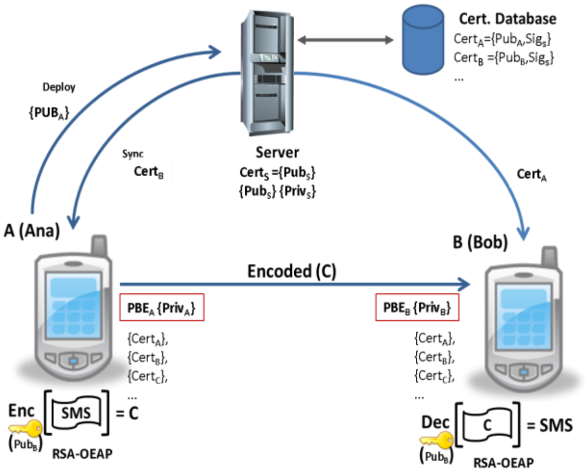


Figure 3. Cryptographically secure SMS.

A. *Cryptographically secure instant message*

The current technology standardized by industry for Instant Messages is the eXtensible Messaging and Presence Protocol (XMPP), the IETF's formalization of base XML streaming protocols for instant messaging and presence, which were originally developed within the Jabber community [15]. The communication architecture supported by XMPP does not allow direct machine-to-machine communication, but requires a server (the XMPP Server) that acts as both a proxy and a mediator among all client applications. This way, the CryptoMsg application is a XMPP client.

An interesting side effect of having chosen XMPP is that CryptoMsg can talk through a proprietary server as well as communicate via Google or Facebook chat servers, as a contingence service, if the primary server is down. Neither Google nor Facebook block encrypted traffic encoded as text, so that two CryptoMsg clients can talk to each other through two Google or Facebook accounts.

CryptoMsg uses a variant of Diffie-Hellman Protocol for key agreement called Station-to-Station (STS). A secret key is negotiated for each chat conversation, and once the key is shared between the two participants, all XMPP messages are encrypted with AES in CBC mode, providing end-to-end encryption. XMPP is actually XML over a TCP communication socket, so TLS can be used instead of regular TCP, for a second layer of point-to-point encryption on the communication channel.

Password-based Encryption (PBE) is the cryptographic technology applied to protect saved conversations. More detail on the cryptographic services provided for this implementation can be found later in this text.

By the time of writing, a secure conference (or group chat) for instant messages was being designed and implemented, as depicted in Figure 4. The Organizer or Chair of the conference requests the conference creation to the Server, and the key agreement for the requested conference proceeds as follows, where $Enc_k(x)$ means encryption of x with key k:

1. Server (S) creates the key for that conference (c_k);
2. For each guest (g[i]), Server (S) does:

   a) Opens a STS channel with key k: S $\leftrightarrow$ g[i], key k;

   b) Sends c_k on time t: S $\rightarrow$ g[i]: $Enc_k(c_k; t; C[i])$;

The steps above constitute a point-to-point key transport using symmetric encryption, which is provided by the STS protocol. After that, all guests share the same conference key and the conference proceeds as a multicast of all encrypted messages.

A variation of this design can use the Chair to both generate and distribute the conference key. This extra computation over the Chair can be acceptable under extraordinary circumstances when the primary server is off-line and the number of guests is small.

B. *Cryptographically secure SMS*

Despite the increasing popularity of mobile IM applications, SMS is still useful among those users without a reliable data access. Also, secure SMS can be used as a secure communication channel for other applications.

The solution described in this paper utilizes asymmetric encryption in order to simplify key distribution among users, who may not have data access at the very moment of sending a message. Figure 3 depicts the proposed solution. First of all, users receive from the server, during application installation, the digital certificates of all her contacts, along with server's self-signed certificate. As can be deduced by the reader, this step requires data access. Contacts synchronization and software update also require data access.

A secure SMS can be encrypted and digitally signed, as well. Two implementation issues have to be considered in this scenario. First, the text actually typed by the user, after be encrypted and signed, can result in multiple SMS messages. Upon receiving a series of SMS messages, the application has to be able to sequencing and marshaling the segmented text from various SMS messages. Second, SMS is text only, so a sort of encoding has to be used before transmitting cipher text. The current version of CryptoSMS was designed to use OAEP for encryption and PSS for digital signatures. Multimedia via SMS is not supported in current version.

V. CRYPTOGRAPHIC SERVICE PROVIDER

A cryptographic library for Android was built to provide cryptographic services to be used in the protection of secure communication via mobile devices. In order to be useful, the cryptographic library had to accomplish a minimum set of functional requirements. Each functional requirement generates a set of non-functional or supplementary requirements, mostly related to correctness of algorithms, compliance to industry standards, security, and performance of the implementation.

The design of the current version of the cryptographic library is illustrated in Figure 5 and contains the cryptographic algorithms and protocols described in the following paragraphs. This implementation complies with standard cryptographic API of the Java Cryptographic Architecture (JCA), its name conventions, and design principles [11]. A small glossary of acronyms is in Table I.

In order to provide a fully functional Cryptographic Service Provider (CSP) for secure communication, a minimum set of algorithms had to be chosen. This minimalist construction follows, but is not certified by publicly available standards [13] and provides the following set of services:

a) A symmetric algorithm to be used as block cipher, along with the corresponding key generation function, and modes of operation and padding. The AES algorithm was chosen, along with thee modes of operation: ECB, CBC, and the GCM mode for authenticated encryption. The padding technique for block ciphers is the one defined by PKCS#5.

b) An asymmetric algorithm for digital signatures, along with the key-pair generation function. This requirement brings with it the need for some sort of digital certification of public keys. The asymmetric algorithms are RSA-PSS for signatures and RSA-OAEP for asymmetric encryption. Both of them are probabilistic schemes constructed over ordinary RSA, and are supposed to be more secure than RSA. By the time of writing, RSA-OAEP was not fully implemented;

c) A one-way secure hash function, SHA-1, which is an underling hash function in MACs, digital signatures and Pseudo-Random Number Generators (PRNG). SHA1PRNG was chosen to be used by all the key generation functions;

d) Message Authentication Codes (MAC). HMAC with SHA-1 as the underling hash function, and GMAC, directly derived from GCM mode, where chosen;

e) A key agreement mechanism to be used by communicating parties that have never met before, but need to share an authentic secret and communicate securely. The need for key agreement was fulfilled by the implementation of Station-to-Station (STS) protocol [21], which is based on Authenticated Diffie-Hellman (ADH) [20], and provides mutual key authentication and key confirmation;

f) A simple way to keep keys safe at rest and that does not depend on hardware features. The mechanism for Password-based Encryption (PBE) [3] is based on the Password-Based Key Derivation Function 2 (PBKDF2) [3], and provides a simple and secure way to store keys in encrypted form. In PBE, a key-encryption-key is derived from a password.

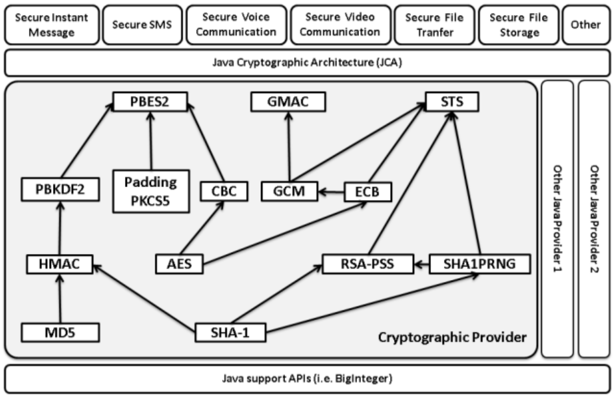


Figure 5. Cryptographic service provider.

A. *How those things are tested, anyway?*

When developing a security-aware application, the first thing to ask is how it will be tested for security. Furthermore, cryptographic software usually requires correctness of basic functions, as well as conformance to specifications and standards. Usually, cryptologists require assistance in writing fast and secure code, because doing it from scratch is almost impossible. Also, canonical implementations, based on standard algorithms, always need optimization and other code transformation in order to be useful in real applications. Code transformation can lead to vulnerabilities, requiring fixes, shipped as software patches, in a never ending cycle.

During this implementation, besides regular functional tests and automated unit test, test vectors were used as automated acceptance tests for cryptographic software. Test vectors are usually available for standardized algorithms [13] and meet halfway between customer and developer, because they come from the problem domain (cryptography) and don't look like source code. They have the benefit to be clear to customer (and to cryptologists) and can be used to reach agreement for when the work is finished. Also, they are not completely freeform and can be used to create automated tests. This approach increases customer’s trust in cryptographic software during algorithm implementation, as well as provides good regression tests as an evidence of correctness after many code transformations.

VI. CONCLUSION AND FUTURE WORK

This short paper presented design decisions taken during the construction of a framework for cryptographically secure communication on Android devices. The early prototypes still have many challenges to be overtaken, as follows. The use of cryptosystems for short signatures, elliptic curve cryptography and pairings-based cryptography are some of the improvements planned for the near future. Another challenge is to preserve usability in the presence of strong security concerns. Layered protections against common software vulnerabilities, such as secure deletion of sensitive information in flash memory, are planned in the roadmap.

Finally, automatic testability of cryptography has to be improved not only for algorithms, but also for protocols.

ACKNOWLEDGMENT

The author acknowledge the financial support given to this work, under the project "Security Technologies for Mobile Environments – TSAM", granted by the Fund for Technological Development of Telecommunications – FUNTTEL – of the Brazilian Ministry of Communications, through Agreement Nr. 01.11. 0028.00 with the Financier of Studies and Projects - FINEP / MCTI.

REFERENCES

[1] A. M. Braga and E. N. Nascimento, “Portability Evaluation of Cryptographic Libraries on Android Smartphones”, Proc. 4th International Conference on Cyberspace Safety and Security (CSS), Dec. 2012, pp. 459-469.

[2] ANATEL – Agência Nacional de Telecomunicações, www.anatel.gov.br [retrieved: Oct., 2013].

[3] B. Kaliski, RFC 2898, “PKCS #5: Password-Based Cryptography Specification”, Version 2.0, tools.ietf.org/html/rfc2898 [retrieved: Oct., 2013].

[4] B. Xuefu and Y. Ming, “Design and Implementation of Web Instant Message System Based on XMPP”, Proc. 3rd International Conference on Software Engineering and Service Science (ICSESS), Jun. 2012, pp. 83-88.

[5] BBC News, “Master Key to Android Phones Uncovered”. www.bbc.co.uk/news/technology-23179522. [retrieved: Oct., 2013].

[6] C. Pereira et al., “SMSCrypto: A Lightweight Cryptographic Framework for Secure SMS Transmission”, J. Syst. Softw, vol. 86, no. 3, Mar. 2013, pp. 698-706.

[7] CERT.br – Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil. www.cert.br [retrieved: Oct., 2013].

[8] D. T. Massandy and I. R. Munir, “Secured Video Streaming Development on Smartphones with Android Platform”, Proc. 7th International Conference on Telecommunication Systems, Services, and Applications (TSSA), Oct. 2012, pp. 339-344.

[9] J. Forristal, “Uncovering Android Master Key that Makes 99% of Devices Vulnerable”, bluebox.com/corporate-blog/bluebox-uncovers-android-master-key [retrieved: Oct., 2013].

[10] J. Reardon, C. Marforio, S. Capkun, and D. Basin, “User-level Secure Deletion on Log-structured File Systems”, Proc. 7th ACM Symposium on Information, Computer and Communications Security (ASIACCS), May 2012, pp. 63-64.

[11] JCA Providers Documentation for Java Platform Standard Edition 7, docs.oracle.com/javase/7/docs/technotes/guides/security/SunProviders [retrieved: Oct., 2013]

[12] N. Saxena, N. S. Chaudhari, “Secure Encryption with Digital Signature Approach for Short Message Service”, Proc. World Congress on Information and Communication Technologies (WICT), Nov. 2012, pp. 803-806.

[13] NIST CAVP, Cryptographic Algorithm Validation Program, csrc.nist.gov/groups/STM/cavp/index.html [retrieved: Oct., 2013].

[14] NSA, Mobility Capability Package - Secure VoIP, v. 2.1, www.nsa.gov/ia/_files/Mobility_Capability_Pkg_Vers_2_1.pdf [retrieved: Oct., 2013].

[15] P. Saint-Andre, K. Smith, and R. Tronçon, “XMPP: The Definitive Guide - Building Real-Time Applications with Jabber Technologies”, O’reilly, 2009.

[16] R. Chavan and M. Sabnees, “Secured Mobile Messaging”, Proc. International Conference on Computing, Electronics and Electrical Technologies (ICCEET) , Mar. 2012, pp.1036-1043.

[17] S. Diesburg and A. Wang. “A Survey of Confidential Data Storage and Deletion Methods”. ACM Comp. Surveys, vol. 43, no. 1, Nov. 2010.

[18] S. Schrittwieser et al., “Guess Who's Texting You? Evaluating the Security of Smartphone Messaging Applications”. Proc. 19th Network & Distributed System Security Symposium, Feb. 2012.

[19] The Guardian, Hot site on “Edward Snowden’s disclosure of NSA top-secret information”, www.guardian.co.uk/world/edward-snowden [retrieved: Oct., 2013].

[20] W. Diffie and M. Hellman, "New Directions in Cryptography," IEEE Transact. on Inform. Theory, vol. 22, no. 6, Nov. 1976, pp. 644-654.

[21] W. Diffie, P. C. van Oorschot, and M. J Wiener, “Authentication and Authenticated Key Exchanges”, Designs, Codes and Cryptography, vol. 2, no. 2, 1992, pp. 107–125.

[22] W. Enck, D. Octeau, P. McDaniel, and S. Chaudhuri, “A Study of Android Application Security”, Proc. 20th USENIX conference on Security (SEC), 2011, pp. 21-21.

[23] Z. Wang, R. Murmuria, and A. Stavrou, “Implementing and Optimizing an Encryption Filesystem on Android”, Proc. 13th International Conf. on Mobile Data Management, 2012, pp. 52-62.

TABLE I. OVERVIEW OF CRYPTOGRAPHIC ACRONYMS

Acronym	Brief Description of the Acronym
ADH	Authenticated Diffie-Hellman is a way of exchanging cryptographic keys among mutually autenticated parties.
AES	Advanced Encryption Standard is a specification for data encryption established by the U.S. government.
CBC	Cipher Block Channing is an operation mode for block ciphers, in symmetric-key cryptography.
CSP	Cryptographic Service Provider is a collection of cryptographic implementations.
ECB	Electronic Code Book is an operation mode for block ciphers, in symmetric-key cryptography.
GCM	Galois/Counter Mode is an operation mode for block ciphers, in symmetric-key authenticated encryption.
GMAC	An authentication-only variant of GCM.
HMAC	keyed-Hash MAC is a function for calculating a MAC involving a secure hash function and a secret key.
MAC	Message Authentication Code is a small piece of data that provides integrity and authenticity for a message.
OAEP	Optimal Asymmetric Encryption Padding is a padding scheme often used together with RSA encryption.
PBE	Password-Based Encryption is a method of deriving a cryptographic key from a password.
PBKDF2	Password-Based Key Derivation Function 2 is a specific technique of implementing a PBE.
PKCS#5	Public-Key Cryptography Standards 5 is a specification devoted to Password-based Encryption.
PRNG	Pseudo-Random Number Generator is an algorithm for generating sequences of numbers that approximate the properties of truly random numbers (a. g. SHA1PRNG).
PSS	Probabilistic Signature Scheme is a secure way of creating digital signatures with RSA.
RSA	RSA is an algorithm for public-key cryptography that is based on the difficulty of factoring large integers.
SHA-1	Standard Hash Algorithm 1 is a cryptographically secure hash function.
STS	Station-to-Station protocol is a key agreement scheme based on ADH that provides mutual authentication.
TLS	Transport Layer Security is a cryptographic protocol for communication security over the Internet.

Enabling Trajectory Constraints for Usage Control Policies With Backtracking Particle Filters

Philipp Marcus, Moritz Kessel and Claudia Linnhoff-Popien
Institute of Computer Science
Ludwig Maximilian University of Munich
{philipp.marcus, moritz.kessel, linnhoff}@ifi.lmu.de

***Abstract*—A number of studies extended access control policies with constraints, aiming at the restriction of mobile users' access to appropriate authorized areas. Recent research proposed to rely on usage control instead, in order to allow for continuous checks of the user's location. A drawback of those approaches is that they rely on crisp trajectory estimates, i.e., spatio-temporal paths, not considering occurring uncertainty. This makes those approaches impractical for indoor applications, where occurring measurement errors are typically large compared to authorized areas. Thus, in this study, we propose extensions for usage control policies to constrain users to an authorized area for the duration of access. We adhere probabilistic trajectories derived from backtracking particle filters combined with WiFi fingerprinting. However, the main contribution is a risk-based model for deriving usage decisions based on risk factors instead of conventional thresholds. Our results show, that particle filters are crucial due to inaccuracy in WiFi positioning. We achieve a true- and false-positive rate of 80% and 6.7%. Finally, this allows to effectively constrain access to appropriate areas in indoor scenarios.**

Keywords*—*Mobile Usage Control; Indoor Positioning; Backtracking Particle Filter; Location-based Access Control

I. Introduction

The significantly increasing popularity of mobile devices offers mobile access to resources from everywhere. However, this arises the inherent risk of access requests to critical resources from inappropriate areas, e.g., from outside a company's site or neighboring offices of foreign companies. To solve this problem, much study in recent years focused on location-based extensions of existing access control models, i.e., role-based access control (RBAC), mandatory access control (MAC) or discretionary access control (DAC) [1]. These extensions allow to refine access rights of mobile users with location predicates. This way, the location of users, accessed resources or both, can be constrained to certain areas or to predefined mutual relations. A drawback of those approaches is that after an access request was granted, the according rights won't be revoked when users move on to possibly inappropriate locations. As a remedy, the change to usage control mechanisms was recently discussed, which focus on the concept of controlling the usage of a resource continuously based on iterative checks [2], [3]. Location predicates applied to this model focus on constraining trajectories, i.e, the covered path of a user in the spatio-temporal space. Typically, trajectories are defined as ploy-lines and created using interpolation on crisp location measurements, for example measured with GPS. Trajectory constraints are used to constrain usage rights to users with trajectories that satisfy predefined boundary conditions. One example is to restrict the path to be contained within a single authorized area (AA), e.g., an office or room. This kind of trajectory constraint is called a containment constraint (CC) for the rest of this paper. The mentioned existing approaches for constraints on trajectories do not account for measurement uncertainty when assuming a crisp poly-line as the user's trajectory. Independently, in the research area of indoor positioning, and tracking in particular, WiFi fingerprinting in combination with backtracking particle filters (BPF), a special Bayesian filter, were shown to yield very promising results for estimating user trajectories [4]. Their performance stems from reducing the negative impact of single location measurement outliers. Additionally, BPFs allow for a probabilistic representation of trajectory estimates over probability density functions (PDF) that are sampled by a set of particles. Every particle represents a hypothesis of the user's past trajectory.

So far, techniques for coping with the probabilistic representation of trajectories in constraints for usage control policies have not been studied. This drawback even makes existing approaches impractical in indoor scenarios, where typically WiFi fingerprinting is used for positioning: Here, when creating a crisp trajectory like required by existing approaches, simply stringing together single location measurements is not sufficient, as occurring errors can easily cause the estimated trajectory to indicate a leaving of the AA erroneously. This causes unacceptable high false decision rates and impractical high risk. Furthermore, existing approaches even prevent trajectory-based usage control to benefit from the promising accuracy achieved with existing BPF and their probabilistic trajectories. Until now, in indoor environments, there exists no means to reliably constrain mobile usage of resources to predefined AAs. This might even make it impossible to obey according legal security and safety constraints in companies.

In order to facilitate CCs on probabilistic trajectories returned by BPFs, our contribution is threefold: At first, Section II presents an architecture for the continuous evaluation of CCs assigned to usage control policies. Here, also the underlying attacker model is defined. Subsequently, Section III first gives a theoretical overview on BPFs and describes how trajectory estimates are derived in our approach. Based on these results, Section IV gives a formal definition of CCs based on the probabilistic representation of trajectories in our BPF. In order to minimize CCs computational overhead, an incremental adaptation is proposed. The main contribution of this section is our risk-based approach for deriving usage decisions, which picks the decision with the lowest risk. Here, risk factors are derived based on time dependent cost functions of false-

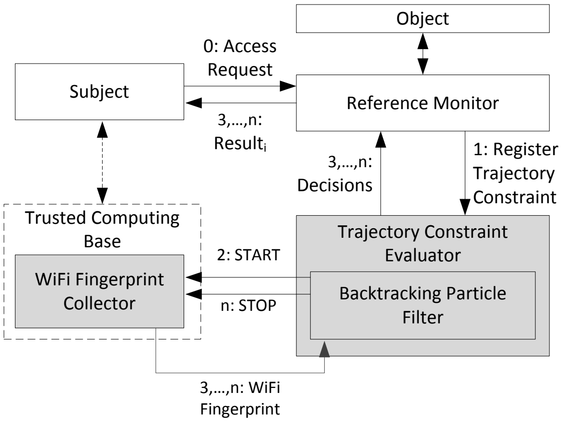


Fig. 1. The new architecture components (gray) continuously collaborate to derive and report usage decisions for the registered trajectory constraints.

negative and false-positive decisions and estimates of a CC's confidence. To the best of our knowledge, only approaches with predefined and static thresholds have been defined up to now [5], [6]. The proposed BPF and evaluation strategies for CCs are evaluated in Section V with 60 trajectories and a database consisting of 206 fingerprints covering $1400\ m^2$. Finally, Section VI concludes the paper.

II. Architecture and Attacker Model

Typical usage control policies incorporate attribute-based access control policies for continuous usage, e.g., UCON ABC [2]. Here, a set of usage rules states which subject is allowed to use a specific object. Decisions about revoking ongoing usage are continuously repeated. However, to the best of our knowledge, there exists only one approach up to now, that explicitly describes the extension of usage control policies with trajectory constraints [3]. Instead of probabilistic representations, only a definition on crisp trajectories is given without a possible underlying architecture. However, we define an architecture that includes components for deriving trajectory estimates and evaluating CCs: Its basic component is the *trajectory constraint evaluator* (TCE), which is able to provide trajectory-aware usage control policies and their reference monitors with Boolean decisions about the satisfaction of applied CCs. Figure 1 depicts the introduced components (gray) along with the necessary message flow. Each time an usage request arrives at the policy's reference monitor, the associated CCs for the responsible usage control rule are looked up and registered at the TCE. In the following steps, the TCE informs the mobile user's *WiFi fingerprint collector* (WFC) to start continuously providing WiFi fingerprints. The WFC is executed directly on the accessing user's mobile device and needs to be under control of a trusted computing base in order to allow for unmanipulated measurements. From now on, the WFC collects fingerprints of the received signal strength (RSS) of surrounding WiFi access points (APs) and sends the measured values along with the time-stamp of their recording digitally signed to the TCE. This needs to be continued until the mobile usage is revoked by the reference monitor or quit by the user. In order to prevent users from appointing WFCs on a foreign mobile device as their own, making trajectory evaluation useless, usage requests need to be constructed by the trusted computing base and additionally have to transmit the service access point of the user's own WFC to the reference monitor. For the rest of this paper, we concentrate on estimating users' trajectories within the TCE and the evaluation of the described CC based on these estimates.

```
function BAYESIAN_FILTER( $bel(x_{t-1})$, $z_t$ )
    for all $x_t$ do
        $\overline{bel}(x_{t-1}) = \int p(x_t|x_{t-1})\ bel(x_{t-1})\ dx_{t-1}$
        $bel(x_t) = \eta^{-1}\ p(z_t|x_t)\ \overline{bel}(x_{t-1})$
    end for
    return $bel(x_t)$
end function
```

Fig. 2. General concept of Bayesian Filters [7].

The presented architecture is conform to our attacker model: An attacker is defined as a mobile user that manipulates estimates about his trajectory in order to obtain usage rights for a given resource by the reference monitor. In our case, we assume attackers that are able to *1)* manipulate sensor data of their mobile device, *2)* delay the provisioning of WiFi measurements to the TCE and, *3)* move freely within and around the AAs referenced in any CC. In contrary, he is not able to *1)* manipulate clock data or the received signal strength (RSS) in WiFi measurements, *2)* replay old WiFi measurements and, *3)* manipulate the WiFi infrastructure.

III. Estimating Trajectories with Backtracking Particle Filters

In this section, our BPF specialized for the evaluation of CCs and its theoretical foundations are presented.

A. Basic Concepts of Backtracking Particle Filters

Particle filters (PF) represent a recursive, non-parametric Bayesian filter with a discrete representation of the posterior by a set of particles of size m. In general, Bayesian filters like the particle filter allow to recursively calculate a *belief* $bel(x_t) = p(x_t|z_{1:t})$ of the system's state for a time-stamp t from already observed measurements $z_{1:t} = \langle z_1, z_2, \ldots, z_t \rangle$ [7]. Internally, the algorithm consists of two parts: The *prediction* step first computes a prior $\overline{bel}(x_t) = p(x_t|z_{1:t-1})$, called the *prediction*, before incorporating the latest measurement. This computes as the integral sum of the *prediction model*, describing the probability of getting to state x_t from a state x_{t-1} and the previous posterior $bel(x_{t-1})$. In the *update* step, the current belief is computed from the *prediction* by incorporating the measurement probability $p(z_t|x_t)$ along with a normalizing constant $\eta = p(z_t|z_{1:t-1})$. The algorithm is depicted in Figure 2. For the field of localization and tracking, PFs implement this algorithm by sampling the posterior probability density function $bel(x_t)$ with a set of m particles $\mathcal{X}_t = \left\langle x_t^{[1]}, x_t^{[2]}, \ldots, x_t^{[m]} \right\rangle$, each one representing a concrete instantiation of the state with information about the position, the velocity and the orientation. The *prediction step* is implemented by applying a mobility model to each particle that predicts its new position, orientation, and velocity. This leads to a updated particle set $\overline{\mathcal{X}}_t$ that approximates $\overline{bel}$. The *update step* is implemented by importance re-sampling: Each particle $x_t^{[i]}$ is assigned a weight $w_t^{[i]}$, called its importance factor, according to the measurement z_t. In order to get a updated set of particles $\mathcal{X}_t$ that is approximately distributed to $bel(x_t)$, m particles are drawn with replacement from the set $\overline{\mathcal{X}}_t$. Here, the probability of drawing a particle is proportional to its weight.

B. Deriving Measurements and Measurement Probabilities

In our system, single location measurements z_t are computed by WiFi fingerprinting: When the user's mobile device measures a RSS for a set of APs, the most likely position is determined from a previously recorded fingerprint database. It computes as the weighted center of mass of fingerprints selected by the k-nearest-neighbors algorithm [8]. Given a location measurement z_t, we derive the measurement probability $p(z_t|x_t^{[i]})$ from a bi-variate Gaussian pdf that describes the error distribution, based on our previous work [8]. When recording the fingerprint database, not only the area of modeled AAs should be covered. Otherwise, as the range of APs must be assumed to be larger than the modeled AAs, a potential attacker could simply stand outside the AA and the positioning algorithm will have no choice but choosing fingerprints as k-nearest-neighbors that were recorded within the AA. This allows an attacker to obtain a position fix that indicates a position within the AA, possibly leading the usage control policy to a false-positive. To solve this problem, we propose to determine the set of those APs that are receivable within the authorized region in at least one point. Next, the union of the coverage areas of this set of APs needs to be covered when recording the fingerprint database. In general, this increases the effort of recording the fingerprint database. For unbiased positioning, the spatial density of the fingerprint database should be uniform throughout the covered site.

In order to annul any of an attackers sensor manipulations, the underlying WiFi fingerprinting and PF algorithm must not employ sensor data. In particular, compass data has been shown to have a positive effect on positioning accuracy to reduce influence of blocking effects of the human body on RSS values [8]. Hence, it is not possible to apply the corresponding technique of recording each fingerprint once for each cardinal direction and choose only those as kNN that have been recorded with the user's current orientation. Therefore, in order to prevent the possibility of an intentional attack, our system needs to accept a possibly lower positioning accuracy.

C. The Application of Backtracking for Refining Estimates

During the *update step* via re-sampling, typically some particles from $\overline{\mathcal{X}}_t$ are not contained in $\mathcal{X}_t$ due to their low weight and are said to *die* whereas others might be drawn a multiple times. In such cases, with each particle, also its assigned hypothesis about a possible user trajectory dies. This allows to refine the knowledge about possible user trajectories up to time t by discarding trajectories associated to dead particles via *backtracking*, which represents a BPF [4]. In BPFs, the knowledge about the past is only refined by future *update steps* as single particles with their assigned estimated trajectory can only be discarded and not newly created. In traditional tracking or positioning systems, for each point in time a single position is computed as the mean value of all particles, leading to one estimated trajectory. This way, information about single existing hypotheses is lost and the single trajectory computed from the means might completely satisfy a CC though none of the estimated trajectories does so. Therefore, trajectory constraints should consult the whole set of trajectories in order to exploit all available information to finally derive confidence values.

D. The Prediction Step: Deriving Trajectory Estimates

To allow for a detailed representation of possible trajectories, we realize the *prediction step* by partitioning the timespan between two *update steps* in sub-intervals of maximally 800 ms, which corresponds to the typical time human users need to take a step. After each passed sub-interval, each particle's trajectory is expanded with a segment. The segments represent a possible movement of the user in this sub-interval. Here, a single segment is constructed as a line by assuming a linear movement from its last position $\mathbf{l}$ with a velocity v in a direction α for the duration of the current sub-interval. Let *Loc* be an arbitrary polygon representing the AA of a CC, we write $\tau(x_t^{[i]})$ *within Loc* iff the segments assigned with $\tau(x_t^{[i]})$ are completely contained within *Loc*. Let $t = 0$ denote the point of time when the usage right was granted and the BPF initialized. Appending the trajectory estimates from all past state estimates a particle $x_t^{[i]}$ was generated from, an estimate for the user's trajectory since $t = 0$ is derived and denoted as $\tau(x_0^{[i]}) \circ \tau(x_1^{[i]}) \circ \ldots \circ \tau(x_t^{[i]})$. This will serve as a key deciding factor for our trajectory-based usage control mechanism.

When constructing a particle's $\tau(x_t^{[i]})$, single segments should be created by a mobility model that is appropriate for the application to usage control in terms of the security implications of our attacker model: Similar to plain WiFi fingerprinting, BPF algorithms were shown to achieve higher accuracy by the application of inertial sensor data, too, and especially benefit from step detection algorithms [9]. This data can be used to perform the *prediction step* based on dead reckoning with the measured sensor values. However, all existing approaches assume benevolent users that do not fake steps by shaking a smart-phone or pretend other movements. If directly applied to our scenario, an attacker could fool the system to assume trajectories that do not leave a prescribed AA though the attacker has left it. Again, it is crucial to accept possibly lower positioning accuracy in order to hamper attacks. Therefore, particles are modeled to follow a random waypoint mobility model, which appends segments to $\tau(x_t^{[i]})$ of each particle $x_t^{[i]}$. One possibility is the model presented by Widyawan et al. [4]. This allows to model linear movement according to the boundary conditions of humans, based on the angle and velocity of the preceding segment. However, basically, any mobility model that is independent of a mobile device's inertial sensor data and based on map matching can be applied. This way, each particle is assigned a hypothesis of the trajectory the user could have walked since the *update step* at time t until the subsequent *update step* is performed.

In order to obtain realistic estimates for trajectories, we also apply the technique of map matching [4], [9]. As for each particle the plausible choices for its next segment are limited by the characteristics of the underlying building plan, we require that particles must not cross walls. During the construction of $\tau(x_t^{[i]})$, this is realized by retrying to infer a valid succeeding segment that is a realistic extension of the trajectory and does not cross walls, until a predefined threshold of maximum tries is exceeded. In such cases, the weight $w_t^{[i]}$

of the particle is 0, despite the probability that might arise from the latest measurement z_t. Hence, the particles' weights are influenced by the *prediction step* and are set to 0 if only wall-crossing segments could be derived within a maximum amount of retries:

$$w_t^{[i]} = \begin{cases} 0, & \text{no valid } \tau(x_t^{[i]}) \text{ found} \\ p(z_t | x_t^{[i]}), & \text{otherwise} \end{cases} \tag{1}$$

Trajectories constructed this way are robust against sensor manipulations of attackers and can finally be supplied for the evaluation of a CC in order to derive usage decisions.

IV. Containment Constraints for Backtracking Particle Filters

The trajectories computed by the adapted BPF are applied to evaluating CCs. In this section, we first discuss the differences to existing approaches and define the concept of CCs formally. Based on this result, an incremental algorithm for their evaluation is presented. Finally, we present a risk-based model for deriving usage decisions based on the current confidence of an evaluated CC.

A. Applicability of Traditional Trajectory Constraints

When using classical location providers like GPS for sampling a user's trajectory, a sample consists of a crisp location and a time-stamp. In the field of moving object databases (MOD), this has been used to define beads, i.e., ellipsoid geometries that contain all points that could be visited during the collection of two samples under the assumption of a maximum velocity [10]. Trajectories are hence affected with uncertainty and described as a sequence of beads. The real trajectory is completely contained within the given beads. With each new sample arriving at a MOD server, the current sequence of beads is extended by a new element. This allows a clear distinction of past and future trajectories. In such cases, trajectory-based usage control can be realized using traditional spatio-temporal queries, assessing to what degree the given beads satisfy the containment within a room. However, as mentioned above, BPFs showed much higher accuracy in indoor scenarios and consequently our trajectory estimates are derived using this method. Those don't form beads and thus classical spatio-temporal queries can't be used directly as an implementation of CCs. Furthermore, re-sampling prevents the clear distinction between past and future trajectories. Even the known representation of beads is not possible without an additional density estimation based on particles' trajectories. These differences of trajectory estimates of BPFs compared to sequentially constructed ones in traditional MODs need to be respected here. Hence, in the following paragraph, we present the formal definition of CCs for usage control policies.

B. Containment Constraints for Backtracking Particle Filters

A *containment constraint cc* is defined as a function mapping from an authorized area *Loc* and the current set of particles $\mathcal{X}_t$ to a confidence value denoting the percentage of trajectory estimates that are completely contained within *Loc* since the usage right was granted at $t = 0$:

Definition 1 (Containment Constraint)
The containment constraint (cc) on a set of particles and an authorized area is defined as:

$$cc\,(\mathcal{X}_t, Loc) = |\mathcal{X}_t|^{-1} \cdot \Big| \Big\{ x_t^{[i]} \in \mathcal{X}_t \,|\, \forall k \in \{0, \ldots, t\} : \tau(x_k^{[i]}) \; within \; Loc \Big\} \Big| \tag{2}$$

with $\mathcal{X}_t$ being the current set of particles since the last update step at time t and Loc being a polygon in $\mathbb{R}^2$ representing the authorized area.

Basically, this constraint needs to be evaluated after every expansion of particles' trajectories during the *prediction* and after each *update step*, as both can influence the percentage of trajectories that satisfy the constraint. Note, that Definition 1 is an adaptation of the well studied *possibly always* spatial query [10]. However, when evaluating CCs, the most computationally demanding step is to check $\tau(x_k^{[i]})$ *within Loc* for each particle $x_k^{[i]}$, for each time-span in history. Employing the discussed properties of our BPF, we define a more efficient, incremental evaluation of CCs: By assigning and incrementally updating a Boolean *is_valid* to each particle in $\mathcal{X}$ it is possible to highly reduce the required number of these checks. This Boolean describes if the trajectory assigned to a particle $x_t^{[i]}$ satisfies or violates the CC that is currently under evaluation. As each single expansion might cause a trajectory to violate the *cc*, in each of its re-evaluations also the property *is_valid* needs to be updated. As single trajectories can't recover from a violation already detected in prior evaluations, the *is_valid* property only needs to be updated for particles with *is_valid* = *True*. The confidence can then easily be computed by counting the percentage of particles that still satisfy the constraint, as depicted in Figure 3. Obviously, this implements Definition 1, as $x_0^{[i]}.is\_valid = \tau(x_0^{[i]})$ *within Loc* and,

```
function INCREMENTAL_CC( $\mathcal{X}_t$, $Loc$ )
    for all $x_t \in \mathcal{X}_t$ do
        if $x_t.is\_valid \wedge \neg\, \tau(x_t)$ within Loc then
            $x_t.is\_valid \leftarrow False$
        end if
    end for
    return $|\mathcal{X}_t|^{-1} \cdot |\{x_t \in \mathcal{X} | x_t.is\_valid\}|$
end function
```

Fig. 3. The algorithm for incrementally computing confidence values.

$$x_{t+1}^{[i]}.is\_valid \; = x_t^{[i]}.is\_valid \wedge \tau(x_{t+1}^{[i]}) \; within \; Loc \tag{3}$$

Thus, in contrast to Definition 1, *incremental_cc* only needs to evaluate $\tau(x_{t+1}^{[i]})$ *within Loc* for the current τ and only for particles with $x_t^{[i]}.is\_valid = True$, i.e., those that did not already violate the constraint in any prior evaluation.

C. Deriving Risk-Based Usage Control Decisions

Based on the confidence values returned by the policy's underlying CC, usage control decisions need to be derived after each *update step* and continuously during each *prediction step*. In detail, the two usage decisions comprise either to revoke the usage right or keep on granting it. We choose that decision with the lowest risk if it was wrong w.r.t. the ground truth. Here, we model risk as the product of probability and costs. Each of these decisions brings costs [11]. The costs for a correct decision, i.e., a true-positive or a true-negative are assumed to be 0. The modeling of costs for false decisions needs to respect that an attacker might retain new WiFi measurements in order to elongate the *prediction step* by an arbitrary time to possibly prevent a revocation of his usage right. Hence, costs of a false-positive are assumed to increase with ongoing time of a *prediction step*, as the occurring damage might increase, i.e., by dumping protected data. Its costs are modeled by a monotonically increasing function $c_{fp}(t) = f(t)$. This also allows to express the intuitive notion of revoking the usage right after a predefined time-off. However, the costs of a false-negative are modeled as $c_{fn}(t) = const$ as we assume the costs that occur from refused usage to be constant with time.

To derive a usage decision directly after the *update step*, first the confidence (c) of the CC is computed. As the *prediction step* has not yet started, here $t = 0$ and the usage right is revoked if the risk of a false-positive exceeds the risk of a false-negative, i.e., $c_{fp}(0) \cdot (1 - c) \geq c_{fn}(0) \cdot c$. However, to derive usage decisions during the *prediction step*, we estimate the maximum risk for both decisions for any point of time t during the *prediction step*: The maximum risk of keep granting is directly influenced by the lowest possible confidence about the CC at time t. Note, that the observed confidence of the randomly moved particles typically will be higher than this lower bound. Contrary, the maximum risk of revoking the usage at time t is proportional to the highest probability that the user could still be within the AA and thus to its initial confidence. In line with existing risk-based approaches to access or usage control [11], [12], again we revoke usage as soon as the maximum risk of a revoke is lower than the maximum risk of a grant. The confidence values used for computing the maximum risk factors are derived from the CC's initial confidence directly at the begin of the *prediction step*: The lowest possible confidence $p_{max\_out}(t)$ of the CC at any point t in the *prediction step* can be computed by assuming all particles that satisfied the CC at begin of the *prediction step* to leave the AA on the shortest path. The highest possible confidence $p_{max\_in}(t)$ occurs when all particles that satisfied the CC at begin of the *prediction step* stay within the AA and thus still satisfy the CC throughout the *prediction step*. This value is a constant. Consequently, the corresponding maximum risk values for false-positive and false-negative decisions compute as $c_{fp}(t) \cdot p_{max\_out}(t)$ and $c_{fn}(t) \cdot p_{max\_in}(t)$ respectively. Directly when a *prediction step* starts, this allows to precompute a time-stamp t_{revoke}, representing the point in time when the risks of a false-positive are too high and thus when the revoke should be issued. The advantage of this approach is that our system doesn't need to evaluate the CC during the *prediction step* and is independent from the number of randomly moved particles that represent an attacker with their movement. Furthermore, this way the maximum time until the next update depends on the confidence of the CC at the beginning of a *prediction step* and revocations can be derived based on occurring risk factors.

V. Evaluation

In this section, the advantages of the proposed approach are evaluated in a comprehensive test set.

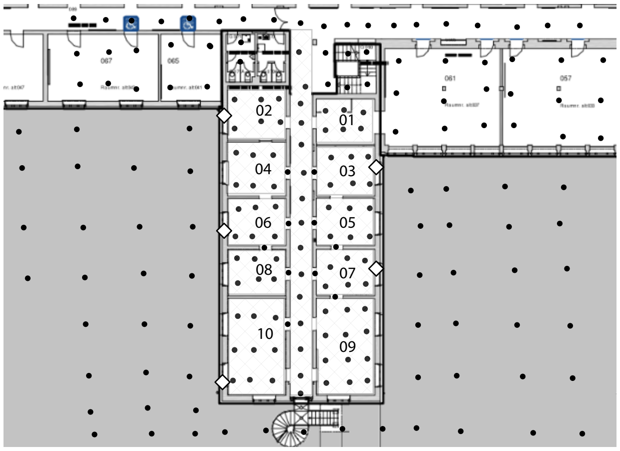


Fig. 4. Reference positions of our WiFi fingerprint database (black dots) and installed APs (diamonds). AAs were defined using the labelled rooms; outdoor areas are drawn in dark gray.

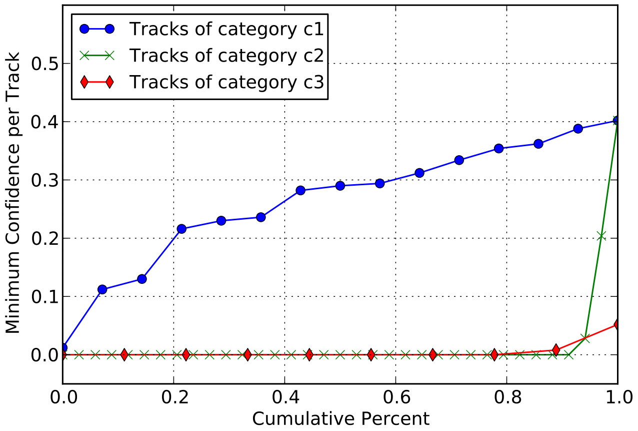


Fig. 5. Minimum observed confidence for each trajectory category.

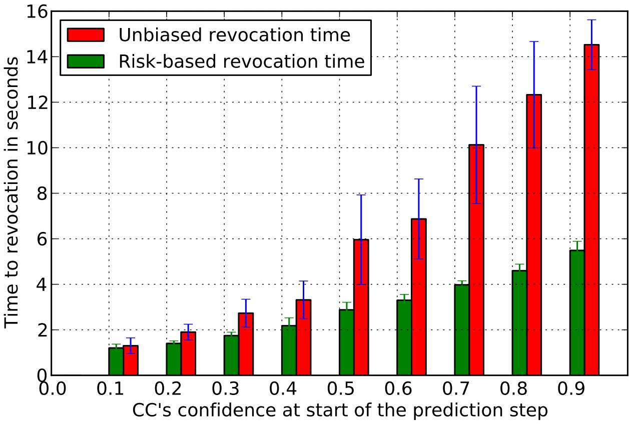


Fig. 6. Revocation times in dependence on the CC's confidence at begin of a *prediction step* for random movement and for the derivation of t_{revoke}.

A. Test Environment and Hardware

For evaluating the approach, we recorded a fingerprint database consisting of 206 WiFi fingerprints, each computed as the mean value of single 20 measurements with 4 for every cardinal direction. We installed 5 APs and employed their RSS values for positioning. The single WiFi fingerprints were collected using a HTC Desire smart-phone. The covered site comprehends about 1400 m^2 and is depicted in Figure 4. Additionally, we defined the following 5 AAs on the rooms contained in the hatched area of Figure 4, here referenced by their depicted identifiers: (01), $(06; 08)$, $(05; 07)$, $(07; 09)$ and $(01–06)$ along with the part of the floor in-between. For each of these areas, 12 possible user trajectories were recorded, each approximately 60 s long and consisting of a sequence of observed RSS values recorded at least every 1.5 s. Each trajectory's ground truth was supplied manually. For each modeled AA, the recorded trajectories can be classified as follows: Three trajectories that run completely within their AA, one with a user standing still inside. Three trajectories that run outside but near to the AA and inside the building, one with a user standing still. Two trajectories that leave and re-enter the AA, and three trajectories that run near the AA outside the building. The recorded trajectories represent three categories w.r.t. their ground truth: satisfying the AA all time (*c1*), violating it from begin on (*c2*), and satisfying the AA at begin but violating it later on (*c3*).

B. Evaluation Results

In order to detect trajectories that satisfy or violate a CC, the minimum observed confidence for a given trajectory should correspond to the trajectory's category. Hence, for each of the three categories *c1*-*c3*, the minimally observed confidence of assigned trajectories was computed and is depicted as a cumulative distribution function (cdf) in Figure 5. Clearly, in over 80% of all cases, the minimum observed confidence for trajectories of category *c1* was greater than 20%. In contrast, the trajectories of *c2* and *c3* showed a minimum confidence of 0% in over 90% of all cases. We considered these proportions in the definition of cost functions according to Section IV-C by choosing the ratio $c_{fp}(0)/c_{fn}(0) = (100\% - 20\%)/20\% = 4/1$. c_{fp} rises compliant with the sampling rate of measurements. The classification results based on our cost functions are compared to those of a single crisp trajectory derived from stringing together observed location measurements, conform to existing approaches to trajectory-based usage control [3]. Table I depicts the results. As the results indicate, the crisp trajectory has a true-positive (TP) rate of 13.3%, which makes its application nearly impractical. However, our approach yielded a TP rate of 80%. The crisp trajectory showed a true-negative rate (TN) of 100% in contrast to our approach, which showed a slightly lower TN rate of 93.3%. Consequently, our approach shows higher false-positive (FP) classifications and slightly higher chances of a misuse but outperforms the crisp trajectory with its TP rate by far, which results in a far higher availability of the usage right if really justified. To assess the benefits of the proposed *incremental_cc* evaluation, we determined the mean count of pruned particles for trajectories of each category. For trajectories in *c1*, *incremental_cc* could prune 22% of all particles in the mean in contrast to 94% for category *c2* and 72% for category *c3*. The outcomes show the strong correlation to the number of violating particles and indicate that the incremental evaluation is expected to prune at least 22% of all particles by mean. Finally, we compare the impact of deriving risks in the *prediction step* from the lowest possible confidence for any point in time instead of adhering the iteratively updated confidence deduced from the particles' random movement. The results are depicted in Figure 6. Clearly, the proposed model of assuming the lowest possible confidence yields more realistic revocation times, as in the other approach, particles often get stuck within certain rooms when following a mobility model with random movement.

TABLE I. CLASSIFICATION RESULTS.

Used Approach	TP	FP	TN	FN
Our BPF	80%	6.7%	93.3%	20%
Stringed location measurements	13.3%	0.0%	100%	86.7%

VI. CONCLUSION

In order to enable usage control policies to benefit from trajectory constraints in indoor scenarios, we proposed backtracking particle filters (BPF) to derive probabilistic trajectory estimates and discussed requirements to complicate attacks. Subsequently, we proposed the concept of containment constraints (CC), which require a user to stay within a certain authorized area (AA) for the duration of his usage of a protected resource. An improved evaluation strategy based on the discrete and probabilistic representation of potential trajectories was presented. In order to allow for a comprehensible revocation of usage rights we proposed to compute risk factors for a false-positive and a false-negative, choosing that decision with the lowest risk. In contrast to existing research, our approach respects occurring uncertainty of trajectory estimates and works on according probabilistic representations. This allows to exploit all available information when deriving usage decisions. Furthermore, to the best of our knowledge, no approach for deriving appropriate revocation times has been proposed before. Thus, the main contribution of this work is a mechanism for enforcing CCs in usage control policies based on probabilistic trajectories represented by particles, constructed by a BPF. In the evaluated indoor scenario, this concept shows a very encouraging true-positive rate of 80% at the price of a false-positive rate of 6.7%. However, for crisp trajectories created by stringing together single location measurements, like required by all existing approaches, only an impractical true-positive rate of 13.3% could be observed. Finally, our approach allows for a robust enforcement of CCs in indoor scenarios and to constrain mobile usage of resources to suitable areas and rooms without high extra expenses for additional positioning infrastructure. Future work should focus on integrating CCs in policies, hampering location spoofing with WiFi fingerprinting and using appropriate implicit authentication methods to couple user and device locations.

ACKNOWLEDGMENTS

Thanks to Sebastian A. Amft for helping with the test data.

REFERENCES

[1] E. Bertino and M. S. Kirkpatrick, "Location-based access control systems for mobile users: Concepts and research directions," in *Proceedings of the 4th ACM SIGSPATIAL Int'l Workshop on Security and Privacy in GIS and LBS*. ACM, 2011, pp. 49–52.

[2] J. Park and R. Sandhu, "The $\mathrm{UCON_{abc}}$ usage control model," vol. 7, no. 1. ACM, 2004, pp. 128–174.

[3] M. L. Damiani, E. Bertino, and C. Silvestri, “Approach to supporting continuity of usage in location-based access control,” in *12th IEEE Int’l Workshop on Future Trends of Distributed Computing Systems*. IEEE, 2008, pp. 199–205.

[4] Widyawan, M. Klepal, and S. Beauregard, “A novel backtracking particle filter for pattern matching indoor localization,” in *Proceedings of the first ACM Int’l Workshop on Mobile Entity Localization and Tracking in GPS-less Environments*. ACM, 2008, pp. 79–84.

[5] C. A. Ardagna, M. Cremonini, E. Damiani, S. D. C. di Vimercati, and P. Samarati, “Supporting location-based conditions in access control policies,” in *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*. ACM, 2006, pp. 212–222.

[6] H. Shin and V. Atluri, “Spatiotemporal access control enforcement under uncertain location estimates,” in *Data and Applications Security XXIII*, ser. LNCS. Springer, 2009, vol. 5645, pp. 159–174.

[7] S. Thrun, W. Burgard, D. Fox *et al.*, *Probabilistic robotics*. MIT press Cambridge, 2005, vol. 1.

[8] P. Marcus, M. Kessel, and M. Werner, “Dynamic nearest neighbors and online error estimation for smartpos,” *Int’l Journal On Advances in Internet Technology*, vol. 6, no. 1&2, 2013.

[9] M. Kessel and M. Werner, “Automated wlan calibration with a backtracking particle filter,” in *2012 Int’l Conference on Indoor Positioning and Indoor Navigation*. IEEE, 2012, pp. 1–10.

[10] G. Trajcevski, “Uncertainty in spatial trajectories,” in *Computing with Spatial Trajectories*. Springer, 2011, pp. 63–107.

[11] L. Chen and J. Crampton, “Risk-aware role-based access control,” in *Security and Trust Management*, ser. LNCS. Springer, 2012, vol. 7170, pp. 140–156.

[12] I. Molloy, L. Dickens, C. Morisset, P.-C. Cheng, J. Lobo, and A. Russo, “Risk-based security decisions under uncertainty,” in *Proceedings of the second ACM conference on Data and Application Security and Privacy*. ACM, 2012, pp. 157–168.

Mobile Agent for Nomadic Devices

Charif Mahmoudi, Fabrice Mourlin and Guy-Lahlou Djiken
Laboratoire d'Algorithmique, Complexité et Logique
University Paris-Est
Creteil, France
{charif.mahmoudi, fabrice.mourlin, guy-lahlou.djiken}@u-pec.fr

Abstract— Today, mobile devices are used as personal objects, which contain own data about our activities, our personal life, etc. Also, these data are essential for the user and it is not acceptable to exchange these data on the network. Then, computing on these data can be done by importing applications. We present our work about moving applications from server to client host. We use mobile agent technology with the difficulties of heterogeneous platforms. Importing agent has to respect features of the device (version, technical interface, business interface, etc.). A first negotiation is settled to ensure that a set of useful agents are selected towards an embedded device. By the end of a scenario, features are recorded to improve negotiation algorithm for the future exchange. So, each exchanged agent is not only used for a mission but also for the next ones. We use our framework for collecting data from nomadic devices. The mobile agents bring back these data to server where they will be treated.

Keywords-mobile OSGi; agent; nomadic device; REST services; transcoding.

I. Introduction

Mobility is a common word, which has different meaning depending on the working context. Mobile feature is often used for device like smart phone, tablet. In that case, mobile aspect highlights that user and device have the ability to move during runtime application. Distributed systems use mobility for actions. Mobile agents are often used for adapting an application to its environment. In that case, it means that the code of a part of an application can move from one node of the network to another one [1].

Two meanings of one word for two distinct contexts seem ambiguous and we should speak about nomadic user but it is too late for changing a so common feature of phone for instance. In our working context both aspects are used to build a distributed application where mobile devices are http client of distributed application.

Development of mobile applications is an activity domain which focuses a lot of designers and programmers. The number of libraries encourages new kinds of interconnection. Several protocols are used depending on the scope of exchanges. Bluetooth [2] and IrdA [3] are exploited for local exchanges, like file transfers with a laptop. A WIFI scanner allows users to connect all WIFI networks. This increases the scope of applications. For example, a client can register a broadcast receiver to perform the scan for new networks and improve its knowledge about environment [4]. Another application context is the developments of Binder framework for inter process communication [5]. It is designed to provide a rich high-level abstraction on top of traditional operating system services

Traditional software architectures use network as input or output data flows. Routes are defined between clients and servers and these are used to exchange information. These routes can be configured dynamically. But when a client is mobile, it enters into an infinite process to adapt current configuration. Also, this update costs time and energy. Another approach could be to update configuration only before communication. But, these events are not always predictable and some of the updates might be useless.

We describe our approach in the next sections of this document. First, we explain in detail our software architecture. Then, we focus to agent server and how agents are exposed over http protocol. Next section is about agent host and the lifecycle of the incoming piece of code. Finally, we describe one of our case studies based on a data collection which gives information to the server about end user activity.

II. Related Work

The Android platform covers a large software solution for mobile devices from an operating system to a set of mobile applications [6]. The two competing: Windows Mobile and Apple’s iPhone allows simplified development environment built on proprietary operating systems that restrict interactions between applications and native data. Android offers an open source development environment built on a Linux kernel [7]. Android offers also a standard API to access to the device hardware; this API allows the application to interact Wi-Fi, Bluetooth, and other hardware components.

The Open Gateway Services interfaces (OSGi) platform [8] defines a standardized, components-oriented computing approach for services. This approach is the foundation of a Service-Oriented Architecture (SOA) based system. The OSGi specifications were created to cover the exposition of residential internet gateways for home automation software [9, 10]. However, the extensible features of OSGi technology contribute to impose it as key solutions. Several devices vendors, like Nokia and Motorola, have choice OSGi standard as a base framework for their smart phones. This choice is due to the features provided secure support for developing and deploying java service component applications packaged in a standard Java Archive (JAR) file called bundles. A bundle contains a manifest file that describe bundle's configuration on the OSGi container. Figure 1 illustrates the OSGi framework, applied to our approach, which provides a shared runtime environment capable of dynamic and hot lifecycle management operations: installs, updates, and uninstalls bundles. No need to restarting the system after a management operation.

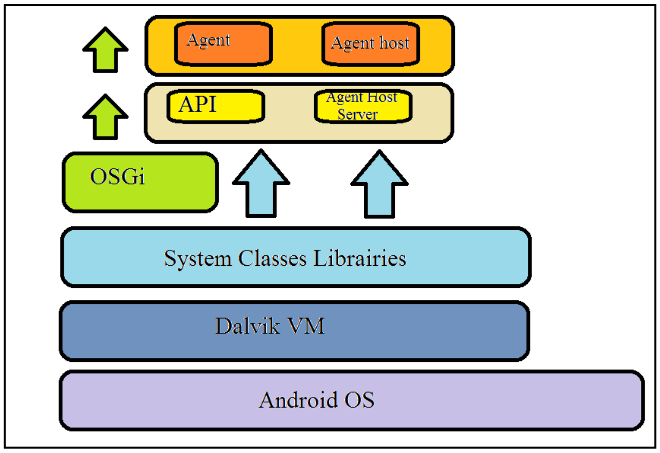


Fig. 1. Agent runtime environment architecture.

Other research works in literature have been focused in the use of agent-based technology on nomadic devices; examples include JADE [11], JaCa-Android [12], AgentFactory [13], and 3APL [14]. Differently from these related works, our approach do not consider the issue of porting agent technologies on limited capability devices, but we focus on the autonomous agent mobility issues, we focus also on the advantages brought by the adoption of an agent-based solution for the development of mobile agents that have decide when and where to move in a complex mobile system.

We propose a new approach based on mobile agents [15] that run into an Android embedded OSGi runtime environment. Our work changes software architecture into a new scheme. When an agent incomes onto a device; it has knowledge about network configuration. This agent collects business data and by the end of its mission, it can return to a server or other clients.

III. Mobile Software Architecture

The software architecture of distributed application is a picture of the system that aids in the understanding of how the system will behave. Such an architecture is depicted by a component graph where each node is a software component. A first observation enhances three main types of components: agent server, agent host, and mobile agent.

The role of agent server is to receive the requests from clients then it records the demands and creates or selects mobile agents. Finally, it exports agent to agent hosts. The role of agent host is to send its demand to a server and then listen to the answer of the server. When the reception occurs, it engages the mobile agent into a state where it is able to execute its own mission. To sum up, a mobile agent is a piece of code which travels over the network. Initially, it is configured by the server for navigating through a set of agent hosts. Its aim is to be as autonomous as possible even if it has to use local resources of host client. Security concerns have to be set first on all the devices which will participate to a case study. A more common definition of mobile agent can be found in Bernichi [16].

A more precise observation stresses technical requirements about network exchange and also message structure between the components. Thereby, a client communicates though the use of REST Web Services [17]. This involves that the underlying protocol is http. However, type of message is considered as a byte stream from the hosts to the mobile agent and vice versa.

A. *A first level observation*

First of all, we provide a deployment diagram where main nodes of architecture are drawn. All connection mobile software support http as the transport protocol, making each compatible for use with the server through a solution on Android devices. No security constraints are applied in the first version of the prototype, but our security approach is developed in subsection III.C.

Server is accessible through WIFI card and it may connect to any standard WIFI router, which is configured as an Access Point (AP), and then, sends the data to other devices in the same network such as basic phones and smart phones. Figure 2 represents the main items: a server which supports an agent server; mobile devices support mobile nodes and agent hosts. WIFI access is provided by antenna or access point. The structure of the graph can be permanently evolved. Of course, other items can be added for example Bluetooth devices.

Once associated with the Access Point, the nodes may ask for an IP address by using the DHCP protocol or use a preconfigured static IP. The Access Point connection can be encrypted, in this case, we have to specify also the pass-phrase or key to the WIFI module:

Nodes may also connect to a standard WIFI router with DSL or cable connectivity and send the information to a web server located on the Internet. Then, users are able to get this information from the Cloud when a static configuration is used.

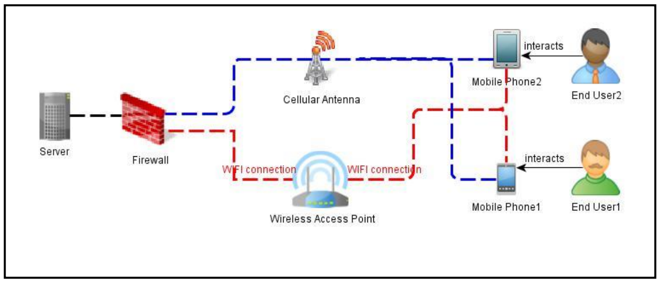


Fig. 2. Deployment diagram of our following case study.

B. *A second level observation*

After a first physical description, we provide a more synthetic architecture description where components are isolated. In this context, a component is a modular unit that is replaceable within its environment. Its internals are hidden but it has well defined provided interfaces. Three main components are designed with their dependencies.

The main component is called Agent Server on the server station. It provides two interfaces. One is used for registering the demands of the hosts. Another interface allows mobile agents to exchange data with Agent Server. This interface is particularly important for the end of mobile agent activity.

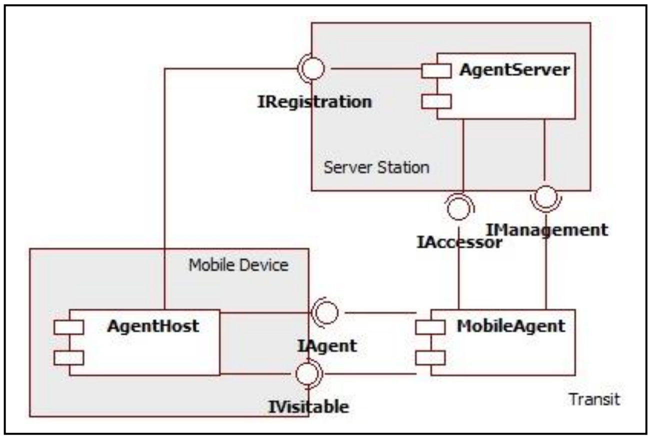


Fig. 3. Component diagram :software architecture.

As illustrated in Figure 3, on each mobile device, a component called Agent Host is installed. It sends demands on mobile services by the use of *IRegistration* interface. This component implements an interface called *IVisitable*. It allows mobile agent to come into the context of agent host.

The component called Mobile Agent, is created and configured of the server station, then it uses *IVisitable* interface of agent hosts for the exportation. It provides an interface called *IAgent*, which is used by the host to launch the runtime of mobile agent on the nomadic device.

The interface *IRegistration* and *IVisitable* are remote but *IAgent* is a local interface. All remote accesses are mapped on http protocol methods. The technology REpresentation State Transfer (REST) [18] is chosen. A resource-oriented approach is well adapted to the interaction scheduling. A resource is a mobile agent for the agent server or an agent host. Or a resource can be an agent host for a mobile agent. It means anything of potential interest that is serializable in some form. The acronym REST refers to the transfer of some bit of information or mobile agent state, as a representation, from a server to a host or back again.

But, another technical constraint appears: the kind of byte code is not the same between the server station and the mobile device. Also, an agent which is built by the agent server on the server has to be transcoded before moving to an agent host. The transcoding means to know the technical features of all agent hosts. Also, we decided to encode mobile agent into an intermediate representation and each local host will adapt the representation though the use of an agent loader. This strategy is close to class loading but the algorithm is not only about loading the byte code of the agent but also on the permissions that are assigned to the agent.

C. security approach

When agent have to access to a resource exposed by current agent host, the agent use URI to address and get the resource. Android SDK 1.0 introduced a security mechanism to manage URI based security. An agent can specify resource URI identifying an XML file. If the agent doesn’t have read permission to the agent host containing the XML file, the agent can use its own URI permission instead. In this case, the agent uses a read flag that grants the agent host access to the resource. URI permissions are essentially capabilities for agent execution. This mechanism allows a least privilege access to the agent host. The tractability of policy is preserved truth the agent host.

In the next section, each component is described to illustrate the mechanisms of code mobility and monitoring the collection of information. The structure of the API allows readers to understand the case study presented is the following example.

IV. AGENT EXPOSITION

The role of each component is crucial in defining this new software architecture. Also in this section, we will focus on the structure of each and how is implemented software mobility.

A. Agent Server component

The *AgentServer* component manages the demands of agent hosts and the pool of mobile agents. Its first role is to receive queries from agent hosts and treat them. To do this, it is necessary to detail the structure of a query. It provides two main records: the need for intervention, a reference for the visit. The intervention of a mobile agent means a remote activity is requested by an agent host. This expression is made by the demand of a technical interface, which is associated to permissions for access to local resources. As part of this work, we have not installed directory services where these interfaces have been reported. This should be done in a real environment.

1) Design constraints

We show in section C the structure of a mobile agent. We have decided that all mobile agents can provide only one business interface for the purpose of simplification and optimization. In fact, offering one business interface allows the server to configure a mobile agent that travels to several agent hosts. Each agent host has provided its reference in its request. Then, a mobile agent manages a list of references to agent hosts.

An agent server not only has the role to create mobile agents as well as receiving mobile agents by the end of their mission. This means having traveled all agent hosts; a mobile agent ends at the agent server. Both activities are dependent and yet receiving mobile agents must take precedence over creation because the pool of agents waiting mission is to be reused for future requests. To implement this constraint, we have defined an agent server as composed of two main threads. The first is to the end of mission, the second for the creation of agents. This thread is of lower priority to the first interrupt by creating a return.

When returning mobile agents to the server, it is responsible for the extraction of data and their backup in the local system of persistence. In our case study, it is a database server. These data are then used by other applications so that future interventions will be planned. For example, when collecting data for monitoring Web server, the messages on anomaly deployment involve the creation of new demands of mobile agent. So, a mobile agent will be exported onto the host where the deployment fails. This update will involve new messages during the next data collection and so on.

In our prototype, all agent hosts are treated equally by the agent server. However, the result of the activity of mobile agent cannot be validated by agent host itself. Also, this could be the role of other mobile agents into a control activity. The goal at the server level is to ensure that any received request by the server is handled by a mobile agent. In another work context, it would be possible to keep track of a set of requests to consider exporting mobile groups of agents in a single export.

We did not secure the exchange of data between the server and mobile agents in the context of this work. We considered that the data collected by mobile agents were visible to all components of the distributed system. This simplification allows us to reach a rapid prototype supporting our functional tests, but also our load tests.

2) Component architecture

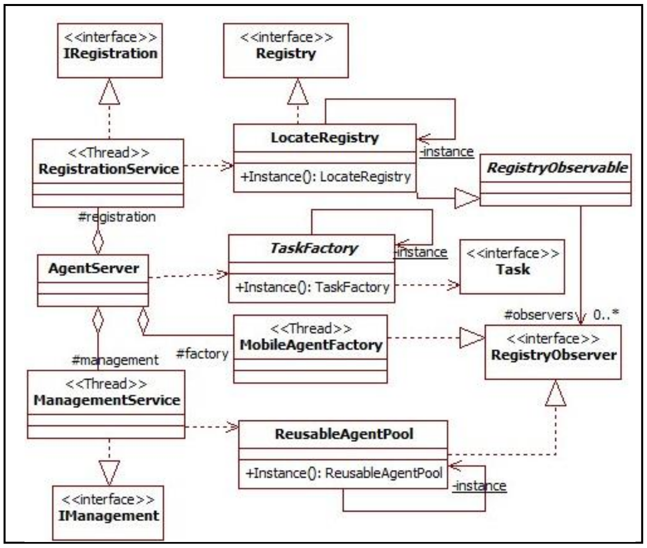


Fig. 4. Class diagram of agent server component

Figure 4 gives a first structure of the classes which belong to the *AgentServer* component. New interfaces are present such as *Registry*, *RegistryObserver*. These interfaces are not exposed by the *AgentServer* component, but are useful as local interfaces within the component itself. Thus, *RegistryObserver* interface is implemented by all classes that are prone to react to requests from a host. Implementation classes are the pool of mobile agents and the class of mobile agent factory.

The class diagram above highlights some design patterns. The data structure that records the host requests is an observable subject by all observers who treat them. The factory of mobile agents is an observer but also the pool of mobile agents ready to be configured for a new mission. The uniqueness of some essential items in the *AgentServer* component is implemented by the use of Singleton pattern. This is the case for *LocateRegistry* class has only one instance for recording all requests from the host.

Technical classes are not present on the diagram in Figure 4 for clarity. For example, management of tasks requested by the host is missing from the figure. In addition, the agent server uses a task set to inject during the configuration of mobile agents. This means the injection of code into an agent, which is already created, but also initiates an initial context for the task performed by the agent.

B. Agent Host component

By definition, the host agent is installed on the nomadic device such as tablets and smartphones. In our current prototype, we have chosen to implement an agent host per device. Thus, the host can be seen in the distribution as representative of the software device that supports system. This component manages the identifiers among other material information.

1) Analysis requirements

A host of agents exhibits a single remote interface to be visited by requested mobile agents. The advantage of this approach is better interactivity. Thus, the interface is published in the directory of the host interface (*Registry*) only when the request from the host has been accepted by the server. In addition, this interface will be removed as soon as the mobile agent has visited the host.

The expression of the need for intervention is difficult because it is necessary that the host is aware of the types of jobs available on the server. As part of this prototype, we made the choice to have a set of interfaces known agent hosts and agent server. It is clear that this global knowledge is not desirable because it all tipsy scalability tasks. More particularly, this approach prohibits the dynamic task creation. This concept will be addressed in the next increment of our prototype.

The structure of an agent host has the particularity to execute a business process but also to involve a mobile agent. This would happen following a request from the host. It is important at this stage to focus on the principle of mobility that we have implemented. A strong constraint is the byte code differences between the platform server and nomadic devices. A host agent cannot execute code from a mobile agent created on the server. It is necessary to transcode this code to adapt to the device from which the request comes. The technical risk was initially important in our project and encouraged us to assign our increment to it. This conversion is done on the server. When the agent host receives the byte code of the class of the mobile agent, it must load it via a local agent loader and configure from the state of the serialized (with its class) agent.

Then, the mobile agent is executed in accordance with the life cycle of the host agents. This means that a thread is dedicated to the mobile agent and activity. The priority of this thread is normal to let the opportunity to agent host to launch higher-priority threads. Finally, specific permissions are assigned to *AgentHost* component in order to perform network communications and access to some useful sensors for the business job.

2) *Design patterns*

The main structure of an agent host is based on the *Command* design pattern. Figure 5 shows that design choice. When a mobile agent invokes *IVisitable* interface to enter the host, the client of this pattern is the *VisitableService* service. It creates a concrete *MobileAgent* instance from input byte data and sets its receiver.

The class called, *AgentLoader* promotes the byte code array into a class which can be used by agent host. The agent thus received is added to the instance of *WaitingRoom* class is the data structure that are placed all mobile agents received awaiting use. This instance acts as an *Invoker*. It asks the concrete mobile agent to carry out the activity. *MobileAgent* abstract class declares an interface for executing the task. *CollectorAgent* and *MonitorAgent* define a binding between a *Host* object and a task. It implements methods of public interface *IAgent*. We use the concrete command *CollectorAgent* into the next case study. Its role is to invoke the corresponding operation on host. In the next example (section V), the data are extracted from data files. The *Host* class knows how to perform the operations associated with carrying out a request. This class serves as a receiver for all the concrete commands.

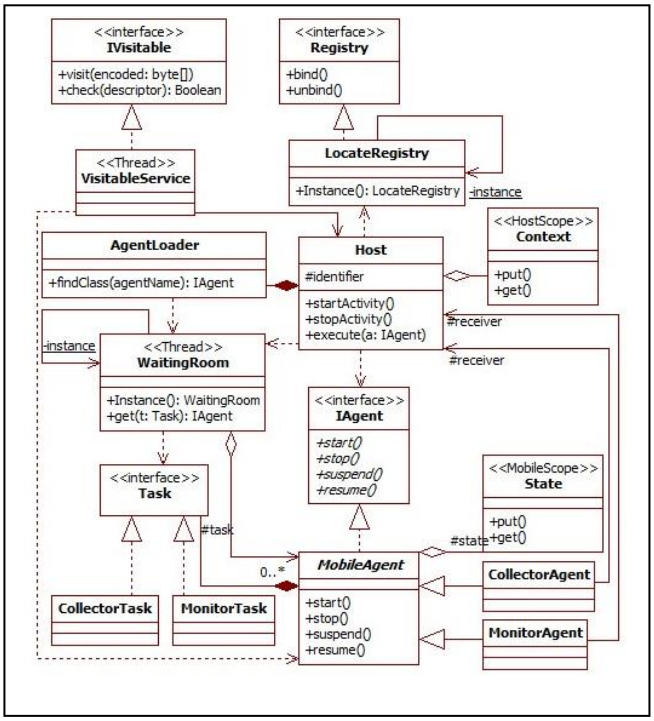


Fig. 5. Class diagram of agent host component

A *Host* object represents a mobile location (any possible place) where a mobile agent may run. Thus it is represented as a physical resource with computing capabilities. An instance of this class is always bound with the corresponding native host. Several local information are stored by the host in order to be recognized as unique in the wireless network without using information specific to the protocol itself.

The *Context* class (Figure 5) shows the data structure used by mobile agents as they pass on the host. Two scopes of data are useful. First data *host* range: they are resident data that the mobile agent can take starting from the agent host. These data are useful for its current activity and the realization of its mission on this host. *Mobile* scope data are those that can move with the mobile agent during its travel from host to host. Usually, these data are the overall mission must make the mobile agent. By the end of the mission, the extraction results come from scope *mobile* data. They allow an external application to validate the success or not of the mission of the mobile agent.

C. *Mobile Agent component*

The concept of mobile code we use; is a mobility on demand, because it is controlled by the applicants. In our case, the server and the hosts are the assets of our distributed system. Hosts request and the server provides services. But in our case the product is not a set of raw data, it is a code that intelligence can make a mobile service. By configuring the server, the agent knows the references hosts and thus moves from host to host in order to apply the only task that knows. Of course, this task can use specific *host* data or specific *mobile* data.

1) *Mobility on demand.*

This expression is often associated to electric vehicle in a city. But, in our working context, this represents use of mobile agent in a distributed system. To be useful the mobile agent needs the host offers an agent loader. Then, its statement is managed by the lifecycle of the host. Thus, it is the host that will decide the launch of its task or its interruption, etc.

The end of normal task of an agent is reached the *stop()* method is executed. In order to comply with a protocol of the easiest possible use, we decided to adopt a balanced approach than the arrival of the mobile agent on a host. Thus, the execution of the *stop()* method is followed by the data backup (local to the host or mobile), and then moving the agent to its next destination. The last destination will be the agent server. Once the migration is done, the code of the agent is discharged. And two successive interventions of an agent of the same type will be considered as two different interventions.

When looking for the next host, the mobile agent performs a search of the reference that it has received from the server during configuration step. There can be only one registry per host. In our case, each host has permissions to manage its own registry but can only lookup into abroad registry. Also, the search is performed by a multi cast on all of these registries to find out the reference of the next destination host.

In the context that such a reference is not found then the mobile agent fails to migrate to the next destination. If this reference is last then he will go on the server that originally configured. The final extraction of the data is made by the server. It calls the *IAccessor* interface that is managed by the component *MobileAgent*. We designed this refund so that all the data is aggregated results in a data structure which only read access is possible through the use of *IAccessor* interface. At its next configuration, the contents of these results will be deleted from the memory of the mobile agent as for all *mobile* data scope.

On the server, the management of mobile agent in done by the data structure called *ReusableAgentPool* which is an iterable structure. We have not developed a technique to eliminate redundant in the waiting pool agents. Another strategy recommended in the work of load balancing is to let the agent on the last host who asked. The underlying idea concerns the fact that an agent is more useful from a client to a server.

2) *Design and change*.

The mobile agent design is simple as it is important that the agent is small so it can be easily encoded and decoded. By construction, a mobile agent is a running activity in the context of remote work. All types of work are not present on the charts but simply *Task* interface.

To be independent, a mobile agent must be configured. It comes from the server contains a list of steps that are the mission of this agent. A step is the application of a task on a host. By the end of the application of a task is achieved and added to the result list (class called *ResultManagement*).

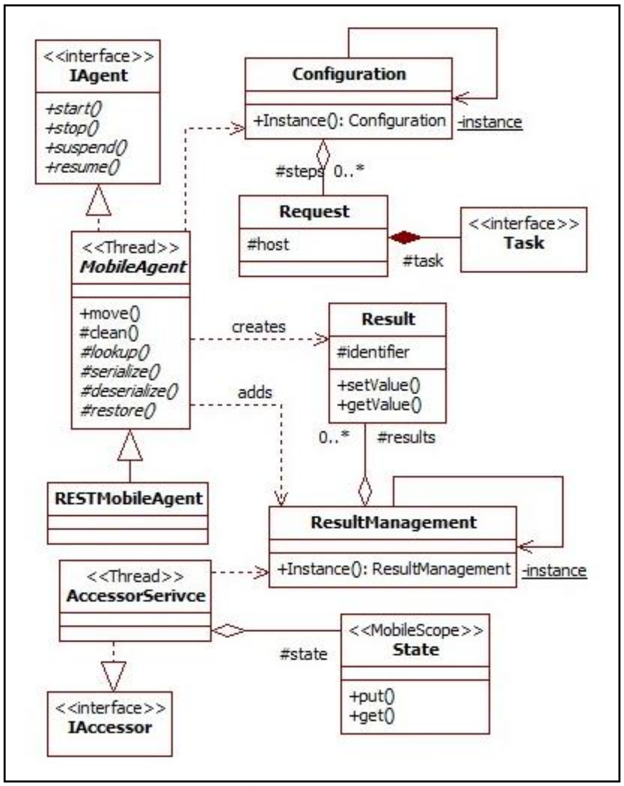


Fig. 6. Class diagram of mobile agent component

Thus, by the end of the mission, the list of steps is empty while the list of results is full. Each of the support structures of an iterator that enables an enumeration of the data structure in the same order as that of the host list.

The strategy of migration is achieved by the *move()* method. This is a step algorithm which is realized by the use of *Template* design pattern. These steps are implemented using abstract methods. Subclasses change the abstract methods to implement the real actions. Thus the general algorithm is saved in the class called *MobileAgent* but the concrete steps may be changed by the subclasses. The refined implementation is done in *RESTMobileAgent* class, where all the steps are implemented as REST web services (Figure 6).

The REST approach is oriented around resources. The resources support often access though get, post, put or delete actions. We built a whole REST prototype, but our design could support other kinds of remote access if necessary.

Next section is about how mobile agent activities are inserted into the life cycle of agent hosts.

V. AGENT HOST LIFECYCLE

Deploying the set of components, they are started at installation. And any agent host is ready to receive a mobile agent after an initial setup phase. Management of mobile agents or constraints specific to each host and belong to the configuration of the host. This is done through the use of XML descriptor.

A. *Component descriptor*

The Mobile Component Descriptor (MCD) describes the properties of agent host. The MCD contains following required information such that the technical name of the component, the version and a description of the component. Specific features are added such that component type (Server, Host, Mobile), runtime environment, etc. Next, the behavior of the component is described as phases of an automaton. The way to specify them is extremely simple and is divided into 4 parts: parameter, transport receiver, transport sender, phase order. These parts are included into `<agentHost/>` tag.

1) *Parameter*

A parameter is a name-value pair which is used by the component. Each and every top level parameter is transformed into properties in Configuration instance of component. The correct way of defining a parameter is as follows:

```
<parameter name="identifier" value="HostD1"/>
```

2) *Transport Receiver*

Depending on the underlying transport on which agent hosts are going to run, different transport receivers are defined as follows:

```
<transportReceiver name="http"
class="fr.upec.lacl.device.host.ReceiveController">
    <parameter name="protocol" value="http"/>
    <parameter name="port" value="8888"/>
    <parameter name="version" value="HTTP/1.0"/>
</transportReceiver>
```

The `"name"` attribute of the `<transportReceiver/>` element identifies the type of the transport receiver. It can be HTTP, TCP, etc. But, because we use Android device, HTTPf is selected.

When the host starts up the "class" attribute is for specifying the actual java class that will implement the required interfaces for the transport. Any transport can have zero or more parameters, and any parameters given can be accessed via the corresponding transport receiver.

3) *Transport Sender*

Like the previous section, Transport senders are registred in the configuration of the host. And later at the runtime, the exportation of mobile agent will follow this feature, we defined HTTP as transport.

```
<transportSender name="http"
 class="fr.upec.lacl.device.host.SendController">
    <parameter name="protocol" value="http"/>
    <parameter name="port" value="8890"/>
    <parameter name="version" value="HTTP/1.0"/>
</transportReceiver>
```

The sender can have zero or more parameters. In the frame above, the port is defined and also the schema of the transport protocol.

We have chosen the same protocol in both cases, but we think about protocol adapter between nomadic devices. For instance, we think about protocol adapter between nomadic devices and the use of Bluetooth protocol as transport protocol and OBEX serialization.

4) *Phase Order*

Specifying the order of phases of an agent host in the execution chain is essential to know when mobile agent can interrupt the host.

```
<phaseOrder name="lifecycle"
         package="fr.upec.lacl.device.host">
 <phase name="init" type="start" class="Init"/>
 <phase name="business" type="loop">
  <handler name="observer" class="Display"/>
  <step name="step1" type="exec" class="Wave1"/>
  <step name="step2" type="import" class="Import"/>
  <step name="step3" type="exec" class="Wave2"/>
  <step name="step4" type="export" class="Export"/>
 </phase>
 <phase name="end" type="stop" class="Close"/>
</phaseOrder
```

Each phase can have a handler, which observes or displays details about the phase. In the example above, the class *Display* has a method called handle(), which does the observation.

A phase is defined by a name and a type which corresponds to an event in the behavior of the agent host and an implementation class. For instance, the phase named *init* has a type called *start*. It means that the *startActivity()* method of *Host* class (Figure 5), triggers its behavior defined in class called *Init*, and its *doWork()* method. All the phases are defined in the same manner. Thus, the phase named *business* represents the core of the agent host. Its type implies that this is an infinite loop which is subdivided into a sequence of four steps. The first one, called *step1*, is defined by a class called *Wave1*. This is triggered by the *execute()* method of *Host* class. The following step allows host to import a mobile agent. This step is leaded by the *Import* class. The third step corresponds to the end of the business activity of the agent host. This is defined by the *Wave2* class. Finally, the fourth step allows host to export agent if its mission is ended. Otherwise, this step is blocking until the end of the behavior of the mobile agent.

The end of that loop is achieved by the use of interruption from *execute()* method. Then, the phase named *end* is achieved. As before the type called *stop* is bound to *stopActivity()* of *Host* class. So, it triggers the behavior coded into the class called *Close* and its method *doWork()*. This brings a set of technical classes which are not on the figure 5, but they represent a *State* design pattern where each state of the state chart is defined by a class and polymorphic methods. Two steps are dedicated to mobile agents: import and export. Between those events the mobile agent is used by the host.

B. *Mobile agent as activity resource*

During the execution of a mobile agent, we can consider it as a thread. As all threads on the nomadic device, it has access to resources, which are internal or external, depending on the permissions it possesses. Because it is not possible to add permissions avec the agent is arrived, it is important to prepare its arrival. This step is called negotiation between the mobile agent and the agent host.

From the side of the agent host, it needs to receive an agent able to do a technical interface. On the side of the mobile agent, it needs to have access to resource which belongs to the device. If the mobile agent comes onto the host and discovers that it cannot do its job, time is wasted and computing resource are used for nothing. Also, mobile agent has a description of the resource, it needs to read or write. As an example, we give below the resource description of the collector agent (Figure 5).

```
<resources agent="ca1" class="CollectorAgent"
    package="fr.upec.lacl.device.mobile">
 <resource name="contacts" mode="read"
    uri="content://contacts/people/"/>
</resources>
```

The resource list contains only one resource which is used as a reader. This resource is defined by an URI which is parsed by the host to know whether it is possible to read. When the condition is validated then the migration can happen.

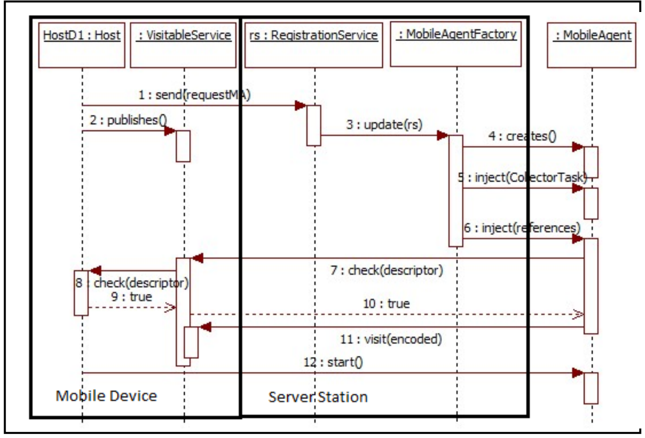


Fig. 7. Interaction diagram of negotiation protocol

To sum up our negotiation protocol, we show the following interaction diagram which is applied at each migration action. After sending its request, the host registers its proxy and waits until a mobile agent is ready (Figure 7).

After configuring a mobile agent, the server injects a task and a list of references into the mobile agent. Then, it can enter into negotiation with the first host of its reference list. The stimulus 7 (Figure 7) tests whether the collector activity is possible onto the host. Because all resources are declared with uri string, the host parses each of them and decides whether its resources are available.

In Figure 7, the answer of the host is true; also the mobile agent can starts the visit (stimulus 11). At that point the mobile agent is viewed as a binary resource which is read by the read and transformed into a class of agent. This is the role of the *AgentLoader* class which is not design on the Figure 7.

During its runtime, the mobile agent will access to the resources, which are declared into its own descriptor. Because the previous check is satisfied by the host, it means that not only the host provides a way to access to them, but also it ensures that the permissions of the host are sufficient to realize this resource accesses. At that point the roles are opposite, the host can be considered as a resource manager for the mobile agent.

When the task of the agent is finished, it saves its result and looks for the next host reference. If the reference exists, then the stimuli from 7 to 12 will occur again (Figure 7) with this new host, and so on.

C. Instrumentation

In order to identify the root cause of bad behavior, instrumentation has to be introduced. We are interesting into two kinds of anomalies. First kind is about phase tracking. Because hosts are defined as automaton, we want to ensure that the event traces are precisely what is predicted. Second kind is about performance of mobile agents. It is not easy to predict the number of mobile agents is necessary for a set of tasks. If twenty hosts send the same request, only one mobile agent can be sufficient. But, depending on what it has to do, the size of this agent can grow and the serialization and deserialization will cost more time. Moreover, the work of one mobile agent will last more time because, all visit will be done sequentially. Also, it could be interesting to create and configure a set of mobile agent, but how many mobile agents?

These measures need to use external libraries. For the host phase tracking, we use JMX API [19], which is a standard for management and monitoring of resources such as hosts and mobile agents. During the following case study, we observe phases, firing transition, configuration of agents and the notification of state changes during the data collection. We have defined *MXBean* classes which are managed remotely by the *MXBeanServer* of the JVM (Java Virtual Machine) of the server. The tool *jvisualVM* (Sun/Oracle) is used to display their results and allows users to interact with our distributed system.

The time spend within code fragments of mobile agent is interesting to find a limit into the use of mobility. It is important to detect the threshold where one mobile agent costs more time than two. Again, we can observe the impact of the *mobile* data on the migration action. This can involve changes in the management of mobile agents on the server. We use JETM [10] Java Execution Time Measure, which perfectly fits in this kind of time measure. It can be used declaratively and programmatically and collecting data are recorded in a flat or nested manner. An advantage is an http console on port 40000 which is used to visualize execution timings in the form of a dynamic report. We have injected *EtmPoint* instances into the methods of mobile agent and task and we follow all the steps of the behavior of a collector agent.

VI. Case Study

Our work has several technical aspects which are necessary to validate through a case study. This case study has to be understandable even by a non-developer. Also, we have chosen to collect data about the personal contacts recorded into a smartphone or a tablet. The example has several advantages. First, everyone knows the concept of contact into a phone book; this is a tuple of string and number. The size of a set of contacts can be big enough to raise exceptions during the data transfer.

Secondly, this resource is easily used through the use of uri, permissions about it are well known and a whole contact is serialized automatically. Finally, it is easy to check whether our tests are checked, failed or in an error status.

A. Synospsis

An experiment in developing small mobile phone application is not new, but in our context the architecture is more complex. There are a standard server workstation and four nomadic devices. Software is installed on the server to deploy Web services in REST technology. It means Apache Tomcat and Jersey libraries.

First, we have defined a test suite composed on four tests; each of them managed a different strategy of mobility. The first test uses only one mobile agent which travels aver the four devices.

Secondly, we have increased the number of contacts on the devices. Again, this test suite contains four tests where the size of contacts is higher each time.

Finally, we have tested anomaly in case of the descriptor is not compatible with the host. the descriptor is not compatible with the host.

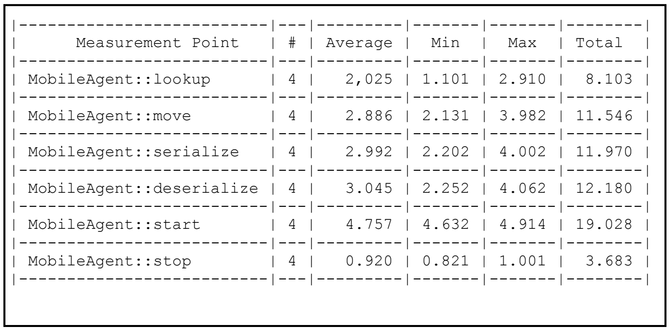
B. Measure and trace event

All the time measures are expressed into millisecond (ms). Because data set are difficult to read, we present only extremes.

1) *First test suite about mobility strategy*.

a) One mobile agent for 4 agent hosts: this array shows only the bounds; we can note that the serialization costs quite the same time as the task of the mobile agent itself. The same remark is true about deserialization. Also, in that case, it could be interesting to limit the sequence of actions of the mobile agent. A better solution could be to launch in parallel several mobile agents.

b) Two mobile agents per two agent hosts: in that context, the measures are more difficult to exploit because each mobile agent has its own array of result, also it is necessary to aggregate the results and use a global reference to the clock of the agent server. If the whole data collection spends less time than in the first case, the number of serialization are strictly the same but distributed over the 4 agent hosts. We observe that a global time measure from agent server shows that two mobile agents work faster than only one. But, this gain comes from the distributionof mobile agents. Now, a whole time measure is not a basic sum of all steps. Two partial collects are done in parallel and interesting results come quickly with two mobile agents.



```
|--------------------------|---|---------|-------|-------|--------|
|      Measurement Point   | # | Average |  Min  |  Max  | Total  |
|--------------------------|---|---------|-------|-------|--------|
| MobileAgent::lookup      | 4 |   2,025 | 1.101 | 2.910 |  8.103 |
|--------------------------|---|---------|-------|-------|--------|
| MobileAgent::move        | 4 |   2.886 | 2.131 | 3.982 | 11.546 |
|--------------------------|---|---------|-------|-------|--------|
| MobileAgent::serialize   | 4 |   2.992 | 2.202 | 4.002 | 11.970 |
|--------------------------|---|---------|-------|-------|--------|
| MobileAgent::deserialize | 4 |   3.045 | 2.252 | 4.062 | 12.180 |
|--------------------------|---|---------|-------|-------|--------|
| MobileAgent::start       | 4 |   4.757 | 4.632 | 4.914 | 19.028 |
|--------------------------|---|---------|-------|-------|--------|
| MobileAgent::stop        | 4 |   0.920 | 0.821 | 1.001 |  3.683 |
|--------------------------|---|---------|-------|-------|--------|
```

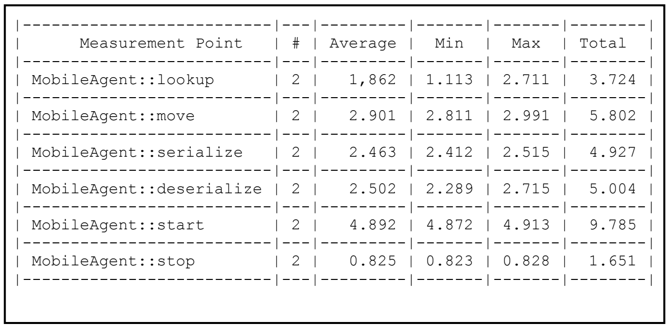
Fig. 8. Data results for one mobile agent

2) Second test suite about volume of data set.

Now there are four tests where each agent host has the same number of contacts. But, for the second test, we have doubled the number of contacts per agent host. For the third test, we have multiply by three, and so on. All the data collections are done by two mobile agents as before.

a) Each agent host manages 100 contacts: two mobile agents collects them. This case corresponds to the last experiment (Figure 9).

b) Each agent host manages 400 contacts: the same number of mobile agents collects data. We observe that the duration of the task is quite the same in all the test cases but the serialization and deserialization steps are more expensive. Also in case of the data size is huge, we note that it is essential to increase the number of mobile agents. Thus, the size of mobile data will be bound and the cost of the serialization could be constant.



```
|--------------------------|---|---------|-------|-------|--------|
|      Measurement Point   | # | Average |  Min  |  Max  | Total  |
|--------------------------|---|---------|-------|-------|--------|
| MobileAgent::lookup      | 2 |   1,862 | 1.113 | 2.711 |  3.724 |
|--------------------------|---|---------|-------|-------|--------|
| MobileAgent::move        | 2 |   2.901 | 2.811 | 2.991 |  5.802 |
|--------------------------|---|---------|-------|-------|--------|
| MobileAgent::serialize   | 2 |   2.463 | 2.412 | 2.515 |  4.927 |
|--------------------------|---|---------|-------|-------|--------|
| MobileAgent::deserialize | 2 |   2.502 | 2.289 | 2.715 |  5.004 |
|--------------------------|---|---------|-------|-------|--------|
| MobileAgent::start       | 2 |   4.892 | 4.872 | 4.913 |  9.785 |
|--------------------------|---|---------|-------|-------|--------|
| MobileAgent::stop        | 2 |   0.825 | 0.823 | 0.828 |  1.651 |
|--------------------------|---|---------|-------|-------|--------|
```

Fig. 9. Data results for two mobile agents

C. Interaction between agents

In the previous, tests we developed cases where there is only one mobile agent per host. Also, no conflict is possible between their activities. But if the data set is too important, we can think about test case where more than one mobile agent will be received by an agent host. So, the first mobile agent could collect a part of the contacts and the second one could collect the other part.

This scenario involves knew ability for agent host and mobile agent. First, if several mobile agents are present on an agent host, each of them should have to manage its own data without any perturbation from the other agents. In that case, the agent host should have to have one agent loader per agent host. So, each mobile agent will be separated by construction. Secondly, if more than one mobile agent realized a task, they have to exchange information or share flags about their own activity. For instance, the first agent collects the first part of the contacts (from one to hundred) and the second collect the next hundred contacts. The cost of the serialization becomes predictable, but mobile agents have to exchange messages during their execution.

This concept of message is implemented in Android framework through the use of Intent service. But, mobile agent comes from an agent server which is not under Android. Also, we have to develop a layer of exchange on the agent hosts to allow mobile agents to have better cohesion.

VII. CONCLUSION

In this paper, we have shown that it was possible to use mobile agents which are interoperable between a JVM and a Dalvik virtual machine. Our work was applied in the context of a data collection. This is a famous example useful in a lot of cases. We have applied an approach of transcoding to adapt byte code from JVM to DVM and vice versa. Measures are computed to highlight that it is essential to configure precisely the pool of agents.

Finally, we have stressed that it was useful to have a message system local to agent host to allow synchronization between mobile agents. The use of a message system global to the device seems to be a solution to explore in future experiments.

REFERENCES

[1] P. Braun, and W. Rossak, “From client-server to mobile agents, mobile agent basic concept, mobility model and the Tracy toolkit” Heidelberg university, Germany, Morgan Kaufmann Publishers, pp. 419–441, 2005.

[2] Bluetooth, S. I. G. Specification of the Bluetooth System, version 1.1, 2001, http://www. bluetooth. com, Retrived October 2013.

[3] S. Williams, “IrDA: past, present and future”, Personal Communications, IEEE, vol. 7, no 1, pp. 11-19, 2000.

[4] S. Li, “Professional Jini: from programmer to programmer”, Wrox Press Publishers, August, 2000, pages.1000.

[5] T. Schreiber. Jacobs and C. P. Bean, “Android Binder, Android inter process communication” Ruhr University, thesis Academic, 2011, pages. 154.

[6] Android Platform Official Site, http://www.android.com, Retrived October 2013 .

[7] J. Chen, P. H. Chen and W. L. LI, “Analysis of Android Kernel,” Modern Computer, Vol. 11, 2009.

[8] OSGi Alliance, OSGi service platform, core specification release 4. Draft, July 2005.

[9] C. Lee, D. Nordstedt and S. Helal, “Enabling smart spaces with OSGi”, IEEE Pervasive Computing 2 (3), pp. 89–94, 2003.

[10] K. Myoung, J. Heo, W.H. Kwon and D.S. Kim, “Design and implementation of home network control protocol on OSGi

for home automation", in: Proceedings of the IEEE International Conference on Advanced Communication Technology, vol. 2, pp. 1163–1168, July 2005.

[11] M. Berger, S. Rusitschka, D. Toropov, M. Watzke, and M. Schlichte. "Porting distributed agent-middleware to small mobile devices." In AAMAS Workshop on Ubiquitous Agents on Embedded, Wearable and Mobile Devices .

[12] S. Andrea, M. Guidi, and A. Ricci. "JaCa-Android: an agent-based platform for building smart mobile applications." Languages, Methodologies, and Development Tools for Multi-Agent Systems. Springer Berlin Heidelberg, pp. 95-114,2011.

[13] C. Muldoon, G. M. P. O'Hare, R. W. Collier, and M. J. O'Grady. "Agent factory micro edition" A framework for ambient applications. In Int. Conference on Computational Science (3) , pp. 727–734, 2006.

[14] F. Koch, J.-J. C. Meyer, F. Dignum, and I. Rahwan. "Programming deliberative agents for mobile services" The 3apl-m platform. In PROMAS , pp 222–235, 2005.

[15] F. Mourlin, C. Dumont, "Implementation of a fault-tolerant system for solving cases of numerical computation", ICIBET 2013, International Conference on Information, Business and Education Technology, ISBN: 978-90-78677-56-7.

[16] M. Bernichi, F. Mourlin, "Two level specification for monitoring application", The Fifth International Conference on Systems, Proceedings of ICONS 2010 - Menuires, The Three Valleys, French Alps, France.

[17] L. Richardson, "RESTful Web Services," O'Reilly Media Publishers, Book pages 220, May 2007.

[18] R. Fielding, "Representational state transfer" Architectural Styles and the Design of Netowork-based Software Architecture, pp. 76-85, 2000

[19] B. G.Sullins, and M. B. Whipple, "Manning JMX in action," Manning Publications Co. pages 424 April 2002.

[20] J. Jenkov, "Java Exception Handling," ProWebSoftware Publisher pages 288, March 2001.

Source Mobility Support for Source Specific Multicast in Satellite Networks

Esua Kinyuy Jaff, Prashant Pillai, Yim Fun Hu
School of Engineering, Design and Technology,
University of Bradford
Bradford, United Kingdom
ekjaff@student.bradford.ac.uk, p.pillai@bradford.ac.uk, y.f.hu@bradford.ac.uk

Abstract—**With increasing human mobility and demand for ubiquitous communications, the growth for satellite communications is likely to continue. Recently, IP multicast support over satellites has witness significant increases. Mobility support for multicast receivers as well as sources within a global multi-beam satellite network is very important. Not much research has been done on this area compared to IP multicast mobility support in the Internet. In this paper, we propose a new mechanism to support multicast source mobility for Source-Specific Multicast (SSM) based applications in a multi-beam mesh satellite network with receivers both within the satellite network and in the Internet. In the proposed mechanism, the mobile source remains transparent to the various SSM receivers at all times despite the fact that its IP address keeps on changing as it changes its point of attachment from one satellite gateway (GW) to another. The uniqueness about this proposal is the absence of encapsulation (tunnelling) and triangular routing paths throughout the system and its compliance with DVB-RCS/S2 specifications.**

Keywords-SSM; Mobile Multicast Source; Transparent Satellite; Multi-beam Satellite Interactive Network.

I. INTRODUCTION

IP multicasting is a technology in which a single copy of IP data is sent to a group of interested recipients and the routers (or hosts) in the network replicate the data as required for delivery until a copy reaches all intended downstream group members. In IP multicasting, there may be a many sources sending data to a single multicast group for example: group voice chat. In source-specific multicast (SSM), the group member of such a multicast group, G requests to receive traffic only from one specific source, S. Hence SSM is usually denoted as (S, G) [1].

The handover of a mobile multicast receiver from one point of attachment to another has a local and single impact on that particular receiver only. However, the handover of a mobile source may affect the entire multicast group, thereby making it a critical issue. A mobile multicast source faces two main problems; transparency and reverse path forwarding (RPF).

In SSM, a receiver subscribes to a multicast channel (S, G) [1]. During a handover, as the source moves from one network to another, its IP address will change. When the source uses this new IP address i.e., care-of address (CoA) [2, 3] as source address to send traffic, the multicast router in the foreign network cannot forward the multicast packets until a receiver explicitly subscribes to this new channel (CoA, G). This is known as the transparency problem.

A multicast source-specific tree is associated to source location i.e., the source is always at the root of the source-specific tree. The RPF check compares the packet's source address against the interface upon which the packet is received. During handover, the location of the source will change (and consequently its IP address), thus invalidating the source-specific tree due to the RPF check test. Hence, the RPF problem relates to the fact that the mobile source cannot use its home address in the foreign network as the source address to send packets as this will result in a failure of the RPF mechanism and the ingress filtering [2].

IP multicast over satellites can be used to communicate important service information like the weather conditions, on-going disaster zones and information, route updates, etc. in long haul flights, global maritime vessels and continental trains. Multicasting this information to all the interested parties rather than individually informing them (i.e., unicast) would save a lot of satellite bandwidth resources.

II. PREVIOUS STUDIES ON SSM

A few mobile source support techniques for SSM have been proposed for terrestrial Internet. These are far from being applicable in a satellite scenario. Due to the problems of transparency and RPF, remote subscription [2, 3, 4] cannot be applied to mobile multicast sources for SSM.

Home subscription [2, 3, 4] in terrestrial Internet on the other hand, can support both mobile receivers and sources (including SSM senders) by use of bi-directional tunnelling through home agent without the problems of transparency and RPF. As shown in Figure 1, once the mobile source leaves its home gateway (GW), it must release the resources in its home GW as it acquires new set of resources in the new GW during the GW handover (GWH) [5]. Following the home subscription mechanism, if bi-directional tunnelling between the home GW and the target GW is used to maintain source identity, the mesh communication concept (i.e., a single hop over the satellite) will be lost and could also results in RPF issues. More so, it is practically impossible for the mobile source to make use of resources under its home GW after handover to a new GW [5]. This implies that bi-directional tunnelling through home agent as mobility support technique for a mobile source in a mesh transparent multi-beam satellite scenario is also not suitable.

ISBN: 978-1-61208-313-1

In [6] and [7], the authors using the shared tree approach proposed "Mobility-aware Rendezvous Points" (MRPs), which in effect replace the home agents in their role as mobility anchors. There is at least one MRP per domain. The MRPs rely on triangular routing and tunnelling to fulfil their role as mobility anchors during inter-domain tree setup and also re-introduce rendezvous points, which are not native to SSM routing. The introduction of new entities/messages for example, the MRP, new registration message (of mobile sources to MRPs whenever they move into a new domain), MRP Peer-to-peer Source Active (SA) and keep-alive messages (required to track the source's MRP attachment point changes) during inter-domain multicasting, coupled with the modification of the standard Multicast Forwarding Table (referenced by the two addresses home address and CoA instead of a unique IP address) make this approach very complicated. Also, large number of signalling messages as proposed in this mechanism is not good for satellite networks as they consume the scarce and expensive satellite bandwidth.

Authors in [8] and [9] introduced Tree Morphing and Enhanced Tree Morphing (ETM) respectively, which are routing protocol adaptive to SSM source mobility. The concept of the source tree extension or elongation as the source moves from the previous designated multicast router (pDR) to new designated router (nDR) is not applicable in satellite scenario because the delivery tree rooted at the source in one GW cannot be extended to that same source when it moves to a different GW. This makes the fundamental design concept of these extensions not consistent with the nature of satellite networks.

SSM source handover notification approach proposed by authors in [10] suggested adding a new sub-option in the standard IPv6 destination binding option known as SSM source handover notification. During handover, the source after acquiring new IP address will notify receivers to subscribe to the new channel. The problems here are the large amount of signalling traffic over satellite air interface and the fact that some receivers may be unsynchronized to source handovers, leading to severe packet loss.

A mobile multicast source support for SSM in proxy mobile IPv6 domain has been proposed by authors in [11]. One of the drawbacks here is that there are no mechanisms to supress upstream forwarding to Local Mobility Anchor (LMA) [12] even when there are no receivers. Triangular routing is also a problem here when a mobile receiver and a source, all having different LMAs are attached to the same Mobility Access Gateway (MAG) [12]. In such a situation, the MAG has to forward traffic upstream to the corresponding LMA of the mobile source, which will tunnel the traffic to the corresponding LMA of the mobile receiver which then tunnels the traffic back to the same MAG for delivery to mobile receiver, causing waste of network resources in the whole domain. The fact that in proxy mobile IPv6 domain, the LMA is the topological anchor point for the addresses assigned to mobile nodes within the domain (i.e., packets with those addresses as destination are routed to the LMA), the role of the LMA and MAG does not fit well into a global interactive multi-beam satellite network with many Transparent/Regenerative Satellite Gateways [13], each having different IP addressing space.

This paper proposes a new solution consistent with the DVB–RCS/S2 satellite network specifications that supports SSM source mobility within the satellite network. The provision of solution to the problems of transparency and RPF without creating triangular routing paths and making use of encapsulation (tunnelling) are what make this approach unique. The solution is divided into two parts; support for receivers within the satellite network and those in terrestrial Internet.

III. PROPOSED MULTICAST SOURCE MOBILITY MECHANISM FOR SSM IN SATELLITE NETWORK

A. Network Architecture

Figure 1 shows the network architecture, where a mobile multicast source is present in beam 1 and the receivers are in beams 1, 2, 3 and 6. GW_A1 serves beams 1 and 2, GW_A2 and GW_A3 serve beams 3 and 4, respectively and GW_A4 serves beams 5 and 6.

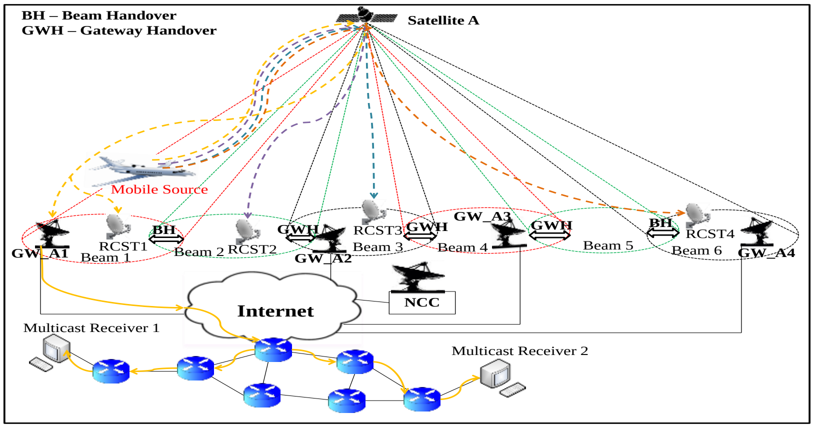


Figure 1. Mobile Source at Home Network (GW_A1)

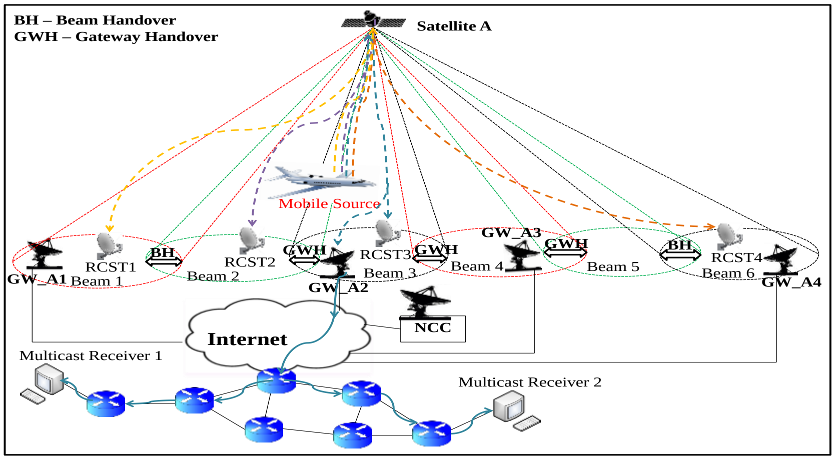


Figure 2. Mobile Source at Foreign Network (GW_A2)

The multicast receivers in the terrestrial network are connected through GW_A1. The mobile source sends out four identical streams of multicast traffic, each for one of the four beams that has interested receivers. This is because the satellite is a transparent one with no IP layer 3 capabilities on-board the satellite to replicate multicast traffic.

Figure 2 shows the mobile source now in beam 3 after successful handover. Here, the terrestrial SSM receivers are now connected through GW_A2 which is the serving gateway to the mobile source in beam 3. This illustrates how the multicast delivery tree to the terrestrial receivers changes whenever the mobile source changes its point of attachment to the satellite network from one satellite gateway to another.

B. *Source Adressing scheme*

According to the DVB specifications each GW has its own IP addressing space different from every other. This proposal leverages on the fact that each mobile Return Channel Satellite Terminal (mRCST), can be reserved a specific IP address under each GW. It is proposed that the IP addresses of the mRCST (in this case, mobile source) under various GWs are made known to the listening RCSTs/GWs as soon as they subscribe to the SSM. This can be made possible by associating the mRCST physical (MAC) address to a specific IP address as illustrated in Table 1.

TABLE I. MOBILE SOURCE IP ADDRESS UNDER EACH GATEWAY

Mobile RCST(Source)	**Mac Address**	**IP Unicast Address**			
		GW1	GW2	GW3	GW4
mRCST1	mac 1	a 11	a 12	a 13	a 14
mRCST2	mac 2	a 21	a 22	a 23	a 24
⋮	⋮	⋮	⋮	⋮	⋮
mRCSTn	mac n	a n1	a n2	a n3	a n4

It is assumed here for simplicity that there are 4 gateways under the control of the Network Control Centre (NCC). However, this can be easily extended. If this allocation can be pre-assigned by the NCC at the time of joining the multicast group, then this would remove the need for the use of tunnelling between GWs. Instead, native forwarding along the source-specific delivery tree throughout the network can be supported.

C. *Support for multicast receivers within the Satellite Network*

It is assumed here that:

- The transparent satellite has on-board multiplexing capability to provide connectivity between different beams provided at the physical layer and mainly responsible for forwarding MF-TDMA carriers or groups of carriers in an uplink beam to different downlinks beams.
- Each mRCST knows all its IP addresses under various GWs as described in the previous section.
- The NCC will act as the the Internet Group Management Protocol (IGMP) querier for the satellite network [13] and contains IP addresses of all mRCSTs under each GW in addition to its normal functionalities
- The NCC enables the establishment of point-to-multipoint connection between source and listening RCSTs/GWs.

The NCC acting as satellite IGMP querier keeps control of the multicast groups and also builds the SSM tree based on on-board connectivity between different beams. Periodically, the NCC sends out the Multicast Map Table (MMT) [14] to all multicast receivers. The MMT which contains the list of IP multicast addresses each associated with a specific Program Identifier (PID) enables listening RCSTs/GWs to receive multicast traffic from groups which they are members of. When the NCC receives an IGMP join report for SSM, it checks the source-list and if some sources are identified as mRCSTs, it will immediately respond with a unicast message to the RCST/GW which requested listening by stating that the multicast source is a mobile source, as well as giving out the source IP addresses under each GW (e.g., Mobile source; IP addresses: a11, a12, a13, a14). This message prepares the receiver to expect multicast traffic from any of these IP addresses knowing that it is coming from the same source, thus, solving the problem of transparency. For receivers on LAN behind RCSTs [14], the RCST acting as an IGMP Router and Querier on its user interface (i.e., interface towards the internal LAN) and an IGMP Host on the satellite interface, after receiving the mobile source details, will take up the role of notifying any interested user terminal in its LAN of the multiple IP addresses of the mobile source. The listening RCSTs' user interface delivers the traffic according to channel subscription (S, G) to user terminals.

D. *Support for multicast Receivers on the Internet*

It is proposed that all GWs in the Interactive Satellite Network (ISN) should be equipped with a new Multicast Mobility Management Unit (M3U) that is responsible for control plane signalling to provide support for mobility for multicast sources. This new M3U entity contains the following:

- A database of information regarding all mRCSTs in the entire ISN, each identified by its physical (MAC) address and the fixed IP addresses it has under each GW as shown in Table 1.
- A message chamber which can generate three new types of messages shown in Table II.

TABLE II. PROPOSED NEW MESSAGES

Message Name	Interface state message	Target GW message	Channel update message
Type	Unicast	Unicast	Multicast
Source	Serving GW	Target GW	Serving GW
Destination	Neighbouring GWs	Serving GW	All SSM Receivers in the Internet
Content	SSM reception interface state i.e., Multicast addresses, filtering mode and source list.	IP address of the Target GW	IP address of mobile source in target GW. Instruction to update channel subscription to new mobile source IP address
Purpose	To identify the mobile source in preparation for GW handover	To notify serving GW which neighbouring GW will be the target GW	For the Internet receivers to start building the new delivery tree to the target GW

When a GW receives the first IGMP join report for SSM, a service interface (socket; interface; multicast address; filter-mode; source-list) [15] is created against the interface that received the join report. While forwarding this join report to the NCC, the M3U as shown in Figure 3, checks the multicast source-list in the report against the data base containing the list of all mRCSTs. If some sources on the source-list are identified as mRCSTs, then the M3U of the serving GW will send the new proposed Interface State Message (ISM) to all neighbouring GWs. The neighbouring GWs extract the mobile sources (mRCSTs) from the source list after consulting the database in their M3U.

When the mobile source moves and a handover procedure is initiated by the NCC by sending the SNMP Set-Request message to the target GW, the target GW issues the new proposed Target GW Message (TGM) to the serving GW. The TGM enables the serving GW to identify which of its neighbouring GWs will be the target GW for the mobile source.With the knowledge of the target GW identity and upon consulting its database, the serving GW will then issue the Channel Update Message (CUM) to all SSM receivers in Internet/LAN.

It should be noted here that only the serving GW can reach all SSM receivers in the Internet since it is located at the root of the SSM delivery tree to the Internet. Upon reception of CUM by SSM receivers in the Internet, a new SSM delivery tree construction to the target GW is triggered as shown in Figure 2. Target GW issues IGMP join report to NCC as soon as it gets the updated channel subscription request (PIM Join) from Internet and at the same time, its M3U will issue ISM to all GWs neighbouring target GW in preparation for the next GWH. The target GW now becomes part of the mesh receivers within the satellite network described in section IIIC above as it assumes the responsibility of serving receivers in the Internet. This therefore makes the IP address changes during a GW handover transparent to all SSM recipients. Eventually, the old tree to old GW will be torn down as it becomes inactive (no traffic).

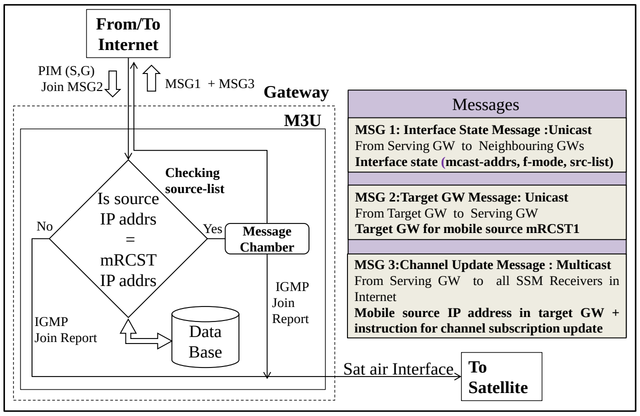

From/To Internet
PIM (S,G) Join MSG2
MSG1 + MSG3
Gateway
M3U
Checking source-list
Is source IP addrs = mRCST IP addrs
No
Yes
Message Chamber
IGMP Join Report
IGMP Join Report
Data Base
Sat air Interface
To Satellite
Messages
MSG 1: Interface State Message :Unicast
From Serving GW to Neighbouring GWs
Interface state (mcast-addrs, f-mode, src-list)
MSG 2:Target GW Message: Unicast
From Target GW to Serving GW
Target GW for mobile source mRCST1
MSG 3:Channel Update Message : Multicast
From Serving GW to all SSM Receivers in Internet
Mobile source IP address in target GW + instruction for channel subscription update

Figure 3. Multicast Mobility Management Unit (M3U)

E. *Message sequence for multicast source mobility*

Figure 4 shows the signalling sequence during GW handover.

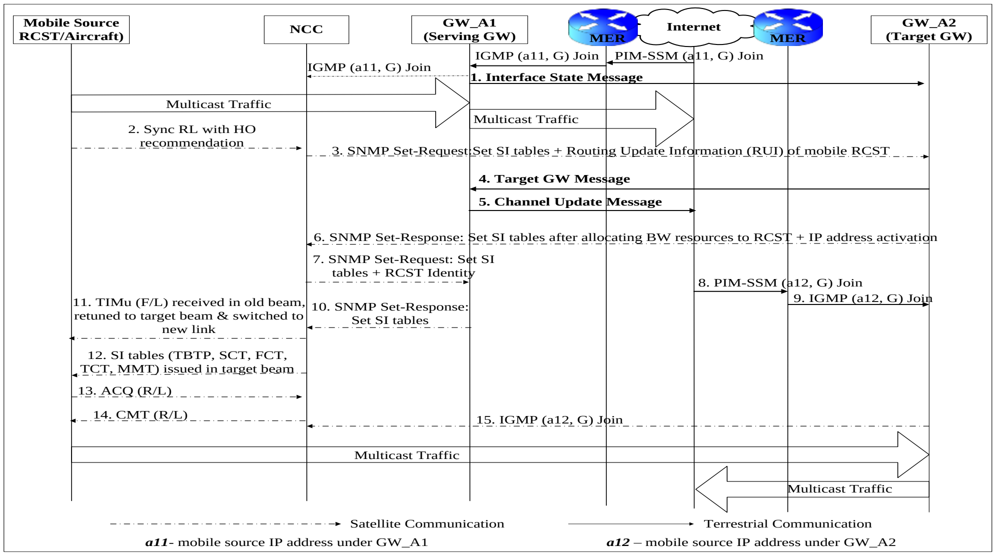

Mobile Source RCST/Aircraft)
NCC
GW_A1 (Serving GW)
MER
Internet
MER
GW_A2 (Target GW)
IGMP (a11, G) Join
PIM-SSM (a11, G) Join
IGMP (a11, G) Join
1. Interface State Message
Multicast Traffic
Multicast Traffic
2. Sync RL with HO recommendation
3. SNMP Set-Request:Set SI tables + Routing Update Information (RUI) of mobile RCST
4. Target GW Message
5. Channel Update Message
6. SNMP Set-Response: Set SI tables after allocating BW resources to RCST + IP address activation
7. SNMP Set-Request: Set SI tables + RCST Identity
8. PIM-SSM (a12, G) Join
9. IGMP (a12, G) Join
11. TIMu (F/L) received in old beam, retuned to target beam & switched to new link
10. SNMP Set-Response: Set SI tables
12. SI tables (TBTP, SCT, FCT, TCT, MMT) issued in target beam
13. ACQ (R/L)
14. CMT (R/L)
15. IGMP (a12, G) Join
Multicast Traffic
Multicast Traffic
Satellite Communication
Terrestrial Communication
a11- mobile source IP address under GW_A1
a12 – mobile source IP address under GW_A2

Figure 4. Signalling sequence at gateway handover

This signalling sequence contains the proposed new messages integrated into the standard GW handover signalling sequence given in [5]. Figure 4 is the detailed illustration of what is described in section IIID. From the signalling sequences in Figure 4, the sizes of the signalling messages determine the total time required to complete a GWH.

IV. GW HANDOVER PERFORMANCE EVALUATION

The messages sizes used in table III are derived from [16, 17] and [18]. It is assumed that the routing update information table (RUI) contains at least 100 bytes of routing data.

TABLE III. GW HANDOVER SIGNALLING MESSAGES SIZES

Messages	Packet Length (in Bytes)
Message 1 (containing at least 3 sources)	45
Synchronisation (SYNC) Burst	16
SNMP Set-Request: set SI tables + RUI	736
Message 2	28
Message 3	28
SNMP Set-Response: set SI tables	636
SNMP Set-Request: set SI tables + RCST Identity	640
PIM-SM Join	64
SNMP Set-Response: set SI tables + RCST Identity	640
TIM (Terminal Information Message)	35
SI Tables (TBTP, SCT, FCT, TCT, MMT)	152
ACQ (Acquisition Burst)	12
CMT (Correction Message Table)	30
IGMP Join	64

The time taken to transmit a single message between two relevant network entities over any given link under ideal conditions i.e., lossless conditions is given by (1) and the time required to send a message during handover relevant signalling entities under lossy conditions is given by (2).

$$T_{lossless} = T_{trans} + T_{prop} + T_{proc} \qquad (1)$$

$$T_{total} = T_{lossless} + \left(T_{lossless} + T_w\right) \times \left[\frac{q}{1-q}\right] + T_{INT} \qquad (2)$$

Where,

- T_{trans} = transmission delay = message size ÷ link bit rate.
- T_{Proc} = average processing time at any node. This is assumed to be 5 ms [18] for all nodes.
- T_{prop} = propagation delay due to the communication link.

$$T_{prop} = \begin{cases} T_{prop\_wireless} \\ T_{prop\_wired} \\ T_{INT} \end{cases} \qquad (3)$$

-$T_{prop\_wireless}$ = propagation delay due to wireless link
-T_{prop_wired} = propagation delay due to wired link
- T_{INT} = propagation in the Internet. Since it is impossible to know the rout taken by the packet in the Internet with certainty, it is assumed to be 8 ms as suggested in [18].

- q = probability of a failure transmission over satellite.

The data rate in the satellite link is assumed to be 144Kbps [18] and the gateways are assumed to be using 100 Base-T Ethernet supporting a data rate of 100Mbps. The propagation speed in LAN Ethernet is assumed to be 2/3 the speed of light i.e., 2 x108m/s [18]. The distance between the gateway and the multicast edge router MER is assumed to be 4m.

Figure 5 shows that the total time required to complete GWH and to reconstruct the SSM delivery tree to target GW increases with increasing probabilities of failure. It should be noted that the new SSM tree construction starts when any receiver in the Internet issues the first IGMP/PIM –SSM join message to the target GW upon reception of channel subscription update message. It can be deduced from Figure 5 that for probabilities of failure 0% - 10%, the average time taken to complete:

- GWH with only mesh multicast support is 2.83 seconds
- GWH with mesh and terrestrial (internet) multicast support is 3.22 seconds
- New SSM delivery tree construction to target GW and the first IGMP reaching the NCC is 1.85 seconds.

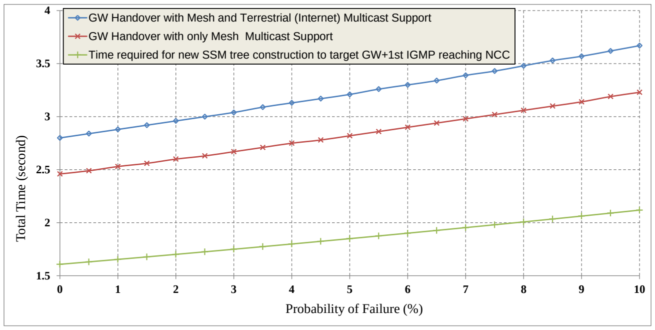


Figure 5. GW handover time delay at probabilities of failure 0% – 10%

Note should be taken of the fact in this proposal, no additional time delay (compared to the standard in [5]) is incurred for source mobility support for mesh receivers during GWH and that the actual time between switching links from serving GW to target GW is very small compared to the total GWH time shown above

V. Conclusion

IP multicast based applications are very important for satellite networks in order to share vital information between various receivers without the wastage of expensive satellite resources. Multicast sources that may move from one point of attachment to another will result in the breakage of the multicast tree. While some solutions to support multicast source mobility have been proposed for the internet, it was seen that these are not very suitable in a satellite network. This paper proposes a suitable solution for multicast source mobility in a multi-beam satellite network. It presents the network architecture and the proposed address management scheme. A new Multicast Mobility Management Unit (M3U) has been proposed within the GW along with three new control messages that provide support for mobility for multicast sources. The proposed solution has made it possible to solve the mobile multicast source transparency and RPF problems. Time delay analysis was carried out in order to evaluate the performance during the gateway handover. The results obtained from GWH performance evaluation above show that source mobility support for SSM has very little or no impact on handover latency on receivers within the satellite network.

References

[1] H. Holbrook, B. Cain, and B. Haberman, "Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source-Specific Multicast," RFC 4604, IETF, August 2006.

[2] I. Romdhani, M. Kellil, H. Lach, A. Bouabdallah, and H. Bettahar, "IP Mobile Multicast: Challenges and Solutions," IEEE Communications Survey & Tutorials, vol.6, First Quarter 2004, pp. 18 - 41.

[3] T. Nguyen, "IP Mobile Multicast: Problems and Solutions," Ph.D. Dissertation, EURECOM, France, March 2011.

[4] G. Xylomenos and G.C Polyzos, "IP multicast for mobile hosts," IEEE Communications Magazine, vol. 35, January 1997, pp. 54 – 58.

[5] ETSI TR 102 768, "Digital Video Broadcasting (DVB); Interaction channel for Satellite Distribution Systems; Guidelines for the use of EN 301 790 in mobile scenarios," vol.1.1.1, April 2009.

[6] I. Romdhani, H. Bettahar, and A. Bouabdallah, "Transparent handover for mobile multicast sources," in proceedings of the International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies, April 2006, pp 145.

[7] B. Fenner, M. Handley, H. Holbrook, and I. Kouvelas, "Protocol independent multicast—sparse mode (PIM-SM): protocol specification (Revised)," RFC 4601, IETF, August 2006.

[8] T. C. Schmidt and M. Wählisch, "Extending SSM to MIPv6—problems, solutions and improvements," in selected papers from TERENA Networking Conference, Computational Methods in Science and Technology, Poznań, May 2005, pp. 147 – 152.

[9] T. C. Schmidt, M. Wählisch, and M. Wodarz, "Fast adaptive routing supporting mobile senders in Source Specific Multicast," Springer Telecommunication Systems, Vol 43, February 2010, pp. 95 – 108.

[10] C. Jelger andT. Noel, "Supporting Mobile SSM Sources for IPv6," in IEEE Global Telecommunications Conference (GLOBECOM), vol. 2, November 2002, pp. 1693 - 1697.

[11] T C. Schmidt, S. Gao, H. Zhang, and M. Waehlisch, "Mobile Multicast Sender Support in Proxy Mobile IPv6 (PMIPv6) Domains. draft-ietf-multimob-pmipv6-source-03," Internet-Draft, MULTIMOB Group, IETF February , 2013.

[12] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, and B. Patil, "Proxy Mobile IPv6,". RFC 5213, IETF, August 2008.

[13] ETSI TR 102 603, "Satellite Earth Stations and Systems (SES); Broadband Satellite Multimedia (BSM); Connection Control Protocol (C2P) for DVB-RCS; Background Information," vol. 1.1.1, January 2009.

[14] ETSI TS 102 429-2, "Satellite Earth Stations and Systems (SES); Broadband Satellite Multimedia (BSM); Regenerative Satellite Mesh - B (RSM-B); DVB-S/DVB-RCS family for regenerative satellites; Part 2: Satellite Link Control layer," vol. 1.1.1, October 2006.

[15] B. Cain, S. Deering, I. Kouvelas, B. Fenner, and A. Thyagarajan, "Internet Group Management Protocol, Version 3," RFC 3376, IETF, October 2002.

[16] G. Maral, M. Bousquet, and Z. Sun, "Satellite Communications Systems: Systems, Techniques and Technology", John Wiley & Sons Ltd., 2009.

[17] ETSI EN 301 790, "Digital Video Broadcasting (DVB); Interaction channel for satellite distribution systems," vol. 1.5.1, May 2009.

[18] P.M.L. Chan, R.E. Sheriff, Y.F. Hu, P. Conforto, and C. Tocci, "Design and evaluation of signaling protocols for mobility management in an integrated IP environment," Computer Networks: The International Journal of Computer and Telecommunications Networking , Vol 38, March 2002, pp. 517 - 530.

mTADA: Mobile Tracking and Advanced Driver Assistance

Alexiei Dingli and Christopher Bartolo

Department of Intelligent Computer Systems

Faculty of ICT, University of Malta

Msida, Malta

alexiei.dingli@um.edu.mt, chris@chrisbartolo.com

***Abstract*—Safe driving requires knowledge of the surrounding traffic environment. Such knowledge will enable the driver to predict possible dangerous situations and events. Driver distraction and reckless driving are two cases, which reduce the ability of a driver to perceive these situations. This is the reason why Advanced Driver Assistance Systems (ADASs) are being introduced. In this paper, we will tackle how mobile devices may help out in ADASs. Many companies, such as BMW and Audi have been developing their own ADASs, but offer them as optional accessories at an additional cost, being available only with new vehicles. The aim of the system we are developing is to help out in decreasing the amount of motor vehicle accidents by developing a low cost solution, which can be easily downloaded to and installed on a mobile device, and which may be used on any vehicle. The final results prove that the system was quite successful during tests, with many people claiming that they would use such a system on a regular basis.**

Keywords-active driver assistance, driver behaviour, motor accidents, driver safety

I. Introduction

In this section, we will introduce the subject matter of the paper. We will first identify the problem, and then we will continue to explain the major causes of motor traffic accidents.

A. The Challenge

Road traffic accidents have become one of the leading causes of death. This costs countries billions of dollars a year for damages, insurances, and health bills. The most worrying, though, is the fact that the increase of these accidents has brought the cause up to the same ranks as HIV/AIDS, Malaria and Tuberculosis, and most communities are still practically ignoring it. Lord Robertson from Make Roads Safe argues that every six seconds we have another reason to work with campaigns such as Make Roads Safe as an international community. Make Roads Safe present statistics claiming that 1.2 million people are killed, with a minimum of 50 million people left injured, every year. They predict that by 2015, “road crashes will be the leading cause of disability for children aged four and above in developing countries” [1]. L.G. Norman argues that accidents are or will become the number one cause for deaths, mainly in highly industrialized nations where roads are crowded with vehicles. He continues to argue that road traffic accidents can be considered as the new epidemic. The paper by L.G. Norman was published in 1962, which only goes to show that the situation of road traffic accidents has not been improved as dramatically as was needed and still has a lot of work to be done [2].

Many drivers are over-optimistic, over-confident, and make above-average biases, leading to excessive risk taking, a delay in the uptake information, and social norms of bad driving, which lead to a high number of traffic accidents. These accidents also lead to moral hazards for damages, which are usually covered by insurances, as many damages are not paid for by the drivers, which make and take the risks. Knowing this, drivers are willing to take more risks. We can easily see that many accidents can be easily avoided. Most accidents are caused by the victims themselves, caused by human errors, such as “imperfect perception, insufficient attention, and inadequate information processing” [3].

B. Major Motor Accident Causes

Cause 1: Alcohol; Alcohol is one of the most well-known and most advertised causes of accidents. There have been many campaigns launched to tackle this cause and increase the awareness of such health issues. In fact, the amount of accidents caused by drunk driving has been declining in many countries throughout the past few years [4].

Cause 2: Fatigue; We all know that we need to sleep seven to nine hours a day to optimize our performance, and to be able to concentrate and work. It is not a matter of choice. The longer we remain awake, the more we feel the need to sleep and the harder it gets to resist the urge to do so. By not sleeping the right number of hours, we are becoming more prone to extreme short term sleepiness, which could easily lead to chronic sleepiness. Sleepiness reduces the person’s reaction time, vigilance, alertness, and concentration, hence resulting in impairment of attention-based activities. These activities include driving. Sleepiness reduces the speed at which information is processed and the quality of all decision-making, badly affecting any driving activity. A study was held by the Sleep Research Centre in the United Kingdom on how drivers are affected by fatigue. The study reported a finding which indicates that 20% of all accidents are caused by driver fatigue on long, monotonous roads [5].

Cause 3: Distraction; Driver distraction is when a driver pays attention to other activities while driving. It is another major problem, which is also well-known to be one of the biggest causes of road traffic accidents. Most widely used in campaigns is the risk of driving while using mobile phones and in fact fines issued by police due to the use of mobile phones while driving have been increasing rapidly during the past few years. However, research in the area has proved that road traffic accidents related to driver distraction are actually being caused by other methods and other forms of multitasking. Distraction impairs the driver's decision-making and observation skills, making him take worse decisions about how to control his vehicle's and surrounding vehicles' safety [6].

Cause 4: Behavior; The driving behavior of drivers reflects how the driver perceives situations and his decision-making skills when encountering different situations. Irresponsible driving behaviors have been given the term 'road rage' to describe them. Road rage is becoming increasingly popular especially with the young and middle aged population because of the stress, anxiety, anger, antagonism, and fear experiences on today's roads and traffic situations. Surprisingly, most problems which result from bad driving behavior are actually caused by the belief that the law cannot or will not do anything and would not take action about any fleeting transgressions [7].

II. Technologies and Driver Assistance

In this section, we the focus will be around around what has already been published related to driving assistance systems and the technologies used.

A. *Computer Vision*

One of the most broad and vast research areas in computing is image processing. It involves not just filtering, but also compression and image enhancement. Computer vision is where image processing is used and implemented, and one of its main tasks is to understand what the image is actually depicting. This is handled by making use of complex algorithms to analyse the content based on recognition. All functionalities provided by computer vision such as reconstruction of 3D scenes, motion capturing and object recognition. The field of computer vision is rapidly growing, with microprocessors being improved and becoming more powerful [8]. The results of these improvements came with a number of computer vision libraries being made available, some of which are even open source and free to use, while others are commercial products. Some of the libraries we investigated were libCVD, openCV, Matrox Imaging Library and NI Vision [9].

B. *Advanced Driver Assistance Systems*

These systems have recently become very popular, with many companies researching and developing their own versions. Their aim is to decrease the amount of accidents and decrease the gravity of consequences caused by accidents, while also making the driving experience easier and more efficient. These systems are intelligent systems, and support the driver in performing one or more elements involved in the driving task [9]. A number of surveys have been conducted regarding ADASs, and the results conclude that 40% of traffic accidents can be prevented with their use. However, even though ADASs can prevent many accidents, injuries and fatalities, they are still rarely used. This is believed to be true because many people have yet to understand their usefulness [10].

As previously discussed, many manufacturers have been and are still developing and implementing their own ADASs. These companies include Mobileye, Nvidia, Continental, BMW, VolksWagen, and Audi. Many of the features which they have implemented include night vision, lane departure warning, curve and speed limit information, collision warning, adaptive cruise control, local hazard warning, lane change assistant, obstacle and collision warning and avoidance [11].

III. Tackling the Main Causes

A. *Cause 1: Alcohol*

The current idea to reduce the problem of driving under the influence of alcohol is to "separate drinking drivers from their vehicles" [12]. This can be done by installing an ignition interlock device on the vehicle, which is connected to the starting system. The setup is connected to a breathalyzer, which enables or disables the ignition interlock device. When the driver needs to drive, he is required to blow into the breathalyzer tube [12]. Smart Start Inc. is a company based on Irving, TX, and has developed ignition interlock devices for vehicles and also for homes [13].

B. *Cause 2: Fatigue*

VolksWagen and Smart Eye AB have developed their own solutions to tackle this cause. VolksWagen group have developed an attention control system, which includes a camera in the cockpit that monitors the driver's eyelids. When the system finds any signs that the driver may be tired, it makes a noise and requests the driver to take a break [11]. SmartEye AB developed the SmartEye Pro 3.0, which is a system that estimates head pose by using methods based on tracking of facial features and a three dimensional head model. However, they do not manufacture any algorithms to monitor the drowsiness [14].

C. *Cause 3: Distraction*

A number of companies have introduced various systems and ideas to solve this cause. Many of these systems are not intelligent, providing feedback if, for example, the driver is intentionally changing lanes. Some of the best solutions are those presented by Mobileye. Their main solution is a camera-based safety solution, which helps out in accident prevention and mitigation. It notifies the driver on lane departure with visual and audio warnings and measures the distance between the tires to the lane markings on both sides of the vehicle [15].

D. *Cause 4: Behavior*

Companies developing driver behavior monitoring include Teltonika and Cellocator. These two companies offer vehicle tracking hardware with in-built algorithms based on accelerometer values. They take into consideration over-speeding, acceleration and braking, while also detecting, processing, logging and reporting a wide set of events concerned with hazardous driving behavior [16][17].

IV. Aims and Objectives

ADASs aim to prevent motor accidents mainly caused by distractions, drowsiness and behavior. In this work, our main goal is to provide a low-cost driver assistance system without the need of installing any additional hardware and which works on any vehicle. In order to produce such a solution, we have investigated a number of existing solutions.

We aim to provide an easy to use, and an easy to understand interface, which notifies the driver of dangerous events based on priorities (sorted according to their importance) without distracting the driver. The notifications we aim to provide the driver with are related to over-speeding, driver distraction and drowsiness, lane departure, oncoming obstacles, and suggested speeds for oncoming curves.

V. Design

A system has been developed and released. In this section, we will go through the system's design and the chosen platform.

A. *Mobile Platforms*

Mobile devices have improved radically, especially with the progress of the internet on these devices. This caused the usage of smartphones and their services to "become more and more popular" [18]. Some of these services include Location Based Services and transport information services. "Mobile Services make life easier, simpler and more effective" [18]. There are a number of Mobile Platforms available for mobile devices, including Android, Windows, iOS, WebOS, Symbian, and RIM [18].

Developing the system to work on a commonly used mobile platform will satisfy one of the aims we set out to achieve – we have tried to produce an easy to use application, which can be installed and made use of by anyone of any age.

The system was developed on the Android platform. It is the mobile platform, which is on the rise, has the highest market share and is one of the easiest to develop on since applications can be developed using JAVA.

B. *ADAS to Android mobile application*

A lot of time was spent researching existing driver assistance systems and solutions, which are available to the general public. All the systems found are powered by quite powerful hardware, specifically designed and built for the use of ADASs. The systems also have plenty of sensors installed around the vehicle, such as RADAR infrared sensors, and cameras with night vision enhancement technologies. For our implementation, we made it a point not to use any third party hardware peripherals, and use only the built-in hardware sensors provided by the device, which made it harder to implement.

Analyzing the causes and how they have been previously tackled, all of them require additional hardware, such as breathalyzers. In our implementation, we developed workarounds, which make use of estimates. This will satisfy one of the aims, which states that we will try to provide a low cost solution with no need to purchase and install any external devices. However, for causes such as drunk driving, workarounds could not be developed since there is no actual control over the vehicle.

1) *Tackling Cause 1: alcohol;*

As previously mentioned, developing functionality to tackle this cause is close to impossible without third party external peripherals. One of the ideas which came to mind was to implement an accelerometer sensor listener, which analyses user activity patterns and checks the driving behaviour. This however would turn out to be very inaccurate, and can only be used to warn the driver not to drive. Warning a person under the influence of alcohol not to drive would prove to be close to useless. This is because as previously explained, alcohol encourages a competitive behaviour, which will only encourage the driver to be dangerously daring.

2) *Tackling Cause 2: fatigue; and cause 4: Behavior*

The methods to tackle driver fatigue and driver behavior are very similar. The idea is to analyse the user activity patterns to identify the status of the user. The data is collected from the device's in-built accelerometer sensor, and then filtered to achieve the data needed. The current implementation identifies the status of the user, whether walking, driving, or standby. It also identifies any sharp 'hits' exerted on the device, such as pot holes in the road, harsh turns, harsh braking and harsh acceleration. However, the implementation assumes that any sharp 'hits' are only harsh braking or harsh acceleration, and does not understand the difference between the others (harsh turning, etc.). To tackle fatigue, the current implementation constantly provides feedback to the driver by both audible tones, and by speaking out instructions. This keeps the driver attentive and pesters him if he is driving too fast and therefore recklessly. The implementation to continually give feedback to the user will also help to satisfy the third aim, which states that we will provide the user with warnings to reduce his possibility of being in a traffic motor accident.

Other parts of the implementation include a curve detection and notification function, which analyses the road the vehicle is currently on, and using digital maps, extracts and calculates the angles of the rest of the road in front. The algorithm then takes the user's speed from the GPS sensor, and compares the angles to the suggested speeds. These implementations will help reduce motor accidents, which occur due to over speeding in curves by notifying the user earlier to give him time to take preventive action.

3) Tackling cause 3: Distraction; Tackling this cause is a combination of all the functionalities and methods used to tackle the other three causes. We tackled driving distraction with the idea that we need to constantly pester the driver to keep his attention and focus on the road ahead. The alerts and warnings used in the previously described functionalities already do a big part of the job. The additional methods to tackle this cause have not been fully implemented. However, the essential, most difficult parts have been implemented for proof of concept.

The system makes use of the camera to analyse the lane markings and oncoming obstacles, and tries to estimate (since only one camera is available) the distance between the host vehicle and the obstacle linearly in front. The current implementation can also detect obstacles on neighbouring lanes, which then can be used to provide feedback to the driver if there is a possibility that an obstacle is beside the host vehicle while changing lanes. All the functionalities combined help reduce the risk of motor accidents by alerting the driver and catching his attention if he distracted.

VI. IMPLEMENTATION

The system has been developed in a modular fashion, where each feature is developed in its own separate module. The main modules are the user interface, the activity analyser, the lane processor, the obstacle processor, the directions and curve detector, the GPS location module and the call catching module. It is important to note that our aim was not to develop all features which are available in other existing ADASs, but to give an idea of what can be achieved with a mobile device and how it can help out in this area.

The GPS sensor will be used for geolocation purposes, to record location, speed and acceleration. The 3G connections provided by the device will be used to connect to the server, and the accelerometer will be used to analyse the driver's activities.

1) The service and the main thread

The system has been divided into two main parts: the main thread and the service. The main thread handles the Graphical User Interface (GUI) of the application: the camera view, the car information views, the destination selection and map view, and other small, low memory intensive threads, lane processing and obstacle processing. The service is a background thread, which controls other small background threads. It handles the text to speech, the activity analyser, the GPS location listener, curve detection, turn by turn navigation, location tracking, warning logging and notifications, incoming call blocker and the SMS sender.

2) The activity analyser

This feature analyses the device's accelerometer and GPS information to determine whether the user is walking, driving, or standby. It is mainly used in conjunction with the incoming call block. If the user is in 'standby' or 'walking' mode, the incoming call block is switched off. If the user is in 'driving', the incoming call block is enabled. In Figure 1, we can see the process for analyzing the user's activity.

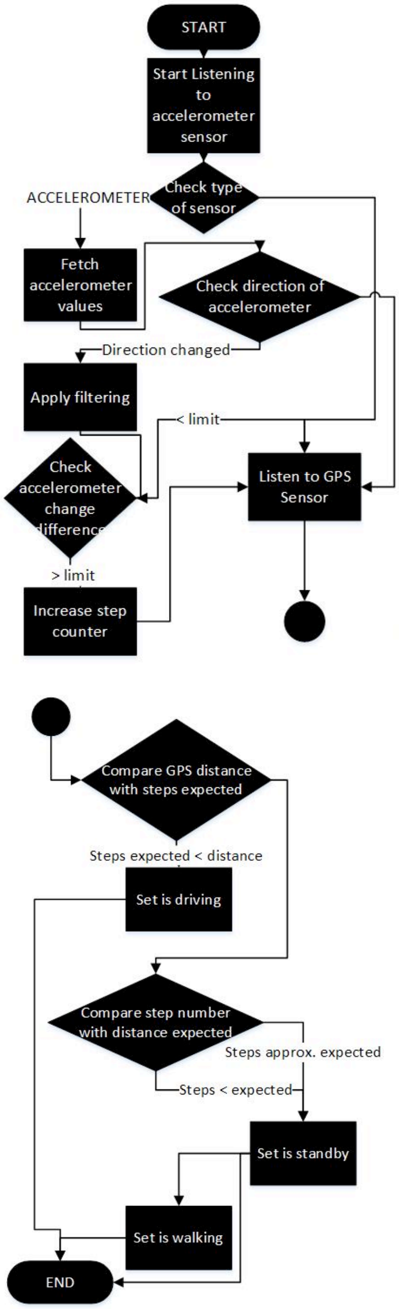

START
Start Listening to accelerometer sensor
Check type of sensor
ACCELEROMETER
Fetch accelerometer values
Check direction of accelerometer
Direction changed
Apply filtering
< limit
Check accelerometer change difference
> limit
Listen to GPS Sensor
Increase step counter
Compare GPS distance with steps expected
Steps expected < distance
Set is driving
Compare step number with distance expected
Steps approx. expected
Steps < expected
Set is standby
Set is walking
END

Figure 1: The Activity Analyzer process

3) *Lane processing*

The lane processing algorithm took a lot of time to figure out; much longer than expected. Many implementations were tried and tested, until a good output was achieved. The main problems were caused because of the faded out lane markings in Malta. Also, many roads do not have lane markings. The result achieved by the lane processing algorithm can be seen in Figure 2.

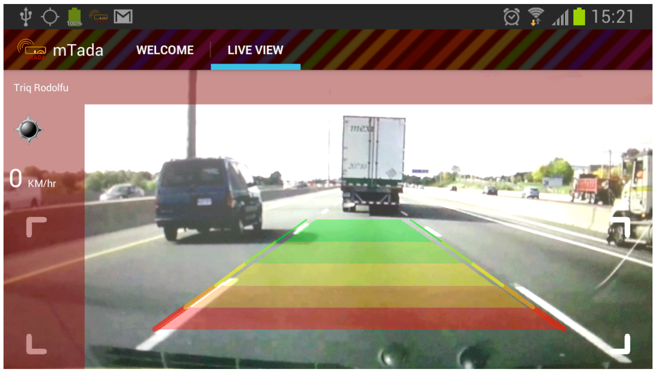


Figure 2: The result achieved from the lane processing algorithm

4) *Obstacle processing*

The obstacle processing algorithm which has been designed and implemented for the paper has to be improved and worked on. The algorithm implemented is one which has been adapted from the algorithm proposed by Shane Tuohy in his dissertation paper: Real Time Distance Determination for an Automobile Environment using Inverse Perspective Mapping in OpenCV.

5) *Directions and Curve detection*

For this feature, we needed to set up a special server with map data. We set up a CentOs 6.0 64bit server and installed a number of applications such as Mapnik, osm2pgsql and Routino. These applications are used for mapping functionalities such as routing. We downloaded the map data from an open source, free to use server: Open Street Maps. We then implemented some methods using PHP and postgresql queries to retrieve road data, such as road names and road points, which make up the road. The road points are sent to the application installed on the device, which in turn analyses them and calculates the difference between each angle. It then issues both visual warnings and sound warnings to the driver depending on the speed of the vehicle and the strength of the oncoming curve. An example of a warning given by the curve detection can be seen in Figure 3.

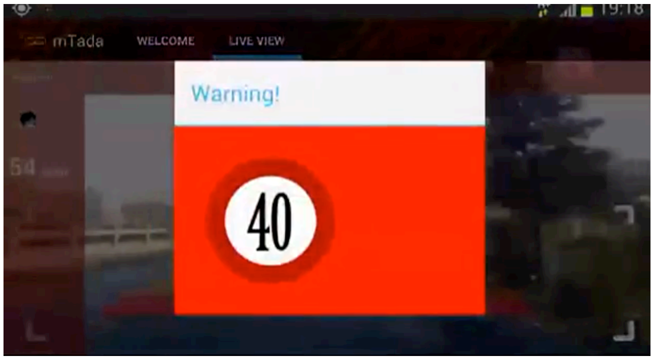


Figure 3: A curve detection speed warning

6) *Incoming calls and SMSs*

The main purpose of this feature is to reduce driver distraction. It listens for any incoming calls, and can be set to either answer the call and send it to loud speaker, or block any incoming calls, and send an SMS automatically to the caller, with a message saying that the other user is driving.

VII. RESULTS AND EVALUATION

The system was made available on Google Play, free of charge. We created a website which enabled users to register and receive a username and password. As soon as we released the system for private testing, we sent out an email to all registered users to inform them that they could start testing it. We had some limitations, such as the fact that user must be located in the Maltese Islands to be able to test it out. After installing the application on the user's mobile, the mobile had to be installed on the dashboard of a vehicle, with the camera facing towards the road in front.

We evaluated the system in two ways. As explained above, we allowed users to download it and test it on their own. For the second method, we chose a number of individuals from different age groups and went on test drives with each one of them individually.

1) *Private testing*

The website for registering and downloading the application was launched on www.m-tada.com. After given some time to evaluate, each user was sent a short questionnaire. The questionnaire was filled and submitted by twenty-one people, 48% being between the ages of eighteen and twenty, 33% being between the ages of twenty-one and twenty-five, 5% being from the age group of twenty-seven to thirty-five and the last 14% being older than thirty-five. Six of the applicants have been in a motor traffic accidents, four admitting that it was their fault, mainly caused by driver distraction and driver drowsiness. One of the applicants also said that the accident was partially caused because of driver behavior. Eleven people said that they never considered making use of ADASs, while four never felt the need, and the other six may have considered them. Also, four of the applicants did not know that many modern vehicles come with driver assistance options, while two never considered it.

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-313-1

With the questions regarding the application, three of the applicants did not have a compatible device and therefore could not make use of it. The others who managed to make use of the system, found the notifications and warnings it provided very accurate, with 15 claiming that it helped them improve their driving behavior, and another three saying that they do not know if it helped. The last question in the questionnaire asked whether the applicants would buy such an application. Fourteen of the applicants said they would, while three said no. One of the applicants comments that he would prefer the basic application to be made free of charge, with only additional features being put up for optional purchase.

2) Individual test drives

We went on test drives with 7 people of different age groups. We carefully analyzed their reactions to the feedback provided by the system to determine their compliance and acceptance of it. The device we used was a Samsung Galaxy S3 with Android version 4.1.2. It had an average frame rate of 10 frames per second.

Age group 1: 18-20; We tested 2 different people. Both were very compliant and accepting of the system, and drove very cautiously. They did their best to avoid being provided any feedback from the system. One of the participants said that he complied to the system as it made him feel safer. He said that since he is still a beginner, he is still used to following instructions given by the instructor, and he felt that the system was replacing him. He also claimed that the curve detection and warning system helped him travel along roads, which were new to him.

Age group 2: 20-30; The person from this age group drove very cautiously and slowly. He was given close to no feedback at all. When asked how he felt driving with the system, he said that he obeyed it because he did not want to be pestered by it, and also wanted to make the best of it. He used it to aid him while driving. He also claimed that he thought that the visual warnings defeated the scope of auditory warnings as it would take your eyes off the road unnecessarily.

Age group 3: 30-40; The person from the third age group ignored most of the warnings provided to him by the system, and only complied to it when it seemed to be the only choice he had. He gave us the impression that he did not comply because of pride and ego. We later asked him why he did not comply, and he simply answered that he wasn't interested in being given any feedback. He also said that he wouldn't mind having such a system in his vehicle, as long as he could choose the type of warnings and what warnings he would be given.

Age group 4: 40 -50; For this age group we went on 2 different test drives, a female and a male. The female claimed that she finds driver assistance system useful, and makes use of them regularly. She said that they help her in heavy traffic roads, and high speed roads. When using the system, she said that the warnings were very helpful and quite accurate. The male did not want any feedback given to him. Out of pride, he went to the extent of doing exactly the opposite of what the system suggested him to do. He constantly criticized the system and said that the only feedback he wants is to know if there is a speed camera up ahead. He also claimed that he also ignores feedback from passengers as he is a very good driver because of his experience he gained over the years.

3) Results

At first glance, we can say that the system, although still requiring a lot of work, was very effective with the majority of the users. The system provided the users with accurate warnings. Most of the warnings were regarding high speeds when approaching curves in the road. The system pestered the drivers, encouraging them to drive cautiously and slowly while approaching these curves. A limitation in evaluating such a system is that the best way to evaluate it would have been to have a higher number of test drives and users evaluating it. We couldn't analyze exactly the number of near-collision situations, and how much of them would have been reduced with the aid of our system. We also believe that as can be seen from the test drive with age group 1, assistance needs to be given from the beginning of a driver's experience. Starting to provide assistance to experienced drivers will only encourage a competitive attitude and show one's pride and ego.

VIII. Conclusions and Future Work

1) Improvements over existing systems

We tackled two main points, which we aimed on improving over existing systems. One of these is that existing systems are very costly, because they require the driver to purchase external devices unless the vehicle has existing sensors. These devices then need to be installed by an auto mechanic, and set up by an auto technician or engineer. To buy a vehicle with pre-installed ADASs also is very costly as each feature is an 'optional', which has to be paid for extra. An existing system which works on mobile phones, as mentioned previously, Mobileye, also needs the user to purchase external devices and install them in various locations around the vehicle. The solution mTADA, which we have developed however does not need any external devices, but only makes use of the sensors provided by the host mobile phone itself. Another improvement is that the application also provides audible feedback with the voice of a human being, while also categorizing the different types of warnings according to the gravity of the situation. The system also provides driver behavior feedback while providing turn by turn navigation instructions, which no existing system currently provides.

2) Future Work

Plenty of ideas had been brought forward for the system proposed, however due to time restrictions it proved very difficult to even research them. In this section, we will be identifying the main ideas and explaining how they could be implemented and their possible uses.

Bug fixes - The first work to be done on the system would be to fix the existing known bugs. These include the bugs related to the turn by turn navigation and finding a solution for the harsh braking and acceleration algorithms.

Obstacle detection and tracking - Obstacle detection has already been implemented, however as explained in the testing section, in real-life scenarios a lot of false positives are also being detected due to the different road shades, glares and the need for further pre-processing on the image feed to help identify the vehicles in dark situations.

Lane departure warnings - On fixing the obstacle detection and tracking algorithm the lane departure warning system can be enabled as the basics for this feature have already been implemented. By getting the obstacles from the image and tracking them, it would be possible to analyse whether there are any possible vehicles in blind spots, at the sides of the vehicle. Doing so will make it possible for the system to provide feedback when the host vehicle is deviating from the lane and a vehicle might be on the other lane.

Driving behavior - Another idea is to improve the user activity algorithm to be able to analyze the driving patterns to deduce the driving behavior of the user. This data then can be used to offer constructive criticism to the user based on these patterns. Another idea was that after deducing the driver behavior, feedback is provided to the user on how to reduce fuel consumption – eco driving.

Settings screen - A number of people provided additional comments and feedback in the questionnaire. The idea we liked the most was to have a ‘settings’ screen, which would allow the user to customize the user experience, and deactivate certain functionalities such as lane detection and tracking.

Map data **-** Another idea would be to have the map data downloaded and stored on the device instead of requesting the data every few seconds. This would help to reduce data costs and to preserve battery life. To accomplish this, however, the application would have to store the logs in internal memory, and then upload them to the server as soon as a valid connection is found.

Web application and traffic data **-** A part of the system which we did not manage to work on was a web application, which provides real-time data of the location of the mobile and real time logs. The application would enable many possible uses, such as vehicle tracking for businesses, and also driver analysis for personal use. With the implementation of this real-time system, it would enable the possibility of crowd-sourced traffic analyses and indication systems. This system would be able to analyze the data being transmitted by the devices, consider the different speeds on the different roads to estimate where there would be possible traffic. It could also be improved to indicate where there would be an accident. In fact, another possibility for future work on the system would be to implement collision detection, which would automatically notify the authorities to take action, and neighboring traffic to avoid the road.

3) Conclusions

Road traffic accidents are a major problem, which takes major toll on life in all countries, mainly in industrialized areas and developing countries. The major causes of such accidents are alcohol, driving behaviour, fatigue and driver distraction. When tackling a cause like driver distraction and driver behaviour, it is very difficult for the driver to actually acknowledge that he is doing something wrong since there is no good source for feedback except for passengers or being in a near-accident situation. To solve this, we attempted to develop a system, which pesters the driver, keeping his attention and focus on the road, while also providing feedback based on his behaviour, such as excessive speeding, harsh braking and harsh acceleration.

We did extensive research, and presented the most relevant information. Based on this information we developed a system, which we named mTADA: Mobile Tracking and Advanced Driver Assistance. We implemented mTADA on the Android mobile platform so that it can be easily installed and used by a majority of the population. The system does not make any use of any third party peripherals but only makes use of the sensors already available on the device – GPS sensor, accelerometer, back camera and the touch screen. We managed to implement lane tracking and detection, curve detection, a turn by turn navigation system, and a warning system. However, we encountered some problems while tackling the obstacle detection algorithm because of different colours in the roads. We also encountered some small bugs during testing, which we have mentioned in the testing section previously and pointed out as future work in the previous section. To top it all off, after fully testing the system, we evaluated it in a real world scenario and it proved to be useful. Many people who we questioned claimed to have found it useful, and claimed that they would make use of such an application on a regular basis. We also went on test drives with people from different age groups and analysed their reactions, acceptance and compliance towards the system. Younger age groups were more compliant and accepting of the system, while males from older age groups did not want to be provided feedback, and were neither accepting nor compliant.

We found that the system development was quite difficult at first without making use of external devices, especially in night situations. We managed to design and implement a number of workarounds to tackle most of the 4 main causes. However, the system still has a number of bugs to be fixed, and a number of features and functionalities to be implemented. Even in the current state the system managed to achieve the results we set out to achieve, together with the aims and objectives we identified at the beginning of this paper.

REFERENCES

[1] M. R. Safe, “Road Crashes Are The No 1 Killer Of Young People Worldwide It Is Time to Act,” 4 12 2010. [Online]. Available: http://www.makeroadssafe.org/publications/Documents/no1_killer.pdf. [Accessed 10 02 2013].

[2] L. Norman, “Road Traffic Accidents,” World Health Organization, Geneva, 1962.

[3] P. A. Hancock and P. A. Desmond, Stress, workload, and fatigue, Mahwah, N.J: Lawrence Erlbaum Associates, 2001.

[4] European Transport Safety Council, “Reducing Traffic Injuries Resulting From Alcohol Impairment,” European Transport Safety Council, Brussels, 1995.

[5] T. . R. S. f. t. P. o. Accidents, “Driver Fatigue and Road Accidents,” The Royal Society for the Prevention of Accidents, 2001.

[6] T. R. S. f. t. P. o. Accidents, “Driver Distraction,” Road Safety Information, December 2007, pp. 1-7.

[7] R. L. Dukes, S. L. Clayton, L. T. Jenkins, T. L. Miller and S. E. Rodgers, “Effects of aggressive driving and drvier characteristics on road rage,” The Social Science Journal, 2001, pp. 323-331.

[8] Y. Wu, “An Introduction to Computer Vision,” Northwestern University, Evanston, 2012.

[9] S. Tuohy, “Real Time Distance Determination for an Automobile Environment using Inverse Perspective Mapping in OpenCV,” NUI Galway, Galway, 2010.

[10] O. Gietelink, J. Ploeg, B. D. Schutter and M. Verhaegen, “Development of advanced driver assistance systems with vehicle hardware-in-the-loop simulations,” Delft University of Technology, Netherlands, 2006.

[11] Volkswagen, “Driving. Safety. Powered by Innovation.,” Volkswagen Group of America, 2009.

[12] J. Mejeur, “Ignition Interlocks, Can Technology Stop Drunk Driving?,” State Legistlatures, 2007, pp. 16 - 21.

[13] Smart Start Inc., “Smart Start Inc.,” Smart Start Inc., [Online]. Available: http://www.smartstartinc.com/index.php/products/about-interlock-and-the-advantages/. [Accessed 10 1 2013].

[14] L. Barr, S. Popkin and H. Howarth, “A Review And Evaluation of Emerging Driver Fatigue Detection Measure and Technologies,” U.S. Department of Transportation, Washington, DC, 2009.

[15] Mobileye, “Mobileye C2-270,” Mobileye, 2009.

[16] Cellocator Division, Pointer Telocation Ltd., “Cellocator Cello-IQ,” Pointer, Rosh Haayin, 2013.

[17] Teltonika, “Teltonika,” Teltonika AB, [Online]. Available: http://www.teltonika.eu. [Accessed 12 4 2013].

[18] H. Sugiharto, “Current and Future Mobile Platforms,” Berlin Institute of Technology, Berlin, 2010.

Obesity Management in Children on an iOS Device

Alexiei Dingli and Gary Hili
University of Malta
Msida, Malta
emails:{alexiei.dingli@um.edu.mt, gary.hili.10@um.edu.mt}

***Abstract*- The main objective of this paper is to create an application that makes use of gamification principles in order to teach and guide children and teens towards a healthier life and encourage them to lead it. The application also allows users to track their overall progress and it instructs them on how to proceed. While doing so it gives them tips and facts about improving their lifestyle. The application also tries to apply gamification techniques like having a sense of competition and providing feedback so that tasks like exercise are more engaging. To evaluate the application, a group of people made use of the application as they saw fit, for a period of time. The gamification techniques used were successful in giving them the desired experience. With the competitive element between friends encouraging them to make more use of the application and the tips and facts helping them towards a healthier lifestyle. The biggest concern was that the application did not encourage the users to keep using it as much as we'd liked.**

Keywords-obesity; healthy; iOS; gamification; lifestyle

I. INTRODUCTION

The problem of obesity is a growing concern that has reached epidemic scales in most developed countries within Europe and the United States of America [16], [18]. Recent studies show Malta as having the most obese adult and child males in Europe [1], [2].

The biggest concern is that obesity puts children and teens at risk of developing more serious health issues, which are normally seen at an older age [17]. It is also a strong indication that the child or teen will be obese once adulthood sets in.

This helped motivate us into trying to find our own solution to the problem; so we set out to develop an iOS application that would help the users tackle obesity. With current smartphone devices, the application's potential is endless, even more so since current generations spend so much time playing some kind of game on some type of device or console [19].

After reviewing various publications and current technologies, we came up with the ***My Lifestyle App*** for the iOS. It makes use of gamification concepts like using a points and a leader board system in order to engage users and help them maintain healthier lifestyles. The application also attempts to challenge and teach users so that they can lead a healthier lifestyle.

Once we developed the application, we also had a group of people trying out the application for a period of time before evaluating it. After they evaluated it, we analysed all the results and came to a conclusion.

In this paper, we give a short analysis of different papers that we reviewed with regards to current technology used for nutritional purposes and different gamification concepts. We then move on to stating our aims and objectives and how we plan on designing our proposed application in order to reach these objectives. From there, we move on to the Implementation stage. There we give a detailed outlook on how the application was developed. Finally, we give a report of our findings before coming to our conclusion. In the conclusion, we try to summarise our findings and state different ways of how we plan on improving the application for future reference.

II. LITERATURE REVIEW

The literature review was split into two main sections. The first section revolved around papers, publications, and applications regarding technology that is being or will be used in helping to improve ones nutritional values and help to combat obesity. The second section was with regards to gamification and the different techniques that can be used in order to make mundane and tiring tasks more fun, engaging and meaningful.

A. Technology and Nutrition

The basis of how we plan on tackling obesity, as said by Tsai et al. [*3*] is to modify the two main modifiable aspects of an obese or overweight person's life. These aspects are their diets and their physical inactivity. So, we sought out to examine the different technologies that were able to help in improving these aspects of someone's life in some way or another.

With the advancements in smartphones, monitoring food intake has become much easier, where various systems allow you to scan barcodes and retrieve the item's nutritional values very easily. Yet, for this to function properly, we need to have large amounts of data stored in large online databases, with food items for all kinds of societies since different societies will have different brands and different food items. This would also make the application more dependent on an Internet connection, which might not always be available. It might also have efficiency issues in order to retrieve certain information from the database, so it will need to be optimized to enable quick searches.

Although such a system would allow an application to tell you what kind of food you would need during the day, and what's good and what's not, simply allowing them to monitor themselves still would not be enough. This is mostly due to the fact that even if you simply scan an item to input it's nutritional values, some users will still either forget or not bother to input the items, which would result in the application giving them bad instructions. Therefore, we will also need to teach the user about why and how they should eat healthier, so that they would not necessarily need to monitor their food intake. Instead they will know what is good for them and what's not.

One way of teaching nowadays is through games, and this is becoming especially common in younger generations. One simple example can be seen in [*4*], where Piziak et al. make use of a Pictorial Bingo Game, which would help teach children about healthy foods, either through riddles instead of calling up numbers like in a normal bingo game or by simply showing them the image of a healthy food item for them to cross out on their bingo card. Piziak et al. also conducted a study about how effective such a method was. Parents of the children were asked to state if they saw a change in their child's eating habit after using the game, and they reported some good results. Parents stated that they saw an increase in vegetable consumption, also showing that children were actually learning about their food in a fun and enjoyable way. In a different publication (Yien et al. [*5*]) they developed a course that made use of different games to teach children about different ways of living a healthier life. The course was split into 4 sections, with each section making use of a different computer game. It included sections like learning about food for psychological and physical needs, where they learnt about the different food groups within the food pyramid. An illustration of the food pyramid can be seen in Figure 1.

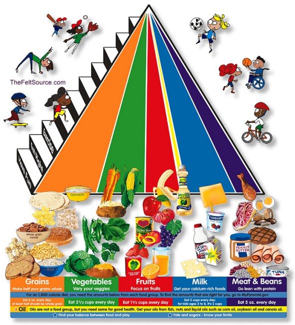

TheFeltSource.com
Have Fun
Eat Right
Grains
Make half your grains whole
Vegetables
Vary your veggies
Fruits
Focus on fruits
Milk
Get your calcium-rich foods
Meat & Beans
Go lean with protein
For an 1,800-calorie diet, you need the amounts below from each food group. To find the amounts that are right for you, go to MyPyramid.gov
Eat 6 oz. every day: at least half should be whole grain
Eat 2½ cups every day
Eat 1½ cups every day
Get 3 cups every day: for kids ages 2 to 8, it's 2 cups
Eat 5 oz. every day
Oil
Oils are not a food group, but you need some for good health. Get your oils from fish, nuts and liquid oils such as corn oil, soybean oil and canola oil.
Find your balance between food and play
Fats and sugars - know your limits

Figure 1. Food pyramid indicating the six main food groups
Source: [*6*]

They also learnt about how certain environmental factors that came into play when eating healthy foods. They learnt about eating less fast food and about eating alternative, healthier foods. Their last lessons involved a one-day diet where they monitored what they consumed for a whole day, so that for a single day they were more aware of what they were actually eating. After this one-day diet, each subject compared their personal lifestyle to that of a healthy lifestyle in order to see how they were doing. They also learnt about how they can improve their lifestyle, by making use of certain rules to develop a healthier example of living.

So, by making use of simple games to teach children, it can have very positive effects. Yet, as mentioned earlier, we not only want to teach our users about healthy eating, we also need to find ways of how to help increase our users' physical activity during the day. Different means were researched, like use of active video games and use of smartphones and other technologies, which can hopefully have some effect on a child or teen's level of activity.

With regards to active video games, there were two publications that generated the most interest in order to help us design our application. Both these papers conducted studies on the use of active video games and on how effective they really are. The first publication is by Maddison et al. [*7*] and they conducted a controlled study with regards to active video games used by children, where they had advised their intervention group to perform a minimum of 60 minutes of moderate to vigorous exercise every day. The children involved (both control and intervention groups) would also be monitored three times during the entire study. At the start, half way through and at the end of the study period, a series of tests were conducted to see if there was any improvement in the children's Body Mass Index (BMI) and other aspects. The study resulted significant difference in the BMI percentile of the intervention group against that of the control group. With the intervention group losing an average of 0.33% of their BMI, whereas the control group averaged a loss of 0.08%. Yet in our second paper by Baranowski et al. [*8*], they conducted a more naturalistic study, meaning that there was less intervention within the subject's life. They simply gave their subjects a Wii console together with some active games and let the children use them in any way they wanted, without any kind of instructions being given to them other than what the video game itself instructs them to do. This study gave conflicting results when compared to the paper by Maddison et al. [*7*]. They came to a conclusion that when the children were given a more natural setting, they either did not play the games that they were given or they compensated by being less active during other times of the day.

Here, we can see the conflicting result where when given the more natural setting, children were not making much use of the active video games. This resulted in no significant improvement on the overall health of the subjects [*8*], unlike what we saw in the more controlled setting, which is more unrealistic [*7*].

As for smartphone devices, we can make use of the device's in-built sensors like an accelerometer, a gyroscope, and Global Positioning System (GPS) to help track the device and the user's movement. So, with such a device there is great potential to create great applications that can help increase the user's physical activity. Such an application can be seen in [*9*], where Arteaga et al. developed an application that recommends third party health and fitness apps that it deems suitable for that particular user. Such an application works by first making use of the Big 5 Personality Theory to help get an idea of what kind of applications and tasks the user would mostly like to perform. Then, when the application recommends a certain application for the user, the user can accept or decline the recommendation, thus helping the app to continue learning about the user in order to better tailor the apps for him/her.

From all the publications, it was very evident that by making use of smartphone technology, we are able to give the user a better interactive experience, and we are able to track their daily activity without having to be too intrusive in their daily routine, yet still manage to be part of this routine. The device is also a great medium to help teach our users about healthy living.

B. *Gamification*

Gamification is the art of adding game mechanics to a system in order to make it less tedious and more enjoyable to use. Such a technique is ideal for making exercise seem less tedious and more like a game where you can compete against your friends. With our gamification techniques, we aim to achieve a Rate of Flow [*10*], a theoretical state developed by Mihaly Csikszentmihalyi, where the user is balanced between a state of anxiety and boredom. It is at this particular state that we are the most concentrated on performing a certain act, making the task feel easy and effortless to perform.

In order to achieve this rate of flow and have a successfully gamified application, we tried to find what makes a successfully gamified system and what we should avoid from doing. We saw various frameworks and theories about how and what needs to be done in order to reach our goal. With all the different theories we analysed, they all came down to the Mechanics, Dynamics, and Aesthetics of the system, also known as the MDA Framework [*11*]. The Mechanics of the system or the application are the individual aspects and background workings of the application. It includes the rules, conditions, game states, and user representation of the game or the system. The Dynamics of the game take into consideration the game's interaction with the user through the different mechanics during run-time. It does so in order to react to the user's inputs, enabling the user to make use of the mechanics. The dynamics also considers how intuitive the application is for the user and it also enables consistency within the application, which is important in order to give the user a good experience. The final aspect is the Aesthetics, which helps to evoke the desired emotions that we try to create with the mechanics and the dynamics of the system. These are the means by which the user will give his/her input and how he/she will receive feedback from the application.

By making use of this framework when gamifying our system, we will be able to give our user a more engaging experience. It will also help to give the application more meaning, which is essential if we want our users to keep on using the application.

From the publications that we reviewed we were also able to come up with the kind of gamification concepts that we can use within our application. One aspect is social integration within the application. This means that we make use of social networks like Facebook, so that users can share their progress and all their achievement on the social network for all their online friends or other app users to see, like and comment so that they could help encourage the users in reaching their goal [*12*]. Another aspect is the feedback system. Here, we would make use of a points system so that for every achievement or completion of a task we give the user an amount of points, thus giving the user an indication of how well or poorly they are doing. Also, by making use of the user's total points we can create a global leader board that would introduce a more competitive aspect. This would help to motivate our users by making them compete against their friends, thus resulting in them making more use of the application.

What we would not want to happen is to have the application be meaningless for the user. Such a thing happened to the Foursquare system that would allow the user to claim mayorship of certain destinations, which they would often visit. The concept of being a virtual mayor of a real destination did not really give the user much meaning. This is why, back in 2012, they completely redesigned the system and went from a game oriented application to a recommendation application, which it was intended for in the first place.

After reviewing some publications on successful and unsuccessful gamified systems, we came to a conclusion that whenever we need to gamify any system, our aim has to be that we try to fully immerse the user in the system. For this to happen we need to give meaning to the system, so that the user is actually doing something with purpose. Without having some form of meaning, users will get bored and stop using the system or the application.

III. AIMS AND OBJECTIVES

The main objective behind the development of this application is to help users engage in a healthier lifestyle by making use of a mobile app on an iOS device. The application will specifically target children and teenagers between the ages of 12 to 19 years. Since not a lot of time is available in order to have a full trial where we can see whether or not the application makes any significant difference, we will only be aiming to get our users on the right track towards leading a healthier lifestyle.

To make the application more fun and engaging for our users we will be making use of gamification in order to help teach and guide them in maintaining a healthier lifestyle. Other than simply teaching them about healthier choices, the application will also try to challenge the user to perform certain tasks that will increase their physical activity, tasks include walking instead of taking the bus, going out for a quick run, taking the stairs instead of an elevator and so on. Therefore, we will be using the iOS device to challenge our users in real life instead of in a virtual environment.

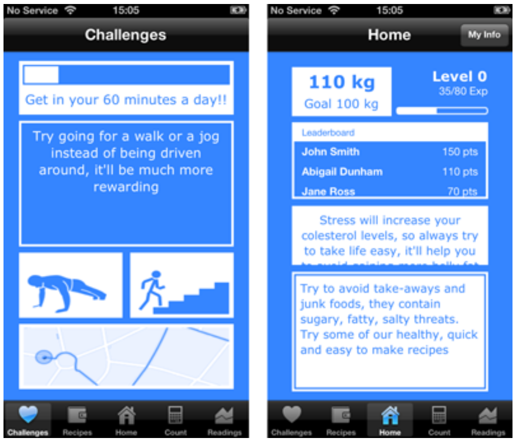
In order to achieve our goals, we try to apply gamification techniques like using a points feedback system to help encourage the user and a leader board system to create a sense of competition. We also incorporated a social aspect to the game by giving users the ability to share their current progress on their Facebook accounts.

Hopefully, by making use of these concepts we are able to engage the users and have them feel a great sense of achievement by having completed the challenges set out for them.

IV. Design

While designing the application, we kept in mind that we wanted to be able to give the user the necessary functionality, but this shouldn't mean that its aesthetic qualities should suffer. First of all, the application will be designed using XCode (https://developer.apple.com/xcode/) and it will be optimized for use with the iPhone/iPod Touch running iOS 6.1 or later, allowing us to make use of the accelerometer and Global Positioning System (GPS) sensors found on the devices.

To start making use of the app, we will need certain information from the user in order to give him/her the appropriate instructions. The information needed will allow us to calculate the user's BMI, their ideal weight and Resting Metabolic Rate. The application will also request Facebook permissions so that the application is allowed to post on behalf of the user.



No Service 15:05

Challenges

Get in your 60 minutes a day!!

Try going for a walk or a jog instead of being driven around, it'll be much more rewarding

Home

My Info

110 kg

Goal 100 kg

Level 0

35/80 Exp

Leaderboard

John Smith	150 pts
Abigail Dunham	110 pts
Jane Ross	70 pts

Stress will increase your colesterol levels, so always try to take life easy, it'll help you

Try to avoid take-aways and junk foods, they contain sugary, fatty, salty threats. Try some of our healthy, quick and easy to make recipes

Figure 2. Application screenshot showing the Challenges and Home Sections, together with the other three sections shown in the tabbed navigation bar
Source: Gary Hili - 2013

The application design is split into five sections, fitting the iOS's tabbed navigation controller as seen in Figure 2, so that we can display all of our content in the different sections for the user to easily navigate through. These sections are the following:

- Home Section – This is the first view seen by the user. It contains a progress bar together with a summary of how the user is doing. It also has an area with some tips and facts about eating and living healthier. An important aspect of our gamification techniques can also be found here, since we included the overall leader board, which displays the top ten users.
- Recipe Section – This section includes a list of healthy, easy to make, tasty recipes for every time of the day that the user can enjoy. It also gives the user the possibility of sharing an image of what they prepared on the social network, by doing so they will also be earning points.
- Weight Readings – Used to track the user's progress in terms of their weight and height in order to get an idea of the user's current Body Mass Index. Here, users will be able to get a more visual representation of how they are doing by having a scatter graph instead of a list of numbers with their past weight.
- Calorie Counting – It allows the user to track the calorie content of the food they consume and calculate the caloric burn of the exercise or physical activity that they perform. By inputting such information, the application will be able to instruct the user on how to proceed by first calculating their current caloric balance.
- Challenges Section – This is a very important section within the application. Here the application will be tracking the user while he/she is performing certain exercises that the app can track. These exercises include push-ups, stair climbing, walking or running. In this section, the application recommends certain exercises depending on the time of day and the amount of exercise performed so far in the day. Users will then be given certain amount of points depending on how well they perform during each challenge.

By having the user make use of these different sections we will be able to help them achieve their goal of leading a healthier life.

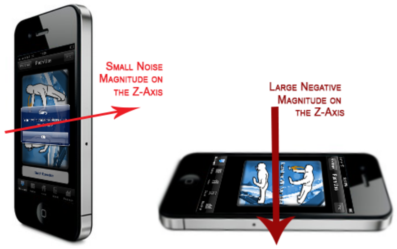
V. Implementation

When we implemented the application, we made use of the Model-View-Controller design pattern, which has the advantage of separating the data and the information from the user, resulting in cleaner, reusable code and a more efficient application.

A. Controller Layer

The Controller layer is where the application communicates from the front end to the back end of the application. Classes in this layer don't have much logical computation performed in them. In our implementation, some of the controllers do perform some of the more complex operations within the Challenges section of the application.

- In the Push-Up challenge, the controller will receive accelerometer data every time the push up button is touched.



Small Noise Magnitude on the Z-Axis
Large Negative Magnitude on the Z-Axis

Figure 3. Use of accelerometer to detect the device's orientation
Source: Gary Hili - 2013

The purpose of this is to detect the device's orientation, as seen in Figure 3, so that it can help detect if the user is cheating by not having the device lying face up but instead he/she is holding it in their hands.

- The Stair Climber challenge also makes use of the accelerometer sensor. In this challenge, the controller is constantly receiving information from the accelerometer.

The controller then calculates the magnitude of the three axes of the accelerometer and if a certain threshold is reached that will be counted as a step. It continues to do this until the user either chooses to stop or the challenge has been completed.

- In the final challenge, the Walk/Run Challenge, the application makes use of the devices GPS tracking capabilities in order to track the user's route, speed and distance. The Walk/Run Challenge also consists of one of the biggest disadvantages within the application. This is the fact that the current version isn't capable of tracking the user once the device's screen has been locked.

Other than the above-mentioned algorithms, the Controller will mostly handle any data being sent by the model layer and any user interactions that are made with the view layer and require some kind of processing.

B. View Layer

The View layer contains all the aesthetic properties of the application. It is this layer of the application that the user will interact with and be given feedback. Implementation of this layer was relatively easy since we made use of the XCode storyboard feature, so designing and connecting the different views was quick. Designing the look of the application still took some time in order to achieve the best possible aesthetic appeal. While designing this layer we tried to make use of a simplistic approach, using a monochrome design with contrasting colours throughout the application. This simple approach was used in order not to overwhelm the user and so that the emphasis can remain on the content of the application yet still be appealing to the user. One of the more complex views was the Weight Readings Section since here we had to include the CorePlot Library [*13*] in order to design the view.

Most of the maintenance of the view layer is done by the iOS itself, so the main job of the views is to simply fire actions each time a user interacts with that view so that any listening methods can perform the necessary action.

C. Model Layer

As for the Model layer, this is the backbone of the application, where all the logical and mathematical calculations are being made. At this layer we have all the data controller classes together with the classes that manipulate this data before being sent to the controller layer in order to be sent to the view layer. The data controllers are used to load and save data in XML format. They are able to either load the necessary data from the application bundle or the devices document directory and they also have basic array manipulation methods that allow other classes to make use of the loaded data. We use the data controllers at various points within the application. They control the challenge difficulty levels, the list of recipes and the handling of weight and calorie count entries amongst other tasks. This layer also makes use of algorithms that help determine what tasks should be recommended or what instruction should be given to help the user. In order to give the necessary instructions to the user, we have certain classes that take the available data and calculate the BMI, Resting Metabolic Rate (RMR), the caloric balance and so on, storing these values for other classes to use. So, these values are then called by another class, which uses them to determine what kind of instruction to give the user. By using BMI and the caloric balance, the application will tell the user how they are doing during that day, instructing them to eat or exercise more, or whether they are overdoing it and eating or exercising too much. The model layer also instructs users on what kind of exercise they should perform. The algorithm used to tell the user what challenges to perform makes use of the time of day together with how much exercise they've performed so far during that day. The main reason why we used this approach is because our main age focus is 12 to 19 years of age, and at these ages they are spending 9 months out of 12 in school, so they are more likely to be home at certain hours, and at school at others. Hence, in the mornings and the evenings the application will be suggesting the Stair Climb challenge since it might be too early or too late to actually go out for a walk. As for the Walk/Run Challenge, it's mostly encouraged after school, maybe walking back home or to some after school activity. As for the push-ups they are mostly being used as a way to working your body and not be sitting down all day, so we create this easy challenge where they don't really need to go out. Being such a low calorie burning exercise, the application would not be recommending it as much as the other challenges. Figure 4 shows what kind of challenge is being recommended during a certain period of time. Taking into account the amount of time spent exercising during the day, the algorithm will no longer recommend the Walk/Run Challenge. Instead it recommends the Stair Climb Challenge, which is easier to perform indoors in a shorter period of time.

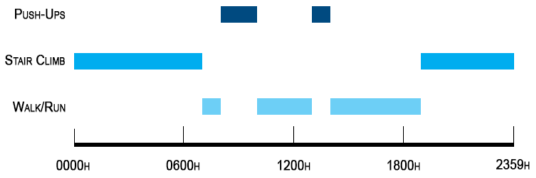


Figure 4. Gantt chart of the times each exercise is suggested
Source: Gary Hili – 2013

The instruction being provided also makes use of positively reinforced statements because certain studies (Woolfard et al. [*14*]) had shown that making use of such messages makes the users feel better about themselves, boosting their self-esteem.

The last thing the model layer is handling is the social integration of the application and any network related functions. This includes retrieving and setting data on the leader board being used in the home section.

Implementation of the artefact took longer than expected since we were new to technology. However, the resulting application was very satisfactory and once testing of the application's features were performed, the beta evaluation of the artefact could commence.

VI. Results and Evaluation

In order to evaluate the application, we first conducted a 20-day trial period, where a group of 25 users were asked to use the application as they saw fit and then evaluate it by filling in a questionnaire. Each user's points throughout this period were also being tracked so that we were able to evaluate how much they actually made use of it.

The aim of the application was to help our users in order to start leading a healthier life, so we tried to improve the users' eating habits by teaching them about healthier alternatives and giving them the possibility of tracking their food consumption for them to be more aware of what they eat. We also tried to increase their physical activity by making use of the different challenges and encouraging them to exercise more during the day. To make it all the more enjoyable we implemented gamification concepts that motivated our users to compete against one another and encourage one another.

The users that evaluated the application did not have any severe problems of obesity. 52% were female and the rest male, with 76% between the ages of 19-21, 16% were between the ages of 16-18 and the rest were over 25 years of age. Although most of the users were not within the age group the application was targeting, their input was still valuable. The reason being that they are able to express themselves better about how the application is helping them and how it can be improved.

In order to evaluate the application they were given a questionnaire, which regarded the use of the gamification concepts within the application. It also regarded how well the application helped the user with regards to their diet and exercise, and finally they were asked to report any bugs or future work they'd like to see.

With regards to the use of gamification, users found the points and leader board techniques to be very helpful and it made them want to use the application. The biggest reason was because of the competition it evoked, which was motivational for them. The social integration factor wasn't as successful as we'd hoped it would be. We had included this feature since in recent years we have seen an increase in how people are constantly posting and sharing statements about their lives on such online networks. With our application, users did not really make much use of this feature, resulting in the majority (68%) not even realizing that using it would earn them more points, yet 70% of those that did not realize this said that they would've made more use of it if they had realized it gave them extra points. Another important aspect of a gamified system is the application's intuitiveness and ease of use, which is highly dependent on the aesthetics of the application. We had 96% of the users finding it very easy to use, while the other 4% gave a neutral opinion on the matter.

Next, the users were asked about how the application helped them increase their physical activity. Most of the users liked all the three challenges that we provided but they were not used as much as they were liked. This meant that the challenges were not effective in helping our users lose weight, but were helping in keeping the users engaged with the application. For example, 80% of the users liked the Stair Climber Challenge, 92% of the users even said that having an application that tracks their exercise is great, yet only 40% of the users said that they made use of Stair Climber challenge on a daily or almost daily basis. The same scenario can be seen throughout the three different challenges (as seen in Figure 5).

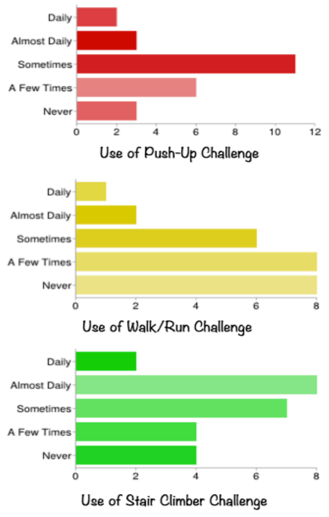


Figure 5. Bar charts showing usage of challenges
Source: Gary Hili – 2013

Then the users were asked about how the tips and facts given to them about living healthier managed to help or in any way affect them. The tips were found to be very helpful and easy to remember as well, with most of the users having recited a tip or a fact that they had read while using the application. As for the calorie counting aspect of our application, we saw that tracking the user's food consumption wasn't very effective. A small percentage said that it was easy to use, yet the majority said that they either did not use it or they couldn't bother to search the web to see how much calorie an item contains. Some users also suggested that we try to make this easier by adding a way that the users can actually search their food's nutritional values and add it to their list.

In the final part of our evaluation, we analyzed the data recorded by keeping track of the users' points throughout the 20-day period. As one would expect, some users made more use of the application than others. Our biggest concern was that even for a small period of time, users were not making use of it daily, with some users even showing long periods of not making any use of the application. By the last 6 days of the trial period, about 90% of the users had stopped making use of the application entirely. Around half way through our trial period, we introduced an update to the application, which handled some bugs that we had found. After making this update available for the users we could see a slight overall increase in the usage of the application even though no new visible content was released.

What we were able to conclude from this was that we were not able to retain the users as much as we'd liked. When we asked some of the more active users why they think this happened various responses were given. Yet most of them stated that they had moved on to another application that was capable of giving them more detailed reports with regards to their current performance and is capable of giving them detailed reports on their personal computer as well.

By making use of gamification, we were able to successfully motivate our users so that they'd make more use of the application by competing against their friends. We also managed to help them learn about ways to lead a healthier life. However, we still were not able to retain the user's interest for a long enough period to make a significant difference in their lifestyle. Other pitfalls with our application included our food tracking implementation, which, although easy to use, isn't able to tell the user the nutritional values of a food item. The application also lacked in the social aspects, since users did not really make much use of the social networking features.

Even if we had some negative results from the evaluation of the application, we still managed to get some results that have helped us get a better understanding of how we can fix and improve the application in order to benefit our users and make a significant difference in the lives.

VII. Conclusion and Future Work

Our aim when we started this project was to develop an application that would help obese children and teenagers start leading a healthier lifestyle.

Although we managed to make a small difference in the lives of our users, room for improvement can be evidently seen, most of all we would need to introduce a better way to track the users' caloric consumption. This can be done using various ways depending on how technology will move forward. One way is to use large databases with different choices that the user can search for. Another way would be to introduce edible RFIDs [*15*], which could contain such information directly on the RFID or via an online database. The application also needs to introduce background processing in order to improve the application's exercise tracking capabilities.

Some other possible ideas were also mentioned in the user evaluation of the application. This included the introduction of more challenges, which would give a better variety. The users also said that one kind of additional feature could be the ability to upload images of their workouts.

We were able to successfully design and implement our application, using gamification concepts throughout in order to help teach the user about healthier foods and motivating them to increase their physical activity. After having analyzed all our findings, we were able to come up with a possible way forward that would improve the application in general. Most of all, it will improve the user retention and hopefully have the users make more use of the application, maybe even on a daily basis.

References

[1] J. Ameen, Times of Malta [Online], http://www.timesofmalta.com/articles/view/20120708/local/Study-confirms-children-s-obesity-problem.427583, July, 2012 [retrieved: October, 2013]

[2] B. deNorre, Eurostat Commission [Online], http://www.ec.europa.eu/malta/news/28.11.22_eu_obesity_rankings_en.htm, November, 2011 [retrieved: October 2013]

[3] C. C. Tsai et al., "Usability and feasibility of PMEB: A mobile phone application for monitoring real time caloric balance", Mobile Networks and Applications, vol. 12, June 2007, pp. 173-184.

[4] V. Piziak, "A pilot study of a pictorial bingo bilingual nutrition education game to improve the consumption of healthful foods in a head start population," International Journal of Environmental Research and Public Health, vol. 9(4), April 2012, pp. 1319-1325.

[5] J.-M. Yien, C.-M. Hung, G.-J. Hwang, and Y.-C. Lin, "A game-based learning approach to improving students' learning achievements in a nutrition dourt," The Turkish Online Journal of Educational Technology, vol. 10(2), April 2011, pp. 1-10.

[6] The Felt Source [Online], http://www.thefeltsource.com/New-Food-Pyramid-Large.jpg, [retrieved: April, 2013]

[7] R. Maddison et al., "Effects of active video games on body composition: a randomized controlled trial," The American Journal of Clinical Nutrition, vol. 94, July 2011, pp. 156-163

[8] T. Baranowski et al., "Impact of an Active Video Game on Healthy Children's Physical Activity," Official Journal of the American Academy of Pediatrics, vol. 129(3), March, 2012, pp. 636-642

[9] S. Arteaga, M. Kudeki, and A. Woodworth, "Combating obesity trends in teenagers through persuasive mobile technology," Sigaccess Newsletter, vol. 94, June 2009, pp. 17-25.

[10] M. Csikszentmihalyi, Flow: The Psychology of Optimal Experience, 1st ed., Harper Perennial Modern Classics, 2008.

[11] D. R. Flatla, C. Gutwin, L. E. Nacke, S. Bateman, and R. L. Mandryk, "Calibration Games: Making Calibration Tasks Enjoyable by Adding Motivating Game Elements," Proc. ACM Symp. User interface software and technology (UIST 11), Santa Barbara, California, October 2011, pp. 403-412, doi: 10.1145/2047196.2047248.

[12] K. Kiili, A. Perttula, P. Tuomi, M. Suominen, and A. Lindstedt, "Designing mobile multiplayer exergames for physical education," Proc. IADIS, International Conference Mobile Learning, Iadisportal.org, 2010, pp. 141-148.

[13] D. McCormack, and B. Wark, Cocoa plotting framework for OS X and iOS [Online], http://code.google.com/p/core-plot/, [retrieved: October, 2013].

[14] S. J. Woolfard, S. J. Clark, V. J. Stretcher, and K. Resnicow, "Tailored Mobile phone text messages as an adjunct to obesity treatment for adoloscents," Telemed Telecare, vol. 16(8), Oct. 2010, pp. 458-461, doi: 10.1258/jtt.2010.100207.

[15] A. Kooser, CNet [Online], http://news.cnet.com/8301-17938_105-20070913-1/chew-on-this-nutrismart-edible-rfid-tags/, June, 2011 [retrieved: April, 2013].

[16] Centers for Disease Control and Prevention [Online], http://www.cdc.gov/obesity/data/adult.html, August, 2013 [retrieved: October, 2013].

[17] Centers for Disease Control and Prevention [Online], http://www.cdc.gov/healthyyouth/obesity/facts.htm, July, 2013 [retrieved: October, 2013].

[18] World Health Organization Europe [Online], http://www.euro.who.int/en/health-topics/noncommunicable-diseases/obesity/obesity, 2009 [retrieved: October 2013].

[19] V. Vahlberg, "Fitting into their lives: A Survey of Three Studies about Youth Media Usage," NAA Foundation, March, 2010.

Towards a Framework for Designing Secure Mobile Enterprise Applications

Basel Hasan
Department of Computing Science
Oldenburg University
Oldenburg, Germany
basel.hasan@uni-oldenburg.de

Jorge Marx Gómez
Department of Computing Science
Oldenburg University
Oldenburg, Germany
jorge.marx.gomez@uni-oldenburg.de

Joachim Kurzhöfer
AS Inpro GmbH
Lufthansa Systems
Oldenburg, Germany
joachim.kurzhoefer@lhsystems.com

***Abstract*— Mobile devices like smartphones and tablets are not only designed for private use, but also for business use as well. Mobile solutions, such as mobile enterprise resource planning and mobile business intelligence, are nowadays becoming more common. However, without strong consideration of security, especially in mobile devices, these solutions would be very risky. Enterprise data are classified in security levels, in which security threats and countermeasures are grouped. These levels indicate the fulfillment degree of the security objectives in each group. From the enterprise point of view, the boundaries between these levels concerning the mobile devices are not clear. In this research, risk analysis with focus on mobile devices is conducted and a framework to design secure Mobile Enterprise Applications (MEAs) is developed. This framework supports developers in the decision making process when designing secure MEAs side by side with promoting the trustworthy usage of mobile devices in business sectors.**

Keywords: Enterprise Mobility; MEAs; Security; Risk Analysis; User Acceptance.

I. INTRODUCTION

Mobile technologies and applications have been greatly improved in the recent few years making the ubiquitous communications a growing reality [1, 2]. It comes to the enterprise mobility concept when the enterprise integrates mobile technologies into its existing IT infrastructure besides giving its employees better possibilities to work on the move effectively [3]. Nowadays, the talk is about MEAs (e.g., Mobility for SAP, which enables mobile access to SAP® CRM, SAP® ERP and various SAP® Workflows via smartphones and tablets [4]).

MEAs are characterized according to A. Giessmann et al. [5] as: "[…] applications that are designed for and are operated on mobile devices and which facilitate business users within core and/or support processes of their enterprises". They are classified into five categories: mobile broadcast, mobile information, mobile transaction, mobile operation, and mobile collaboration applications [6]. Mobile broadcast category facilitates large-scale information broadcast, such as advertisement and promotions. Mobile information category provides information requested by the mobile user, such as job vacancies and timetables. Mobile transaction category eases and executes transactions, such as e-transactions and the transactions of Customer Relationship Management (CRM). Mobile operation category covers internal operational aspects of the business, such as inventory management and Supply Chain Management (SCM). Finally, mobile collaboration category supports collaboration among employees and various functional units. The proposed framework takes these five aforementioned categories into consideration.

Mobility gives enterprises many advantages. It enables the ubiquitous real-time access to critical business information which supports the managers to meet strategic decisions in shorter time to satisfy their customers' need [3, 7]. Consequently, mobility increases worker productivity and reduces business operation costs [7]. Due to these advantages, enterprises demand mobility and flexibility of their workers as inevitable success factors [8]. However, the involvement of mobile technologies and applications has also brought new security challenges and risks, particularly on mobile devices.

In this paper, the most relevant definition of information security is taken from the ISO/IEC 17799 standard that defined it as follows: "Information security is the protection of information from a wide range of threat in order to ensure business continuity, minimize business risk, and maximize return on investments and business opportunities" [9]. Although mobile devices face a wide range of potential security threats [8, 10, 11], mobile applications are developed often without implementing proper security measures [2]. The major security threats related to mobile environments include, but are not limited to: device loss/theft, data interception and tampering, malware, vulnerable applications, compromised devices, mobile operation system vulnerability, and social engineering. Some of these threats are similar to those in a traditional desktop environment. However, the more prominent threats in mobile environments are malware, data interception and tampering, and device loss or theft [2]. Due to the small size and high portability of mobile devices, they can easily get lost or stolen [12]. According to McAfee report, 40% of the surveyed companies had mobile devices lost or stolen and half of these lost/stolen devices contained critical business data [13]. Consequently, unauthorized third parties can make use of these critical data [8]. Moreover, the disclosure of such kind of data might have harmful consequences on enterprise like financial loss or even the loss of its reputation [10, 14].

Integrating mobile devices into enterprises means that sensitive business data could be accessed everywhere and anytime using mobile devices. Conforming to Bring-Your-Own-Device (BYOD) trend [15], where mobile workers use their personal mobile devices, critical and sensitive business information might be located on these personal devices. The

ISBN: 978-1-61208-313-1

more sensitive the data are the higher security level is required. In general, to achieve a certain level of security, appropriate countermeasures must be applied and that might restrict the use of mobile devices. Therefore, mobile workers have to accept all the restrictions on their own devices. As a result, mobile security solutions must hold a balance between the private and business use.

The key concern in MEAs is the mobile application security including information confidentiality, integrity, and availability. This comes from the issue that communications via mobile networks, in which security threats can take place anywhere, are more vulnerable to be attacked than wired networks [6]. Kelton Research had shown that 75% of 250 surveyed companies, which their revenues are up to $100M across the United States and United Kingdom, considered security the major factor that prevents companies from adopting mobile applications [7].

The research in this paper focusses on security issues in mobile environments with emphasis on MEAs. It represents a work in progress to discuss and investigate new ways towards building a framework for secure MEAs. The rest of this paper is structured as follows: Section II presents the research problem. Then, the adopted research process is presented in Section III. Section IV proposes and presents the details of the secure mobile enterprise applications design framework. Related work is then presented in Section V. Finally, the paper sums up with a conclusion and future work in Section VI.

II. Problem definition

Mobile devices are exposed to a wide range of threats that have to be countered. The vital point in this regard is finding and applying appropriate security countermeasures. According to T. Wright et al. [16]: due to the significant resource constrains of the mobile devices, many security countermeasures from traditional computing domains are not translated well to mobile devices. In other words, simply porting standard information security tools from stationary computers, notebooks, and server domains to mobile devices is unlikely to be effective [16–18]. In order to achieve a certain level of security, the mobile user has to accept some restrictions on the features and functions supported by mobile devices. Examples for such restrictions are: specifying exactly which applications are permitted to be installed, or restricting the types of connections that a third-party application can establish. The employee, who wants to access very critical information using mobile devices, might accept a wide range of limitations. However, these limitations might be not accepted in the case that the employee doesn’t need to access this critical information. Generally, a high level of security might be reached on mobile devices by setting a high level of restrictions. On the other hand, this might minimize user acceptance and satisfaction factors. Thus, there is an opposition between security and usability. A balance between them should be carefully taken into account [19]. Achieving a balance between smart phone effective security countermeasures and employee acceptance is a serious dilemma for CIOs and security professionals [17].

Another important issue to be considered in this context is the types of enterprise data. They are normally classified into private or public data. The importance of private data can be defined by the level of security attached to it [20]. Particularly, with regard to the use of MEAs, experts agree that the boundaries between security levels are not clear in the business sectors. Based on the aforementioned security-related problems, the research behind this paper tries to answer the following research questions:

- To what extent can MEAs be protected?
- Which security level can be applied?
- Which security countermeasures can achieve the applied security levels?
- Which data, under which conditions, can enterprises transfer to mobile devices?
- What are the accompanied consequences or restrictions?
- Will these consequences be accepted by the mobile users?

The following section gives a short overview of the research process followed by this work and briefly explains the main outcomes behind the conducted research.

III. Research Process

This work follows the Information Systems Research Framework that is based on seven guidelines provided by A. Hevner et al. [21]. The work began with the business needs, to define the problem and to ensure that this research meets the goal of relevance. This is achieved by the discussion with the experts in the German Lufthansa systems company. The discussion revealed that enterprises need to execute business processes remotely from the mobile devices in a known level of security. The research relies on existing knowledge base within two main fields namely information security and enterprise mobility. The suitable use of this knowledge ensures the rigor of the research. The expected artifact of this research is a generic framework that helps developers to design secure MEAs. Based on that, this research mainly aims at coming up with a framework to guide developers in designing secure mobile applications. The whole process encompasses conducting risk analysis in the mobile environments and classifying MEAs in security levels considering the user acceptance of the consequences arising in each level. Eventually, using this framework will make the boundaries between these levels as clear as possible. This framework is considered as the artifact behind the conducted research. To evaluate this artifact, it will be firstly implemented as a proof of concept. After that, the resulted prototype will be evaluated descriptively by constructing detailed scenarios around the artifact to demonstrate its utility.

IV. Framework to design secure Mobile Enterprise Applications

To answer the abovementioned research-related questions, risk analysis has been conducted to determine the potential mobile security threats and the applicable security countermeasures which overcome them. As a method to analyze the risks, assessment methodology provided by G. Stoneburner et al. [22] is employed, taking into consideration the following three standards: ISO/IEC 27005 [23], BSI-Standard 100-3 [24] and Risk Management Guide for Information Technology Systems [22]. In the proposed framework, each threat defines a mobile security issue that might be overcome by applying one or more security countermeasures called “alternatives”. The security issues and their alternatives are determined based on literature and best practices. The alternatives define a set of reusable decisions made in previous projects that concern mobile application development. The proposed framework is developed based on Service-Oriented Architecture Decision Modeling (SOAD) framework [25], which aims at enhancing the SOA’s architectural style. To reuse the structure of SOAD framework in security and enterprise mobility domains, adaptations have to be made to come up with a new framework, which introduces a security knowledge base to support developers in designing the Security Concept (SC) of MEAs.

A. Structure Overview

The structure consists of three models namely: the guidance model, the decision model and the meta-model. This structure is depicted in Fig. 1. The framework’s meta-model is instantiated into two models: the guidance model to identify required decisions and the decision model to log the decision that had been made [25]. The relations between these two models are the tailoring and harvesting decision log. These relations are considered similar to those used in the SOAD framework. On the one hand, the tailoring relation initiates the creation of the decision model. This relation represents an activity in which the developer of a MEA selects the relevant security threats and its alternatives (security countermeasures) to build a decision model that forms the security concept of MEA. On the other hand, the harvesting decision log relation is about feeding information regarding the decision (or result) made in the decision model back to the guidance model to get it refined in the next version.

B. The Guidance Model

As illustrated in Fig. 1, this model contains a list of security issues that are already identified during the risk analysis process. A security issue informs the developer that a particular security threat exists and a decision is needed. Each threat is accompanied with its likelihood of occurrence and harm consequences on the enterprise. According to G. Stoneburner et al. [22], three likelihood levels: high, medium, and low are defined based on the threat-source motivation and capability, nature of the vulnerability besides the existence and effectiveness of current security countermeasures. Each security issue has a reference to one or more alternatives along with their consequences on the mobile user and known uses in the previous mobile applications. The mobile user acceptance of those consequences is a very important factor to be considered during MEAs design. Evidently, it is insufficient to use a strong technical solution that enhances the security when such solution doesn’t satisfy the user. User acceptance scale can take one of these five values: strongly accepted, accepted, neutral, rejected, and strongly rejected. This model is enhanced with a security evaluation method to classify MEAs in security levels as well.

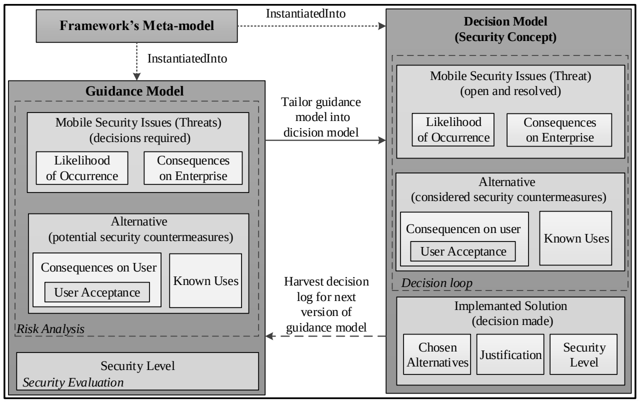


Figure 1. Structure Overview of the proposed Framework (adapted from O. Zimmermann [25])

C. The Decision Model

The decision model is created in a tailoring step, which might involve deleting irrelevant security issues, adding new issues and enhancing relevant ones. After selecting the relevant security threats and one alternative for each threat, the MEAs is evaluated and classified into a specific security level. If the resulted security level does not meet the security requirements, other alternatives can be selected. Therefore, a decision loop is enabled to select other alternatives. Ending the loop means that the decision has been made. This decision (the lower right corner of Fig. 1) contains the chosen alternatives along with justification and security level. After tailoring the guidance model into a decision model, the decision model can feed information about the made decision (result) back to the guidance model in a formal or informal lessons-learned review. The new mobile security issues, which were not considered in the guidance model, besides the enhanced ones could be harvested and integrated back to the guidance model to improve it in the next version.

D. The Meta-model

The meta-model of the proposed framework is shown in Fig. 2 as a UML class diagram. The determined threats during risk analysis are described in the entity Threat and classified into ThreatGroup. Each threat is solved by one or more alternatives which are described in the entity Alternative.

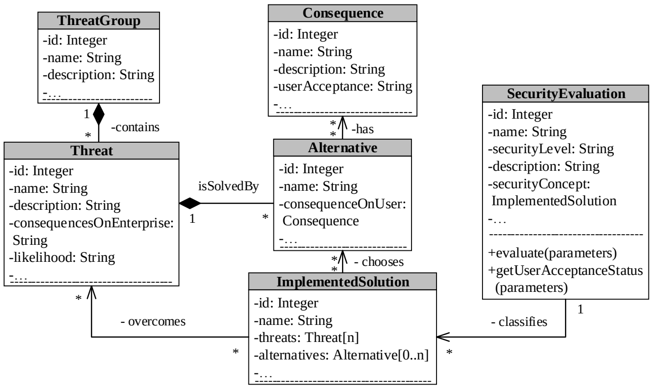


Figure 2. The Framework’s Meta-model

The consequences caused by the alternatives and their user acceptance are described in two entities namely: Consequences and UserAcceptance respectively. The selected threats and their chosen alternative are grouped in the ImplementedSolution entity which represents the security concept of the MEA. At the end, this security concept is classified into security levels provided in SecurityLevel entity.

V. RELATED WORK

Based on an adapted version of SOAD framework, a guidance model to architect secure mobile applications has been created [12]. This model supports decision making process and covers security-related architectural issues during architecting mobile applications. However, in that work, risk analysis is not conducted and the research behind this paper considered such analysis as a vital prerequisite to address and understand all security issues in the mobile environment. A framework to develop MEAs has been presented in [6] to offer a systematic solution for the development and maintenance of mobile application, but it has just highlighted the security as a major concern for the enterprise in developing mobile application without providing more in-depth analysis about it. With regard to security knowledge in the field of software engineering, security patterns are often essential. Security patterns are basically built from best practices and help to solve recurring security problems. However, security patterns don’t support developers in making a proper design decision even if the available patterns can cover all security-related issues [26]. This work addresses such challenge by providing more concrete details about each security-related aspect so that the developers will always have enough arguments to make a proper design decision while designing their secure MEAs.

VI. CONCLUSION AND FUTURE WORK

This paper presented the ways towards building a generic framework to design secure MEAs. Insights to the internal structure of the framework and its building models had been detailed as well. This framework is supposed to provide enterprises and MEAs developers with a security knowledge base needed to comprehend the mobile security issues and their accompanied challenges. Furthermore, this framework will help developers in making proper decisions and keeping a balance between mobile security solutions and user acceptance. Such comprehension tries to make the mobile security issues and challenges as transparent as possible to promote the trustworthy use of mobile technologies in business sectors. The study will be furthered to provide a fully-fledged framework with step-by-step guidelines to show how it works. As a proof of concept, a prototype will be implemented to show the practicability of the overall concept.

REFERENCES

[1] R. Basole and W. Rouse, “Mobile Enterprise Readiness and Transformation,” Idea Group Inc. IGI, 2006.

[2] A. Jain and D. Shanbhag, “Addressing Security and Privacy Risks in Mobile Applications,” IT Professional, 2012, pp. 28–33.

[3] J. Ranjan and V. Bhatnagar, “A holistic framework for mCRM – data mining perspective,” Information Management & Computer Security, 2009, pp. 151–165.

[4] ISEC7 - Mobility for SAP - Mobile SAP. Available: http://www.isec7.com/en/products/mobile-sap, [retrieved: Sep, 2013].

[5] A. Giessmann, K. Stanoevska Slabeva, and B. de Visser, “Mobile Enterprise Applications--Current State and Future Directions,” 45th Hawaii International Conference on System Science (HICSS), 2012, pp. 1363–1372.

[6] B. Unhelkar and S. Murugesan, “The Enterprise Mobile Applications Development Framework,” IT Professional, 2010, pp. 33-39.

[7] H. Hurley, E. Lai, and L. Piquet, Enterprise mobility guide 2011. Sybase, 2011.

[8] K. Detken, G. Diederich, and S. Heuser, “Sichere Plattform zur Smartphone-Anbindung auf Basis von TNC,” D.A.CH Security 2011: Bestandsaufnahme, Konzepte, Anwendungen und Perspektiven; syssec Verlag; Oldenburg, 2011.

[9] ISO/IEC 17799, Information technology - Security techniques - Code of practice for information security management. ISO/IEC, 2005.

[10] W. Copeland and C. C. Chiang, “Securing Enterprise Mobile Information,” Computer, Consumer and Control (IS3C). IEEE, 2012, pp. 80–83.

[11] L. Qing and G. Clark, “Mobile Security: A Look Ahead,” Security & Privacy. IEEE, 2013, pp. 78–81.

[12] W. Schwittek, A. Diermann, and S. Eicker, “A Guidance Model for Architecting Secure Mobile Applications,” in 4th International ICST Conference on Security and Privacy in Mobile Information and Communication Systems, Springer, 2012, pp.12-23.

[13] R. Power, “Mobility and Security: Dazzling Opportunities, Profound Challenges”. Available: http://www.mcafee.com/us/resources/reports/rp-cylab-mobile-security.pdf, [retrieved: Sep, 2013].

[14] BSI, BSI-Standard 100-1: Information Security Management Systems (ISMS). Available: https://www.bsi.bund.de/, [retrieved: Sep, 2013].

[15] P. Rubens, “4 Steps to Securing Mobile Devices and Apps in the Workplace”. Available: http://www.esecurityplanet.com/mobile-security/4-steps-to-securing-mobile-devices-and-apps-in-the-workplace-mdm-byod.html, [retrieved: Sep, 2013].

[16] T. Wright and C. Poellabauer, “Improved Mobile Device Security through Privacy Risk Assessment and Visualization,” Data Engineering Workshops (ICDEW), IEEE 28th International Conference on, 2012, pp. 255–258.

[17] M. Landman, “Managing smart phone security risks,” Information Security Curriculum Development Conference, ACM, 2010, pp. 145-155.

[18] J. Oberheide and F. Jahanian, “When mobile is harder than fixed (and vice versa): demystifying security challenges in mobile environments,” Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications. ACM, 2010, pp. 43-48.

[19] N. Daswani, C. Kern, and A. Kesavan, Foundations of security: What every programmer needs to know. Apress, 2007.

[20] A. Tsolkas and K. Schmidt, Rollen und Berechtigungskonzepte: Ansätze für das Identity- und Access-Management im Unternehmen, 1st ed. Wiesbaden: Vieweg + Teubner, 2010.

[21] A. Hevner, S. March, J. Park, and S. Ram, “Design science in information systems research,” MIS Quarterly, 2004, pp. 75-105.

[22] G. Stoneburner, A. Goguen, and A. Feringa, “SP 800-30. Risk Management Guide for Information Technology Systems.” Technical Report, NIST, 2002.

[23] S. Klipper, “ISO/IEC 27005,” Information Security Risk Management. Vieweg+Teubner, 2011, pp. 63–97.

[24] BSI, BSI-Standard 100-3: Risk analysis based on IT-Grundschutz. Available: https://www.bsi.bund.de/, [retrieved: Sep, 2013].

[25] O. Zimmermann, “Architectural Decisions as Reusable Design Assets,” IEEE Software, 2011, pp. 64–69.

[26] M. Schumacher, Security engineering with patterns: Origins, theoretical models, and new applications. Springer, 2003.

Random Transmission in Cognitive Uplink Network

S. Barman Roy
School of Computer Engineering
Nanyang Technological University
Email: swagato002@ntu.edu.sg

S. N. Merchant
Department of Electrical Engineering
Indian Institute of Technology Bombay
Email: merchant@ee.iitb.ac.in

A. S. Madhukumar
School of Computer Engineering
Nanyang Technological University
Email: asmadhukumar@ntu.edu.sg

***Abstract*—To meet the ever increasing demand for higher data rate, improving spectral efficiency is absolutely essential. Spectrum sharing between several users has been proposed for maximum utilisation of available bandwidth. But whenever multiple users are using the same frequency band and at same time, they are going to interfere with each other resulting in poorer performance as compared to single user scenario. This work proposes a novel scheme for channel capacity improvement in Multiple Access Systems (Uplink Communication) in cognitive radio networks and explores the trade offs involved among the cognitive users and primary user. It is argued that, when the mobile transmitters lack the channel state information, they can't use the broadcast scheduling algorithm to cooperate with each other. Convex-concave properties of the data rate is used to find the appropriate bounds. The corresponding scenario with Broadcast Systems (Downlink Communication) is compared where the transmitter has perfect knowledge of the channel state information.**

Keywords-Broadcast Scheduling; Cognitive Radio; Interference Channel; MIMO Channel; Multiple Access Interference.

I. Introduction

The last decade has seen enormous growth of wireless devices in consumer driven and industrial applications resulting in an exponential demand for data rate through wireless media. Eventually, today's telecommunication infrastructures are severely strained to meet this demand. Particularly, radio frequency is a resource whose availability is limited by physics and hardware technology. Still, a number of survey results, including [1] and [2] have shown that the current spectrum usage is suboptimal from utility point of view. In many cases (paging and amateur radios, for example), the licensed users remain silent rendering the dedicated frequency bands idle. At the same time, the mobile frequency bands are severely overloaded to serve existing users and to meet the increasing demand from new users. To meet the discrepancy between high demand and suboptimal usage, cognitive networking has been gaining popularity in current research.

Two major approaches proposed in cognitive networking (both aiming to improve spectral efficiency) are spectrum *sensing* and spectrum *sharing*. A practical system can use any of the approaches or a combination thereof. In the context, a licensed user is referred as a primary user who got priority access over a spectrum. In addition to the primary user there may be one or more secondary (cognitive) users who will use the spectrum opportunistically.

A spectrum sensing network operates based upon burst nature of primary user transmissions. To ensure that no spectrum band remains idle, many experts (see [3] for example) have advocated dynamic spectrum access where the secondary users will continuously scan for free spectral bands, known as *White Space* and use them for transmission. Certainly, it is necessary to ensure retreat of secondary users once the primary users resume their transmission. Needless to say, sensing the spectral range for a White Space and making correct decision about temporary presence/absence of primary user plays the most important role.

In the paradigm of spectrum sharing, along with the primary user, the secondary users will use the spectrum for their own communication so as to cause minimum degree of interference to each other. The primary user will definitely not have an exclusive right over the spectrum, but cognitive users has to ensure that their harmful interference is kept below a certain threshold so that, in an ideal scenario, the primary user is not even aware of their existence. The description is an extremely generic one and quantitative analysis of interference between users will be dependent upon the specific system itself. In this work, the focus will be on a cognitive uplink network described in Section III.

A. Organisation

This work is organised as follows. Section II gives the background with the current state of the art literature. Section III describes the system model. In Section IV, the proposed technique of random transmission is analysed and the simulation results are given in Section V. Then, Section VI discusses implications of the results and possible applications.

B. Notations

Capital boldface letters stand for matrices and lowercase boldface for vectors. $||\mathbf{v}||$ and $||\mathbf{v}||_1$ give the Euclidean and $\mathcal{L}_1$ norm of a vector $\mathbf{v}$ respectively. $\mathbf{A}^{\mathrm{H}}$ and $\mathbf{A}^{\mathrm{T}}$ respectively denote conjugate transpose and transpose of matrix $\mathbf{A}$. $\mathbf{u} \succ \mathbf{v}$ indicates tuple wise inequality between two vectors valid for each tuple. $\mathbb{R}$ and $\mathbb{C}$ denote the fields of real numbers and complex numbers respectively. $\mathbb{E}(X)$ gives expected value of the random number X and $\mathbb{P}(\mathcal{A})$ gives probability of an event $\mathcal{A}$.

II. Background Work

Multiple Input Multiple Output (MIMO) systems have long been proposed as a way to improve capacity of systems. Effects of multiple antennae (in terms of power allocation and diversity) have been extensively studied in the general context of wireless network. In [4], it is shown that capacity increases linearly with min$\{M, N\}$ where M and N are respectively number of transmit and receive antennae. The flexibility offered by MIMO systems makes it an ideal candidate to meet the challenges of cognitive interference network and various system models have been evaluated in literature. For example, the simplest case of two transmit one receive antenna (a MISO system) has been studied in [5]. A more general approach of user scheduling in a broadcast channel with an objective of throughput maximisation has been undertaken in [6]. The present work concerns with an uplink system model. It will be shown afterwards, the fundamental differences between uplink/downlink models in terms of joint versus distributed receive strategies or centralised power control for downlink will have important implications on performance.

III. System Model

When several users spread throughout a coverage area transmit to a base station, it is called a multiple access channel (MAC). A standard multiple access system model with a cognitive base station (having with M antennae), a single antenna primary base station, K cognitive users and a primary user (each having a single antenna) is shown in Figure 1. The $M \times 1$ channel vector from cognitive user k to the cognitive base station is given by $\mathbf{h}_k, \quad \forall 1 \leq k \leq K$, the channel from primary user to cognitive base station is another $M \times 1$ channel vector $\mathbf{h}_p$. The scalar channels from cognitive user k and primary user to primary base station are given by g_k and g_p respectively. Throughout this work, our assumption is $M \ll K$, implying number of user is much more than number of antenna. Each cognitive user is transmitting a scalar symbol $s_k \in \mathbb{C}$ and the primary user is transmitting a symbol $s_p \in \mathbb{C}$. It is obvious that power usage of cognitive user k is given by $P_k = \frac{|s_k|^2}{T}$ and that of primary user is $P_p = \frac{|s_p|^2}{T}$, where T is the symbol duration. We assume coherent detection with same symbol rates at the receivers. Since the users are transmitting at same frequency and time, there will be interference between them. Additionally, since the symbol rates are same, the symbol itself serves as a measure of power ignoring a constant. So, we will take $P_k = |s_k|^2$ and $P_p = |s_p|^2$ for brevity.

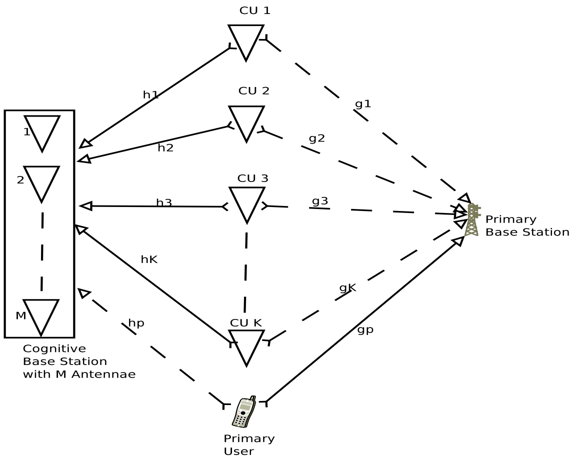


Fig. 1. Multiple Access System Model in Presence of Primary User

A. Multiuser Decoding

With the defined notations, the $\mathbb{C}^{M\times 1}$ vector received at cognitive base station is

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{h}_k s_k + \mathbf{h}_p s_p + \boldsymbol{\eta} \quad \in \mathbb{C}^M \tag{1}$$

and the scalar received at primary base station

$$y_p = \sum_{k=1}^{K} g_k s_k + g_p s_p + \eta \quad \in \mathbb{C} \tag{2}$$

where $\boldsymbol{\eta} = [\eta_1, \eta_2, \ldots, \eta_M]^{\mathrm{T}}$ is the noise vector with independent identically distributed components and η is the scalar noise at the primary receiver.

For brevity again, $\boldsymbol{\eta}$ can be assumed to have a covariance matrix of identity and η also has unit variance. So $\mathbb{E}(\boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{H}}) = \mathbf{I}_M$ (the $M \times M$ identity matrix) and $\mathbb{E}(\eta\eta^*) = 1$. In fact, if the variances are anything apart from unity, the entire expressions of 1 and 2 can be divided by the corresponding factor. That will give rise to the same expression and the constants absorbed in the channel co-efficients.

For the cognitive base station, which receives the vector $\mathbf{y}$, since all the antennae are connected to the same radio front end, joint decoding is possible for each user with a set of K receive strategies. Optimal receive strategies for each of K users can be selected independently [7] with an objective to maximise the individual SINRs. Usually the receive strategy of user k is just to multiply the received signal with a row vector $\mathbf{f}_k$ and the decision statistic is the product $\mathbf{f}_k\mathbf{y}$.

With all the users transmitting simultaneously, the optimal receive strategy of user k can be found using Wirtinger derivative of real functions(see [8]) as $\mathbf{f}_k = \mathbf{A}_k^{-1}\mathbf{h}_k$ where

$$\mathbf{A}_k = \mathbf{I}_M + \mathbf{h}_p\mathbf{h}_p^{\mathrm{H}} p_p + \sum_{i=1, i\neq k}^{K} \mathbf{h}_i\mathbf{h}_i^{\mathrm{H}} p_i \tag{3}$$

Since the primary base station has a single antenna and not connected to the cognitive bases station, its decision statistic is the received scalar y_p. After multiplying with the receive strategies, we get the following SINR expressions for cognitive users.

$$\gamma_k = \frac{|\mathbf{f}_k^{\mathrm{H}}\mathbf{h}_k|^2 P_k}{\mathbf{f}_k^{\mathrm{H}}(\mathbf{I}_M + \mathbf{h}_p\mathbf{h}_p^{\mathrm{H}} P_p + \sum_{i=1, i\neq k}^{K} \mathbf{h}_i\mathbf{h}_i^{\mathrm{H}} P_i)\mathbf{f}_k} \quad \forall k \tag{4}$$

and the primary user SINR is given by

$$\gamma_p = \frac{|g_p|^2 P_p}{1 + \sum_{k=1}^{K} |g_k|^2 P_k} \tag{5}$$

Corresponding maximum data rates (channel capacities normalised by the bandwidth) for all the users users are given by

$$r = \log(1 + \gamma)$$

Albeit the capacity itself is chiefly of theoretical interest, still the logarithmic variation of data rate with SINR is an important measure of performance. In Rayleigh fading AWGN channels, the throughput is closely related to this expression [9]. Also, it intuitively indicates the diminishing marginal return from pumping in more power at high SINR low noise regime and inspires the use of water filling optimisation with power constraint.

B. Comparison with Broadcast Channel

An apparently similar system model has been analysed in [6]. Although the systems are similar in essence, all the communication links are reversed in direction. For the SISO primary user system, this is of relatively little consequence. For the cognitive users, if they have full knowledge of channels (their owns and also other cognitive users), the orthogonal user selection algorithm [6] selects the same set of users with same power to maximise the throughput. The well known rate duality principle [10], [11] dictates that it is possible to achieve the same throughput here.

In practice, however, the channel state information is usually not available at the mobile transmitters. Even if they are available, limited processing capability at the cognitive terminals makes it difficult to schedule the transmissions. If we assume no inter-user point to point communication, the users have absolutely no way to coordinate their behaviours and schedule their own transmissions keeping the broad interest of whole cognitive user group in mind. In Section IV, a novel protocol has been proposed and analysed to address the multiple access channel.

IV. Random Transmission

When there is no coordination among the users, in particular, each user is not aware of others' channel condition, throughput demand or power availability, it is not possible to schedule transmission from a centralised control. Then, if all the users transmit continuously, they not only end up spending a lot of power, at the same time, cause harmful interference to others. The proposed approach consists of random transmission from all the cognitive users. Since the primary user has a single antenna, zero forcing from all K cognitive users is not possible. So it has to tolerate interference to some extent. But, usually for the licensed primary users, the requirement is that the link quality (expressed in terms of SINR) must be above a certain limit. So, it is possible to set up a scheme where every cognitive user will transmit in a particular symbol interval with a probability $p \in (0, 1]$ and stay silent with probability $1 - p$.

To put the scenario in more mathematical form, K independent identically distributed Bernoulli variables $\{b_k\}_{k=1}^{K}$ are introduced and in a particular symbol interval

$$b_k = 1 \Leftrightarrow \text{Cognitive user } k \text{ has transmitted}$$

Since every user is transmitting on a random basis with probability p it is obvious that $\mathbb{P}[b_k = 1] = p$ and $\mathbb{P}[b_k = 0] = 1 - p \quad \forall k \in \{1, 2, \ldots, K\}$.

It is assumed that the primary user is transmitting and using its own spectrum range on a continuous basis and there is no question of probabilistic transmission.

A. Performance Criteria

Since the protocol proposed is not deterministic and there is a degree of randomness involved in the transmission procedure, the performance criteria can no longer be evaluated and/or compared deterministically. From the definitions of $\{b_k\}_{k=1}^{K}$ we can conclude that from the point of view of k-th cognitive user, the interference power vector, received at the base station is

$$\mathbf{I}_k^{\text{cognitive}} = \mathbf{h}_p \mathbf{h}_p^{\mathrm{H}} P_p + \sum_{i=1, i \neq k}^{K} b_i P_i \mathbf{h}_i \mathbf{h}_i^{\mathrm{H}} \tag{6}$$

which is easily interpreted as sum of interferences from all other users (primary and cognitive). Similarly, for the primary user, the interference power is

$$I^{\text{primary}} = \sum_{i=1}^{K} b_i P_i |g_i|^2 \tag{7}$$

The corresponding data rates are, for cognitive users

$$r_k = \log\left(1 + \frac{|\mathbf{f}_k^{\mathrm{H}} \mathbf{h}_k|^2 b_k P_k}{\mathbf{f}_k^{\mathrm{H}} \left(\mathbf{I}_M + \mathbf{I}_k^{\text{cognitive}}\right) \mathbf{f}_k}\right) \tag{8}$$

Here the same receive strategy of 4 is used (assuming the extreme case of all the users transmitting) and the interference is obtained from 6. For primary user, the rate is

$$r_p = \log\left(1 + \frac{|g_p|^2 P_p}{1 + I^{\text{primary}}}\right) \tag{9}$$

Since all the rates are random variables (dependent upon $\{b_k\}_{k=1}^{K}$) to find the true performance measures, expectation must be taken over all the independent variables. So for cognitive users,

$$\begin{aligned} &\mathbb{E}\left[r_k\right] \\ &= \mathbb{E}_{\{b_i\}_{i=1}^{K}}\left[\log\left(1 + \frac{|\mathbf{f}_k^{\mathrm{H}} \mathbf{h}_k|^2 P_k b_k}{\mathbf{f}_k^{\mathrm{H}} (\mathbf{I}_M + \mathbf{I}_k^{\text{cognitive}}) \mathbf{f}_k}\right)\right] \end{aligned} \tag{10}$$

For the primary user, there is no randomness involved in its own transmission, but the interference is random and we get,

$$\begin{aligned} &\mathbb{E}[r_{\text{primary}}] \\ &= \mathbb{E}_{\{b_k\}_{k=1}^{K}}\left[\log\left(1 + \frac{|g_p|^2 P_p}{1 + \sum_{k=1}^{K} |g_k|^2 P_k b_k}\right)\right] \end{aligned} \tag{11}$$

B. Bounds on Rates

For an intuitive idea of how the randomness affects the data rates, properties of the logarithmic function can be used.

Consider the function $f(x,y) = \log\left(1+\frac{y}{x}\right)$ defined for $x > 0$ and $y \geq 0$. It can be shown to be convex in x when y is constant and concave in y for x constant. So, if X and Y are independent random variables (defined on appropriate supports), applying Jensen's inequality (separately as the convex and concave functions of X and Y respectively), gives the bounds on $\mathbb{E}[f(X,Y)]$ as

$$\begin{aligned} &\mathbb{E}_Y\left[\log\left(1+\frac{Y}{\mathbb{E}(X)}\right)\right] \\ &\leq \mathbb{E}_{X,Y}\left[\log\left(1+\frac{Y}{X}\right)\right] \\ &\leq \mathbb{E}_X\left[\log\left(1+\frac{\mathbb{E}(Y)}{X}\right)\right] \end{aligned} \tag{12}$$

1) Cognitive User: To use these inequalities in the context of our multiple access channel, note that, if we set

$$Y = |\mathbf{f}_k^{\mathrm{H}}\mathbf{h}_k|^2 P_k b_k$$

and

$$X = \mathbf{f}_k^{\mathrm{H}}(\mathbf{I}_M + \mathbf{h}_p\mathbf{h}_p^{\mathrm{H}}P_p + \sum_{i=1,i\neq k}^{K}\mathbf{h}_i\mathbf{h}_i^{\mathrm{H}}P_i b_i)\mathbf{f}_k$$

then $\log\left(1+\frac{Y}{X}\right)$ is the random rate of cognitive user k. To find the upper and lower bounds, it suffices to note that according to the above definitions of X and Y,

$$\mathbb{E}[X] = \mathbf{f}_k^{\mathrm{H}}(\mathbf{I}_M + \mathbf{h}_p\mathbf{h}_p^{\mathrm{H}}P_p + \sum_{i=1,i\neq k}^{K}\mathbf{h}_i\mathbf{h}_i^{\mathrm{H}}pP_i)\mathbf{f}_k$$

and

$$\mathbb{E}[Y] = |\mathbf{f}_k^{\mathrm{H}}\mathbf{h}_k|^2 pP_k$$

Define the power allocation vector of cognitive users as

$$\mathbf{p} = [P_1, P_2, \ldots, P_K]^{\mathrm{T}}$$

Obviously, $||\mathbf{p}||_1$ gives total power usage by all cognitive users.

Then our assertion is that for $p \in (0,1]$ and $\mathbf{p} \succ \mathbf{0}$, the following two cognitive uplink systems are identical in terms of power consumption.

- All cognitive users are transmitting continuously with power allocation vector $p\mathbf{p}$.
- Each cognitive user is transmitting with probability p at any symbol interval and the power allocation vector is $\mathbf{p}$.

Based on this assertion, from the first inequality of 12 it is noted that if we replace the random variable X by a deterministic variable $\mathbb{E}[X]$, the rate decreases. In effect, the random variable X corresponds to random transmissions from interfering users and replacing it by $\mathbb{E}[X]$ corresponds to continuous transmissions from interfering users with less power. So, from the inequality itself, it is clear that random transmissions from interfering users is better than continuous transmission so far as tackling the interference goes. But, from the second part of the inequality, it is seen that continuous transmission from user k himself is better so far as its own data rate in concerned. Also, by invoking Jensen's inequality, it can be shown that

$$\begin{aligned} &\mathbb{E}_Y\left[\log\left(1+\frac{Y}{\mathbb{E}(X)}\right)\right] \\ &\leq \log\left(1+\frac{\mathbb{E}[Y]}{\mathbb{E}[X]}\right) \\ &\leq \mathbb{E}_X\left[\log\left(1+\frac{\mathbb{E}(Y)}{X}\right)\right] \end{aligned} \tag{13}$$

A comparison between $\log\left(1+\frac{\mathbb{E}[Y]}{\mathbb{E}[X]}\right)$ and $\mathbb{E}\left[\log\left(1+\frac{Y}{X}\right)\right]$ will depend upon specific values and not possible to carry out in general form. Simulation results in Section V show that this depends upon the power of primary user and for high power primary users it is possible to achieve marginally better performance with the proposed scheme.

2) Primary User: The same form of expressions for $f(X,Y)$ can be used for the primary user. But in this case, the definitions are

$$X = 1 + \sum_{k=1}^{K}|g_k|^2 P_k b_k$$

and $Y = |g_p|^2 P_p$. To be noted here, since the primary user transmits continuously with its own power, Y is not a random variable anymore (in other words, $Y = \mathbb{E}[Y]$). So the only previous inequality for lower bound reduces to

$$\mathbb{E}_X\left[\log\left(1+\frac{Y}{X}\right)\right] \geq \log\left(1+\frac{Y}{\mathbb{E}(X)}\right) \tag{14}$$

So, the primary user is having a clear advantage in tackling the interference from the cognitive users. In light of the simulation results in Section V it will be shown that this advantage can be turned in the favor of cognitive users themselves.

V. Simulation Results

For the simulation, a cognitive system of three antennas at the base station and ten cognitive users is considered. Rayleigh fading is considered with $h_{kj} \sim \mathcal{CN}(0,1)$, $\forall 1 \leq k \leq K$, $\forall 1 \leq j \leq M$ i.e., h_{kj} is a circularly symmetric complex normal variable. With a fixed cognitive users power allocation vector $\mathbf{p}$ and primary user power P_p various data rates have been plotted against variation of probability p (varying from 0 to 1). For comparison, another system is considered where cognitive users transmit continuously with power allocation vector $p\mathbf{p}$. As per the assertion made in Section IV-B1 these two systems are similar in terms of power consumption and the probability p gives the measure of total cognitive user power apart from a constant factor. Figures 2 and 3 show the average rates of all cognitive users, first with a comparatively low primary user power (P_p), and then with a high P_p. Figure 4 gives the corresponding variation of primary user data rate with probability p. From Figures 2 and 3 it is obvious that cognitive users gain in the random transmission scheme if the primary user power is high compared to cognitive users. This assumption is often valid, particularly in wireless sensor networks where sensor motes operating at low power are used for short range communication. As Figure 4 suggests and already shown in Section IV-B2, primary user always gains in terms of data rate.

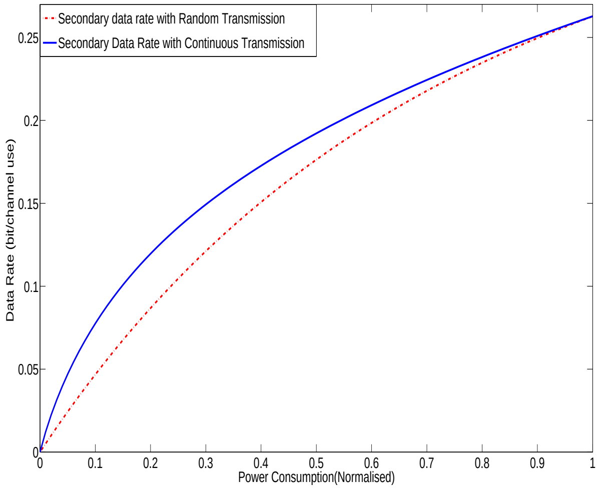


Fig. 2. Average Data Rates of Cognitive Users with Low Primary User Power

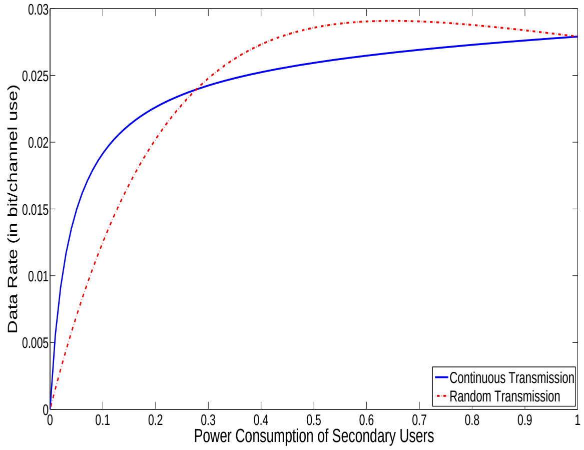


Fig. 3. Average Data Rates of Cognitive Users with High Primary User Power

VI. CONCLUSION

The random transmission scheme is able to outperform the continuous transmission scheme for primary user and possibly the cognitive users as well for certain cases. As the figures demonstrate, tuning the secondary user parameters can be used to limit the data rate of primary users. In certain conditions, higher data rate for primary users may be necessary (with obvious trade-off for cognitive users) and in some other conditions, the primary user can tolerate higher interference from cognitive users. It is shown that the cognitive users can respond to constraints imposed by primary users either by adjusting the actual power or by adjusting the transmission rate p. From the design point of view, controlling the probability p is an easier way than to control the battery power.

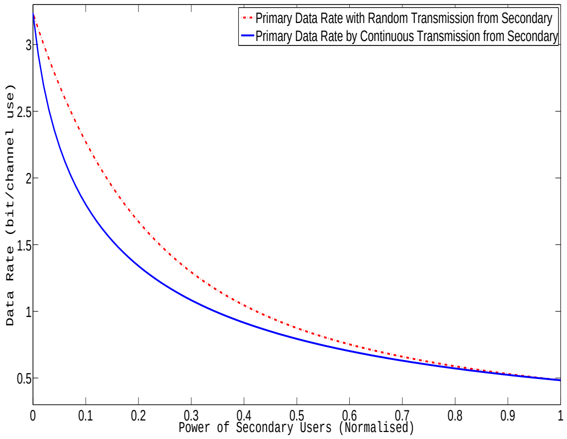


Fig. 4. Primary User Data Rates with Continuous and Discrete Transmissions

REFERENCES

[1] M. Calabrese, “The End of Spectrum ’Scarcity’ : Building on the TV Bands Database to Access Unused Public Airwaves,” *New America Foundation*, 2009.

[2] M. Marcus, J. Burtle, N. Mcneil, A. Lahjouji, and B. Franca, “Report of the Unlicensed Devices and Experimental Licenses Working Group,” Federal Communications Commission Spectrum Policy Task Force, Tech. Rep., 2002.

[3] S. Geirhofer and L. Tong, “Dynamic Spectrum Access in the Time Domain : Modeling and Exploiting White Space,” *IEEE Communications Magazine*, vol. 45, no. 5, pp. 66–72, 2007.

[4] E. Telatar, “Capacity of Multi-antenna Gaussian Channels,” *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.

[5] H. Huang, Z. Zhang, P. Cheng, G. Yu, and P. Qiu, “Throughput Analysis of Cognitive MIMO System,” in *First International Workshop on Cross Layer Design*, Jinan, China, Sep. 2007, pp. 45–49.

[6] K. Hamdi, W. Zhang, and K. B. Letaief, “Opportunistic Spectrum Sharing in Cognitive MIMO Wireless Networks,” *IEEE Transactions on Wireless Communications*, vol. 8, no. 8, pp. 4098–4109, 2009.

[7] M. Schubert and H. Boche, “A Generic Approach to QoS-Based Transceiver Optimization,” *IEEE Transactions on Communications*, vol. 55, no. 8, pp. 1557–1566, 2007.

[8] R. Remmert, *Theory of Complex Functions*, ser. Graduate Texts in Mathematics. Springer-Verlag, 1991.

[9] G. Dimic and N. D. Sidiropoulos, “On Downlink Beamforming with Greedy User Selection: Performance Analysis and a Simple New Algorithm,” *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3857–3868, 2005.

[10] M. Joham, “MIMO Systems,” Associate Institute for Signal Processing, Technological University of Munich, Germany, Tech. Rep., 2009.

[11] S. Vishwanath, N. Jindal, and A. Goldsmith, “Duality, Achievable Rates, and Sum-Rate Capacity of Gaussian MIMO Broadcast Channels,” *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2658–2668, Oct. 2003.

An Empirical Research on InfoSec Risk Management in IoT-based eHealth

Waqas Aman, Einar Snekkenes
Norwegian Information Security Lab (NISLab)
Gjøvik University College
Gjøvik, Norway
{waqas.aman, einar.snekkenes}@hig.no

***Abstract*—Enabling the healthcare infrastructure with Internet of Things (IoT) will significantly improve quality of service, reduce the costs and efficiently manage remote and mobile patients. To be efficacious, IoT and eHealth infrastructure essentials as well as their associated security and privacy issues should be thoroughly recognized to effectively manage the InfoSec risks involved. Unfortunately, there has been a potential lack of research comprehensively addressing these issues jointly while InfoSec risk management solutions are devised for IoT-based eHealth. In this paper, we have highlighted the necessary knowledge while approaching InfoSec risk management in IoT-eHealth as per a standard process, assessed it against standard and proposed requirements and identified the current trends and gaps to set directions for future research.**

***Keywords*—Internet of Things (IoT); Remote Patient Monitoring; Risk Management; Security & Privacy; eHealth.**

I. Introduction

Internet of Things (IoT) is a global internet architecture connecting various wired and wireless technologies designed to meet specific objectives [1]. Beside its anticipated benefits in various private and business domains, enabling IoT in welfare spheres, such as healthcare, will greatly facilitate the society as a whole. A patient can now be monitored remotely in a continuous fashion thus making the health services more mobile, extendable and effective. Though offering a great deal of benefits, IoT is still facing a number of critical challenges such as networking, security and privacy, QoS, standardization, etc., which needs to be sorted out and yet remain open [2]. Among these challenges, the most threatening are the security and privacy concerns. Connecting diverse technologies may lead to new threats with much grander risk of security. These threats become more drastic when considered in the context of a continuous service, such as healthcare, where the concern is not limited to a patient's privacy but, there is a threat to the breach of trust, leading to the exploitation of a welfare service.

Standards, guidelines and good practices concerning InfoSec Risk Management (ISRM), such as ISO 27005, NIST, CRAMM, ISACA RiskIT, etc., recommend to approach ISRM in a methodological fashion, i.e., understand the target business function, service or system, identify the security and privacy (S&P) concerns and threats, analyze the risk faced, and manage the risk to reduce it to an acceptable level. To qualify this process, IoT-driven eHealth as a continuous real-time service will need an intelligent security system that can dynamically predict and estimate the risk faced and mitigating it autonomously to be more resilient and adaptable in the face of changing security threats [3]. A number of architectural designs, security issues, risk management (RM) models and surveys are presented concerning eHealth [1], [2], [4]–[7]. However, such studies are either focused on the mentioned individual topics, target a specific technology or presents abstract modeling. Hence, there is a lack of literature that provides a holistic study of the related topics as per the standard RM process to approach ISRM in IoT-based eHealth.

In this paper, we will highlight IoT-eHealth infrastructure essentials, the associated S&P issues and will explore various ISRM approaches to establish an understanding of how ISRM can be modeled in IoT driven eHealth. Existing literature is evaluated against standard and projected requirements and current trends and gaps are identified. We identified that the current system and S&P modeling are focused only on the primitive requirements and is done in an empirical manner. Whereas vital operations, key system components and necessary S&P services are overlooked. Suitability of various ISRM models and methods is explored and it was concluded that most of them have a subjective influence which makes them difficult to be adopted in a dynamic-real-time environments and lacks intelligent risk analysis and management capabilities, such as context awareness and self-adaptation, which are deemed to be essential for IoT driven eHealth [3]. We strongly believe that this contribution will provide a reference point for future researchers and will enable them to understand the requirements, challenges, options, methods and techniques necessary to consider while approaching ISRM in IoT driven eHealth.

The rest of the paper is organized as follow: In Section II, an overview of the related work will be highlighted. Section III will elaborate the current literature highlighting architectural designs, S&P services, issues and threats and modeling security risks in the perspective of IoT-based eHealth. In Section IV, evaluation of the current literature will be highlighted by aligning them with a set of standard and proposed requirements. In Section V, current trends and gaps will be identified by discussing the evaluated knowledge. Finally, concluding remarks and future research endeavors will be underlined in Section VI.

II. Related Work

A summary of related efforts concerning IoT, remote eHealth, associated security challenges and ISRM modeling are highlighted in this section. The goal is to identify and converse the reviews which are aligned with the theme of ISRM and related topics in IoT-based eHealth.

A detailed description of networking and architectural characteristics of ubiquitous computing used for remote patient monitoring (RPM) is presented in [8]. Sunil et al. discuss the use of mobile networks and the utilization of their mobility features in RPM. 3G and 4G networks characteristics were compared and it was showed that 4G can provide magnified advantages in terms of QoS. QoS requirements concerning wireless networks were highlighted and respective suggestions were discussed to overcome some of the current shortcomings.

Wireless Sensor Networks (WSNs) play a vital role in remote eHealth setup. They enable the notion of a continuous monitoring in remote patient monitoring systems (RPMS). Murad et al. [9] stressed that preventive security measures are not sufficient for WSNs due to the presence of an internal attacker. They provided a comprehensive survey of different intrusion detection systems (IDS) categorizing rule, data mining, statistical and game theoretic based techniques as detective measures to comprehend internal and network attacks dynamically thus enabling a second layer of defense to preventive measures. Similar work is also done in [10] [11].

Latré et al. [12] discussed the importance of Wireless Body Area Network (WBAN) and its applications in remote monitoring of various diseases. Positioning of the WBAN in a RPMS setup is detailed and it is argued that most of the current research is focused on the extra-WBAN communication. Available MAC and Network layer protocols were highlighted and it was suggested that new MAC layer protocols need to be design to accommodate patient mobility. Latré et al. reasoned the current issues like QoS, usability and security are more studied in the WSN and should also be examined in WBAN being a more healthcare focused technology as compared to WSN. The survey however was more emphasized on the networking protocols.

A systematic literature review on S&P issues in an Electronic Health Record (EHR) system is presented in [13]. Literature appraisal was based on the requirements defined in ISO 27799 standard related to achieving security goals through cryptographic techniques, HR security measures, such as training and awareness, and its alignment with compliance and regulatory requirements. Luis et al. concluded that though most of the studies do explicate security controls but are not really implemented in health sectors.

A detail survey on IoT is given in [2]. Atzori et al. explain IoT from three different perspectives: *Things, Internet and Semantics* and converse its overlapping and diverse nature. Different technologies, such as Middleware, WSN, RFID, etc., are recognized to review their possibilities in enabling effective IoT. Extended opportunities of IoT in different application areas are explored and their benefits are traversed. Furthermore, a list of open issues, such as, security, privacy, networking, standardization, QoS and data integrity was highlighted and suggested to be researched to make IoT a more mature and promising technology.

To perform effective risk analysis, it is a difficult task to select the appropriate Risk Analysis (RA) methodology [14]. Vorster and Labuschagne presented a framework of evaluating RM methodologies to assist the business managers in selecting an appropriate method to conduct RA within an organization. A five-point common criterion was used for the comparison. A similar approach is also taken by [15] where RA methodologies were classified based on the involvement of risk analysts or stakeholders and the execution nature of the steps used in the RA process. RA methodologies can also be classified into two groups based on the approach adopted–*Traditional*: where a methodology have a subjective influence of the stakeholder involved and risk is analyzed by the appraisers; *Contemporary*: where risk is estimated based on the target system behavior by inspecting the events it creates, testing it and validating it with formal methods [16].

III. Approaches, Concepts & Issues

This section presents an overview of the current literature in accordance with the standard ISRM process. The selected literature encompasses systems overview, S&P services and threats and ISRM modeling approaches which are necessary to be understood while impending ISRM in IoT driven eHealth. A depiction of the literature organization in line with the standard ISRM guideline is shown in Table I.

TABLE I. Literature Organization & Standard ISRM Process

Standard ISRM Process	Literature Organization
Scope Identification	**IoT-eHealth Infrastructure** – System Overview & Functions – Key Assets – Comm. Medium
S&P Services/Threats	**S&P Modeling** – Threats & Security Services Modeling
Analyzing & Managing Risks	**Modeling InfoSec Risks** – Methods, Models & Frameworks for handling IoT-eHealth Risks

A. *IoT-eHealth Infrastructure*

IoT-based eHealth can be referred to as the global internet of wired and wireless technologies placed to monitor remote and mobile patients. Besides monitoring, patients can also be supervised over the internet and response to emergency situations can be made in a timely manner with the required aid. The infrastructure includes wearable sensors which collect various physiological sensed data from the patient as biosignals, forwards it to a smart device, such as a smartphone or tablet. Biosignals are filtered and are sent to a remote hospital site via mobile network or internet where the medical staff further investigate them and prescribe the patient accordingly. This concept is also portrayed in [17] in which Otto et al.explained a heart patient scenario while presenting their RMPS. A similar model is also described in [18] in which the proposed system, Tele Health Care, is used to monitor blood pressure and heart rate of a remote patient. In abnormal situations the patient is alerted with an alarm and a SMS is sent to the corresponding doctor for instant response. Ambulatory and emergency situations are also discussed. However, Rajan et al. did not discuss the notion of false alerts which may cause panic on both the patient and doctor sides.

Suh et al. proposed a RPMS, *WANDA*, for monitoring congestive heart failure patients [19]. The system is composed of three tiers: sensors, web and back end databases. Mobility is provided through the use of a smartphone carried by a patient. Via Bluetooth the biosignals are transmitted to a smartphone from the sensors and are sent to the second tier through GSM, 3G and/or Internet for further investigations. Health status can be accessed either by using the smartphone or the web services. The database tier is used only for backup and recovery procedures assisted with offline backup schedulers.

Based on the fact that a TV is still the most convenient way of interaction among the older adults, Santos et al. presented a TV based solution, *CareBox*, for RPM [20]. CareBox processes the vital signs only locally. Sensor data is sent to the monitoring unit attached to a TV where the patient can have a look at to his health status displayed on TV. The communication layer of the system is designed to support various protocols and technologies. A VoIP client is used where a patient can connect to a doctor for a video meeting. A survey form is programmed into the TV, which asks health related questions from the patient and can be sent upon submission to the doctor site via an internet connection.

Scacht et al. proposed *Fontane* [21]. In Fontane, medical data sensed by various sensors are transmitted to a home broker via Bluetooth. The home broker, implemented in a smartphone, sends the processed data to a tele-medicine center (TMC) using GSM or UMTS. The live medical data received in the TMC is recorded as the patient's EHR. A J2EE–based SaPiMa module is used at the TMC to ensure EHR interoperability. Medical professionals can access the EHRs via the internet to review health status. Based on specific prioritization rules set by a doctor, the system can review orders for the patients.

Sneha et al. [22] provided a comprehensive set of requirements for RPMS and suggested a three-step framework for RMPS: Sensing the vital signs, Analyzing them and if an anomaly is found, the analysis report is transmitted to the concerned site. A PDA equipped with different agents responsible for various tasks such as location update, collection and processing of vital signs, alarm generation, updating EHR and storage of personal data are utilized. These agents use ontology based on Descriptive Logic (DL) and implements various alerts and alarms as per the patient history. Sneha et al. however, did not discuss the patient-doctor communication within their model.

Wu et al. [23] presented an RIFD based Mobile Patient Monitoring System (MPMS) which they claimed to be the first of RFID driven RPMS. The sensor part of the network is composed of wearable ring-type pulse monitoring tags. The sensed data from the tags are sent to a reader where it is delivered to a smartphone via Bluetooth. The smartphone has the ability to process and analyze the data and anomalies are shared with a remote medical station. The smartphone is also equipped with a GPS, which sends out the patient location to the medical station in case of out-door emergencies. RFID is also used in [24] for an out-patient registration. Though, the title reflects a MPMS but is in-fact a model to facilitate the patient's check-in procedure in the hospital. A patient is registered into the system and an RFID bracelet is given to him. The doctor's PDA connects to the RFID server and retrieves the patient information. After the personal information is read, the corresponding patient history is extracted from the health system and advising is done accordingly.

Van et al. proposed *MobiHealth* system experimented in a number of countries [25]. In MobiHealth, the health information was transferred through the next generation wireless networks. Van et al. argued that beside wearable sensors, devices such as actuators and other wearable devices can also be integrated into the system. MobiHealth, however, was prone to major issues of data loss and low bandwidth drawn from the experiments conducted.

Kargl et al. presented a pervasive eHealth monitoring system, *ReMoteCare* [26]. ReMoteCare consists of a local processing and data collection units, which process and collect local data through sensor motes. The data is then forwarded to a remote or local analysis unit over a communication network through a gateway. A PC is used for local analysis from where analyzed data can be sent to a remote processing and collection unit via SNMP for further investigation.

B. Security & Privacy Modeling

eHealth involves critical information exchange and requires a number of security services to make this information reliable, confidential, available and trustworthy. The objective of this section is to understand the threat landscape, S&P issues and how various security services are modeled in remote/mobile patient monitoring.

RPMSs will no doubt greatly improve the quality of healthcare. However, it still have to face a number of challenges concerning S&P. Meingast et al. [5] discussed the issues concerning data access and storage such as authorization, data retention and the type of data to be stored to meet privacy objectives. Regulatory requirements and conflicts among regulations are also highlighted. They stressed that existing controls such as Role Based Access Control (RBAC), Encryption and Authentication mechanisms should be implemented to overcome these issues.

Extending the notion of threats posed in a MPMS, Leister et al. produced a threat assessment report stating the critical threats faced in an MPM environment using various scenarios [27]. Though, the main focus of the assessment is on the WSNs, they have also considered the long range wireless communication infrastructure and the corresponding threats. They also suggested a few countermeasures and security recommendations which can be considered to circumvent these threats.

A comprehensive analysis of threat faced by the WSNs is presented in [4]. The attacks and threats listed by Kalita and Kar are not specific to eHealth but as WSN plays a vital role in RPMS, these threats should be seriously considered when a secure design or risk analysis of RPMS are intended. The attacks identified are categorized in accordance with the TCP/IP network model so that appropriate measure can be taken at the specified layer. Countermeasures are suggested to avoid some of the common attacks.

Lin et al. presented a privacy protection scheme depicting how patient's privacy can be preserved in an MPMS setup [28]. Lin et al. demonstrated how the privacy of the patient medical information is protected from a global adversary trying to eavesdrop on the messages transferred between the patient and the doctor. Furthermore, they explained the preservation of patient's contextual privacy using the proposed scheme showing that an adversary cannot link a patient to a specific doctor by linking their sources and destinations. They also performed a thorough performance analysis of the proposed scheme demonstrating its efficiency in terms of transmission delays. Ramli et al. [29] provided an insight on four serious privacy issues in pervasive health monitoring systems; eavesdropping, prescription leakages, social implication and abuse of medical information. They argued that these concerns not only affect the health system but also greatly influence patient's life.

Frank et al. described different types of attacks that can be experienced by various network components in RPM as well as the threats corresponding to the information shared between them [26]. They suggested a number of security measures that can be used to prevent internal and external attackers from compromising the confidentiality, integrity and availability of the network components and information. However, privacy and legal issues are just mentioned and are not well elaborated.

Apaporn et al. [30] presented a security framework for eHealth services using two mechanisms: Data and Channel security. Channel security is provided using the SSL on the HTTP layer and data security is provided on the SOAP layer constructed above the HTTP. Apaporn et al. emphasized that RBAC should be used along with multi-factor authentication to ensure proper authorization and authentication. Based on the roles of stakeholders and data sensitivity, communication is divided into different layers where various authentication and encryption settings can be adapted. The framework however dealt only with the web based eHealth services. Multi-factor authentication is also utilized in [31] where Sriram et al. used ECG and accelerometer features from the sensor to perform an activity based biometric authentication.

Elkhodar et al. proposed a Ubiquitous Health Trust Protocol (UHTP) in combination with TLS to authenticate a mobile doctor visiting patients at home [32]. Authentication is performed using three factors based on personal, device and environmental (location) information. During a request to a patient EHR, the doctor uses his smart phone to access the EHR system using his username and password. Beside these personal credentials, the SIM details, IMEI and GPS locations from doctor's phone are validated and access is granted accordingly. The rest of the communication security is ensured as per the TLS negotiated parameters. UHTP, however, doesn't have any application in a continuous RPM orientation.

Simple and secure RPMS is demonstrated in [33]. A mobile set is used as a pulse oximeter where pulse rates are transmitted to a smartphone. The smartphone is equipped with a symmetric cipher and a hashing algorithm to achieve confidentiality and integrity. Shortcomings of this model are ignoring the distribution concerns of the keys and the abstract knowledge of the model, which needs to be detailed.

Timestamps can provide valuable and fresh data for authentication and requires no active involvement of the user [34]. Elmufti et al. used packet timestamps to authenticate a patient/doctor (users) in RPMS. Users are assigned tokens based on timestamps signed by an authenticating server. These stamps are transmitted with individual messages and are compared with a sliding window maintained at the receiving end. User authentication itself is done with digital signature. Elmufti et al. although included sensors in their architecture but did not explore the proposed protocol applications in them.

QoS and event reporting are important requirements in information system. In eHealth, real-time delivery is a must and health status has to be monitored continuously [35]. Rikitake et al. presented an NGN/IMS based ubiquitous health monitoring system in which they addressed the issues of event notification, real-time transfer and data accumulation. Sensor's data is sent to an IMS Client from where it is sent to the observer's site using Realtime Transfer Protocol (RTP). For event notification a SIP base Subscribe/Notify module is utilized that records incidents in an event server connected to the hospital application server. An XML database management system (XDMS) is used that extracts the events from the event server and stores it in an XML format.

Malhotra et al. used Elliptic Curve Cryptography (ECC) to secure the exchange of medical data using mobile devices [36]. Basic ECC methods are used where encryption is done at the user level with a public-private key pair. User is authenticated through a username/password terminal and access to the data is granted based on the user (patient/doctor) role. ECC based digital signature to ensure non-repudiation while message integrity is provided through a cryptographic hash.

C. Modeling the InfoSec Risk

To detect and prevent accidental events regarding a patient's health, an activity based risk analysis framework is proposed in [6]. Collected vital signs events are matched with the patient's history already stored as EHR and the current situation of the patient is predicted. Based on the prediction, risk is calculated and an alert is generated to cope up with the situation. The proposed architecture, although only address the patient health, can be extended to the information security domain as a reference when modeling InfoSec risk analysis is desired.

There are several studies on general S&P issues in eHealth comprising ubiquitous systems. However, it is quite hard to understand and systematically listing down these key issues and design a risk mitigation strategy for them. Oladimeji et al. [37] proposed a framework to model security and privacy objectives, identifying threats and risks and approaching their mitigation strategies. They also discussed how information sensitivity can be characterized as well as how different administrative policies can be refined to protect the patient's privacy.

The attributes that are used to design IT solutions specifically in eHealth are usually complex and interdependent thus needed to be analyzed and prioritize to produce a reliable and trustworthy solution. In [38], it is discussed how these critical attributes and their inter-dependencies can be assessed to reduce the risk after the solution has been deployed. The study can be used for formulating the requirements of designing an automated or real-time risk analysis model as it discuss both the quality and security issues at the requirement engineering level.

Bønes et al. proposed *ModIMob*, a model which can be used to discover the availability of the health experts where their presence is required for an expert opinion [39]. The Australian and New Zealand standard for RM (AS/NZS

4360:1999) is used to discover the risks associated with the use of IM and mobile services used in a healthcare. Though, the scope of their risk evaluation is limited to a specific domain of instant messaging but it can provide an understanding of conducting a RA process in a RPMS.

Abie and Ilangko proposed a risk based adaptive framework for IoT-based eHealth [3]. They argued that based on the real time data collected from the sensors and recent information history, a risk will be calculated, which will further be used in the decision making process of system adaptation. They also provide a detailed literature on various issue concerning system adaptation and risk management and it is deemed that using context awareness and Game Theory techniques, the faced risk can be effectively estimated and predicted.

To provide an appropriate level of privacy all the assets as well as the stakeholders involved in the target system must be considered [7]. A Privacy Risk Model is demonstrated specifically targeting the Ubicomp systems where risks concerning privacy are identified and analyzed by a series of questions. RM is performed by categorizing the risks analyzed and designing architectural strategies for them.

Maglogiannis et al. presented a detail risk analysis of RPMS. RA is performed through the CCTA Risk Analysis and Management Methodology (CRAMM) by considering a case study highlighting the associated key risks [40]. The results of the RA are used in developing a graph using Bayesian Network technique showing the interaction of various critical events that can cause system failure.

Beside the risk posture of the sensitive information processed by the health information systems, the devices used in healthcare have their own inherited risks. With the introduction of pervasive computing and IoTs this risk has grown rapidly. Zhao and Bai described how Failure Mode and Effect Analysis (FMEA) can help in analyzing and managing the risks associated with these devices to circumvent any potential hazards [41]. They showed that Risk Priority Number (RPN) can be used in the context of FEMA to reduce such potential casualties associated with medical devices.

ENISA, using EBIOS tool, performed a detailed RM process of a diabetes case study basing a RPMS [42]. EBIOS is a tool that incorporates the 5-steps RM process developed by the Central Information Systems Security Division of France. The report described a detailed step-by-step procedure of assessing and managing risks indicating the intended audience how to approach the overall process of risk management in MPMSs.

IoT comprises of a complex architecture composed of a variety of technologies due to which the overall threat faced becomes more drastic. There is a need for a sophisticated risk analysis method to assess the risk faced. Lui et al. proposed a mathematical dynamic risk assessment model, DRAMIA, to cope with the threat situation confronted in the IoT space [43]. Enthused by the Artificial Immune System (AIS) their proposed method consist of two components: a Detection Agent that sense and detect the attack environment and evolve accordingly; and a Dynamic Risk Assessment subsystem that computes the risk associated with the attack detected.

IV. Evaluation

In this section, an evaluation of discussed literature is depicted. Evaluation is performed by mapping the reviewed articles onto a set of standard and proposed requirements.

A. System Models Evaluation

System models discussed in section III-A are evaluated against a set of functional requirements proposed in [44]. We believe that these are complete set of requirements which should be included in any RPMS and MPMS. However, we have added an important requirement of *mobility* as it is the only component that makes the health service mobile and assist in out-doors ambulatory and activity monitoring needs [22]. Functional requirements are described below whereas system models evaluation against the requirements is shown in Table II. ($\surd$) mark indicates the presence of a specific function whereas an (–) implies that either the function is absent or not explicitly discussed.

- **Collection and Processing**: Collection and processing of vital signs from the body sensors by a Patient Cluster Head (PCH) or a wireless base station (BS)
- **Real Time Delivery**: PCH or BS should be able to deliver the processed data in real time for analysis to specified destination such a remote hospital site or a smart device.
- **Alarm generation**: The investigating node, a server at hospital or the smart device, should be able to generate alarms based on the real time data received both locally and remotely at hospital.
- **Interpretation**: Local and remote investigating nodes should be able to diagnose and interpret processed vital sign.
- **Correlation**: Local and remote investigating nodes should be cable of correlating various vital sign such as heart rate, diabetes level and blood pressure to diagnose the correct health status
- **Data Request**: Patient health history should be made available whenever requested
- **Communication Interface**: A communication interface should be incorporated locally to enable expert supervision for a remote patient.
- **Actuation**: To assist elder patients or on demand basis sensors or actuators should be able to saturate the essential medicine or trigger the required action.
- **Mobility**: The system should be able to support mobility services to the patient. This includes tracking the location and service availability while the patient is moving.

B. Evaluating S&P Modeling

S&P service modeling literature reviewed in section III-B is evaluated against the networking and communication requirements standardized by the U.S Health Insurance Portability and Accountability Act (HIPAA) of 1996 specified in [45].

TABLE II. IoT-based eHealth Systems Evaluation

Function/ Reference	[19]	[26]	[20]	[17]	. [25]	[18]	[21]	[22]	[24]	[23]
Collection & Processing	√	√	√	√	√	√	√	√	√	√
Realtime Delivery	√	√	√	√	√	√	√	√	√	√
Alarm Gen.	–	√	–	–	–	√	–	√	√	√
Interpretation	√	√	√	√	√	√	√	√	√	√
Correlation	–	–	–	√	–	–	–	√	–	√
Data Request	√	–	–	–	√	√	–	–	√	–
Comm. Interface	–	–	√	–	–	–	–	–	–	–
Actuation	–	–	–	–	√	–	–	–	–	–
Mobility	–	√	–	–	√	–	–	√	–	√

Besides, ensuring health insurance coverage and simplification of administrative policies, HIPAA aims to standardize S&P mechanisms for electronic health information exchange. Requirements stated by HIPAA are: Data Access, Confidentiality, Integrity, Availability, Alarm Generation, Identity Management, Privacy Preservation, Authentication, and Event Reporting. Table III depicts the requirement(s) covered by each study as per HIPAA security requirements and how they are approached in individual study.

C. InfoSec RM Models Suitability

IoT-based eHealth is a continuous service in which response to an adverse situation should be made in a dynamic fashion. Hence, it requires an ISRM solution that can estimate and predict the security risk faced in real time and adapts appropriate security setting accordingly [3]. To capture the requirements of a real time RM in IoT-eHealth below we devise a fitness criteria, which we believe should be met by a given ISRM model in order to fulfill the operational needs of IoT-eHealth and efficiently manage the risk faced.

- **Operational Nature:** The time at which an ISRM process is executed. This can be *on-demand* basis where the process is activated when required. For instance, ISACA Risk IT method can be executed bi-annually or quarterly by an enterprise. ISRM can be performed in a *dynamic* manner where security risks are analyzed in a real-time fashion such as in military setups. For IoT driven eHealth the operational level should be dynamic in order to be in line with the continuous monitoring theme.
- **Context Awareness:** It corresponds to the understanding of an adverse situation in a given time. In most cases, risks are analyzed individually however; in real computing environment, risk can be seen as a combination of different adverse events. These events and risks need to be correlated to understand a given situation otherwise low impact risks might be tagged as critical leading to false positives and unwanted situations.
- **Analysis Complexity:** It should be taken care of that risk analysis method is lightweight and fast in response to facilitate the theme of real time service [3]. RA solutions having low computational complexity can also be integrated in devices with limited resources.
- **Self-Adaptation:** For IoT-eHealth to be dynamic and self-adaptive, an ISRM should have the ability to react to an adverse situation and manage security autonomously. Self-adaptation refers to the autonomous effective reaction of a system to minimize the effect of a risky situation [3].

In Table IV, we evaluate the suitability of the studied ISRM approaches against the above mentioned metrics to see how they address these metrics in order to be implemented in IoT-based eHealth.

V. Trends And Gaps

Key elements of ISRM concerning IoT-eHealth are reviewed in this paper as system, S&P and InfoSec risk modeling and are evaluated as per projected requirements. The objective was to understand and recognize the essential operations, S&P challenges and methodologies for effective ISRM in IoT driven eHealth. A brief discussion on the evaluated knowledge corresponding to individual domains is conferred below to reflect the current trends and gaps in the existing literature.

– System Models
A total of 10 models are studied and analyzed according to the required features in a RPMS or IoT driven eHealth. Some of the models reviewed are focused on monitoring generic vital signs such as ECG, Blood pressure and heart rate [21], [25] while a few targets specific heart [17], [19] and chronic diseases such as diabetes [22]. Systems corresponding to [17], [19], [21], [22], [25] emphasized the use of cellular network (GPRS, GSM and UMTS) for the transmission of sensed data to the hospital site through the use of smart phones. However, simultaneous transmissions on cellular networks can cause performance degradation and may affect continuous monitoring in critical situations [25]. Except for [22], the importance of local analysis of sensed data is ignored in the rest of the models, which enable a patient to view his health status locally and schedule the daily routines accordingly. Similarly, actuation of medical infusions is also overlooked. A vital functionality of RPM is to diagnose the patient at home to save the time and energy spent in regular checkups, i.e., the provisioning of communication interface between a doctor and patient however, an absence is experienced of this feature in most of the systems reviewed. Those who support this functionality did not explicate it in detail. Santos et al. [20] on the other hand fairly explained a patient-doctor communication over a VOIP client, which can also be used in calling the health facilities in case of emergencies as well. Alarm generation is merely explored, except for [22] who detailed each alarm as per the assigned agent's responsibilities.

It can be seen that most of the system models are focused on the basic functionalities of collection, processing and delivery of vital signs to the remote hospital site. Analysis and correlation of various bio-signals are limited to the server side, which is needed to be shifted to the patient side to increase

TABLE III. MAPPING S&P REQUIREMENTS ONTO HIPAA

Author	Data Access	Confidentiality	Integrity	Availability	Alarm Gen.	Identity Mgt	Privacy	Authentication	Event Rep.
Lin et al. [28]	–	Symmetric Encryption	Hash	–	–	PKI based on Patient IDs	Symmetric Encryption & Pseudo ID	Shared Key	–
Apaporn et al. [30]	RBAC	Symmetric Encryption	–	–	–	–	–	Muliti-factor	–
Elkhodar et al. [32]	–	–	–	–	–	–	–	Multi-factor	–
Mona et al. [33]	–	Symmetric Encryption	SHA–1	–	–	–	–	Message Authentication Code(MAC) based on a Secret key	–
Khalid et al. [34]	–	–	–	–	–	–	–	User: Digital Signatures Message: Timestamps	–
Koichiro et al. [35]	AAA over NGN/IMS	–	–	Realtime Transfer Protocol (RTP)	–	–	–	AAA over NGN/IMS	SIP Event Subscribe/Notify Framework
Sriram et al. [31]	–	–	–	–	–	–	–	Multi-modal Biometrics	–
Malhotra [36]	RBAC	ECC	SHA-1	–	–	–	–	ECC based Digital Signatures	–

TABLE IV. ISRM APPROACHES SUITABILITY IN IOT-BASED EHEALTH

Author	Artifact	Analysis Method	Operational Nature	Context Awareness	Complexity	Self Adaptation
Don et al. [6]	Framework	Quantitative Analysis of patient activities	Dynamic	Event Correlation		No
Croll et al. [38]	Framework	Qualitative Investigation of Quality, Usability, Privacy and Safety (QUPS) Attributes	Dynamic & OnDemand	Investigating interdependent critical attributes and events		No
Hong et al. [7]	Model	Qualitative Assessment based on a questionnaire	On-Demand	No		No
Liu et al. [43]	Method	Quantitative RA based on attack detection in network packets using Artificial Immune System	Dynamic	No	Attack detection & RA are done by specific agents	Adaptation is performed only to enhance the detection capabilities.No mechanism of adapting a RM strategy
Maglog et l. [40]	Case Study	Threat Identification is performed using Bayesian Network Modeling whereas CRAMM is used as a RA method	On-Demand	Event dependencies are used to build the context of a specific threat	Unclear to evaluate the actual computation complexity just on the graphical model presented	No
Nes et al. [39]	Case Study	Methodology: Australian & New Zealand Standard for RM ASNZS 43601999. Qualitative Approach is used in the RA process	On-Demand	No		No
Abie et al. [3]	Framework	Monitor, Analyze & Adapt loop.	Dynamic	Game Theory & Context Awareness		Yes
Zhao et al. [41]	Model	Methodology: FMEA . Risk (RPN) is analyzed using Severity, Occurrence & Detection (SOD) values	On-Demand	No	Low: RPN = SxDxO	No
ENISA [42]	Case Study	Qualitative 5-Step EBIOS RA Methodology: Formulating Risk, Asset Valuation, Probability Calculation, Impact Valuation & Prioritizing Risk Levels	On-Demand	No	Low- Risk Calculation: Risk = (Threat x Vulnerability x Impact)	No

patient satisfaction. Mobility features should be well designed to support both in and out door patient and to facilitate ambulatory services [22]. Security and safety alarms are needed to be designed intelligently to support critical patient monitoring. Communication interface and GUIs needed to be constructed in order to enrich a patient-doctor relationship and trust.

– Security & Privacy

Among the HIPAA required services for secure remote and mobile patient monitoring systems, the most addressed are the confidentiality and authentication. However, none of them addresses all the HIPPA requirements. Our objective here is not to criticize this fact but to recognize how these requirements can be approached and to identify the current focus of S&P modeling and the necessary issues to be explored in future.

In most of the literature, Symmetric encryption is used to attain confidentiality [28] [30] [33]; however, asymmetric encryption using ECC is also explored [36]. Multi-factor authentication is used in a few studies where passwords, SIM credentials, GPS location, ID cards [30], [32] and vital signs (as biometrics) such as ECG and heart beats are used as various factors of authenticating patients and doctors [31]. Digital signatures are also used in authentication [34], [36]. Message authenticity is achieved by using packet timestamps and message authentication code [33], [34]. Hashing remained the only method of ensuring message integrity however, discussed by only a few [28], [33], [36]. Anonymity is only discussed in [28], where pseudo patient IDs are used to ensure identity privacy against global eavesdropping. Authorization through RBAC are conversed in [30], [36] but are not explicitly defined.

Some of the security services in a continuous RPM such as event reporting, alarm generation and availability are yet to be researched. These are the services which are used in real-time delivery and emergency situations and are the key attributes of RPMS. Most of the literature summarized targets the extra-Body Area Network (Ex-BAN) security, which includes traditional web services and back end database resources in eHealth. As per our knowledge, there is a very limited literature available on securing inter-BAN communication specific to medical information exchange. Research is necessary to be done to secure these networks as they are the core producers of the medical information in an IoT-based eHealth or RPM. Also, the resources used in such networks have limited capabilities thus there is a need to design lightweight cryptographic solutions as discussed in [36] to be aligned with sensors computational competencies.

– InfoSec RM Models

Managing InfoSec risk in IoT-based eHealth is a tough task because of the diverse nature of technology utilized in it. The evaluation of the studied literature in context of ISRM reveals that almost all of them can be used in an On-Demand basis most of which are analyzing the risk on qualitative grounds [7], [39]–[42]. This is because of the subjective influence in RM process which makes it stiffer to be adapted in a dynamic environment. Those that can be executed in dynamic setups are suggested frameworks [3], [38] and still needs a keen and defined method of quantitative risk analysis. Liu et al. [43] on the other hand provide an effective method for analyzing the risk in a real time manner on a quantitative basis, which make it easier to program and usable for IoT-based eHealth. It also includes intelligent agents to adapt its attack detection capabilities and requires fewer resources as the threat detection and analysis is performed by specific agents. However, the suggested techniques are based on the inputs from signature based IDS, which makes it to generate false positives [9]. Self-adaptation as a risk management strategy is completely absent and needed to be designed intelligently to make IoT-eHealth an autonomous technology.

IoT-based eHealth needs quantitative methods for predicting and estimating threats in a dynamic fashion and should be capable of understanding and analyzing the threat situation and transforming the system security autonomously [3]. Some of the methods and framework discussed such as [3], [6], [43] can be utilized as a reference point to design the desired InfoSec RM methods for IoT driven eHealth.

VI. Conclusion and Future Work

In this paper, we have explored the existing literature in the context of approaching InfoSec risk management in IoT-based eHealth. A common knowledge of RMPSs, S&P issues, security and risk management modeling was established in the light of a standard risk management process. System models are evaluated against a set of required functionalities, models pertaining to security services are aligned with the standard HIPAA requirements and existing RM approaches in the context of IoT driven eHealth are weighed against a fitness criteria. An overall analysis is discussed and current trends and gaps are identified.

Our future work includes devising lightweight real-time InfoSec RM methods for IoT-based eHealth with the abilities of context awareness and self-adaptation. An adaptive security model will be developed that will address the mentioned InfoSec RM requirements. Security metrics and options necessary for the adaptation will be explored. To analyze the foreseen risk, Game and Utility theory will be used to model the dynamic and expected behaviors of adversaries and a comprehensive case study will be formulated to validate the model.

Acknowledgments

The work presented in this article is a part of the ASSET (Adaptive Security in Smart IoT in eHealth) project. ASSET (2012-2015) is sponsored by the Research Council of Norway under the grant agreement no: 213131/O70. Wishing thanks to the project colleagues and anonymous reviewers for their valuable suggestions and comments.

References

[1] R. H. Weber, "Internet of things: New security and privacy challenges," *Computer Law & Security Review*, vol. 26, no. 1, pp. 23 – 30, 2010.

[2] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787 – 2805, 2010.

[3] H. Abie and I. Balasingham, "Risk-based adaptive security for smart iot in ehealth," in *Proceedings of the 7th International Conference on Body Area Networks*, ser. BodyNets '12. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2012, pp. 269–275.

[4] H. K. Kalita and A. Kar, "Wireless sensor network security analysis," *International Journal of Next-Generation Networks (IJNGN),*, vol. 1, pp. 1–10, December 2009.

[5] M. Meingast, T. Roosta, and S. Sastry, "Security and privacy issues with health care information technology," in *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE*, 2006, pp. 5453–5458.

[6] S. Don, E. Choi, and D. Min, "A situation aware framework for activity based risk analysis of patient monitoring system," in *Awareness Science and Technology (iCAST), 2011 3rd International Conference on*, 2011, pp. 15–19.

[7] J. I. Hong, J. D. Ng, S. Lederer, and J. A. Landay, "Privacy risk models for designing privacy-sensitive ubiquitous computing systems," in *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, ser. DIS '04. New York, NY, USA: ACM, 2004, pp. 91–100.

[8] S. Kumar, K. Kambhatla, F. Hu, M. Lifson, and Y. Xiao, "Ubiquitous computing for remote cardiac patient monitoring: a survey," *Int. J. Telemedicine Appl.*, vol. 2008, pp. 3:1–3:19, Jan. 2008.

[9] M. A. Rassam, M. Maarof, and A. Zainal, "A survey of intrusion detection schemes in wireless sensor networks," *American Journal of Applied Sciences*, vol. 9, no. 10, p. 1636, 2012.

[10] C. Karlof and D. Wagner, "Secure routing in wireless sensor networks: attacks and countermeasures," in *Sensor Network Protocols and Applications, 2003. Proceedings of the First IEEE. 2003 IEEE International Workshop on*, 2003, pp. 113–127.

[11] D. Christin, P. S. Mogre, and M. Hollick, "Survey on wireless sensor network technologies for industrial automation: The security and quality of service perspectives," *Future Internet*, vol. 2, no. 2, pp. 96–125, 2010.

[12] B. Latré, B. Braem, I. Moerman, C. Blondia, and P. Demeester, "A survey on wireless body area networks," *Wirel. Netw.*, vol. 17, no. 1, pp. 1–18, Jan. 2011.

[13] J. L. Fernndez-Alemn, I. C. Seor, P. ngel Oliver Lozoya, and A. Toval, "Security and privacy in electronic health records: A systematic literature review," *Journal of Biomedical Informatics*, no. 0, pp. –, 2013.

[14] A. Vorster and L. Labuschagne, “A framework for comparing different information security risk analysis methodologies,” in *Proceedings of the 2005 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*, ser. SAICSIT ’05. Republic of South Africa: South African Institute for Computer Scientists and Information Technologists, 2005, pp. 95–103.

[15] P. L. Campbell and J. E. Stamp, “A classification scheme for risk assessment methods,” last Accessed On: 13-Sept-2013. [Online]. Available: http://prod.sandia.gov/techlib/access-control.cgi/2004/044233.pdf

[16] E. Paintsil, “Taxonomy of security risk assessment approaches for researchers,” in *Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on*, 2012, pp. 257–262.

[17] C. Otto, A. Milenković, C. Sanders, and E. Jovanov, “System architecture of a wireless body area sensor network for ubiquitous health monitoring,” *J. Mob. Multimed.*, vol. 1, no. 4, pp. 307–326, Jan. 2005.

[18] R. S. Rajan, S.P. and S. Vijayprasath, “Design and development of mobile based smart tele-health care system for remote patients,” *European Journal of Scientific Research*, vol. 70, p. 148158, 2012.

[19] M.-K. Suh, C.-A. Chen, J. Woodbridge, M. K. Tu, J. I. Kim, A. Nahapetian, L. S. Evangelista, and M. Sarrafzadeh, “A remote patient monitoring system for congestive heart failure,” *J. Med. Syst.*, vol. 35, no. 5, pp. 1165–1179, Oct. 2011.

[20] A. Santos, R. Castro, and J. Sousa, “Carebox: A complete tv-based solution for remote patient monitoring and care,” in *Wireless Mobile Communication and Healthcare*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, B. Godara and K. Nikita, Eds. Springer Berlin Heidelberg, 2013, vol. 61, pp. 1–10.

[21] A. Schacht, R. Wierschke, M. Wolf, M. von Lowis, and A. Polze, “Live streaming of medical data - the fontane architecture for remote patient monitoring and its experimental evaluation,” in *Object/Component/Service-Oriented Real-Time Distributed Computing Workshops (ISORCW), 2011 14th IEEE International Symposium on*, 2011, pp. 306–312.

[22] S. Sneha and U. Varshney, “Enabling ubiquitous patient monitoring: Model, decision protocols, opportunities and challenges,” *Decision Support Systems*, vol. 46, no. 3, pp. 606 – 619, 2009.

[23] Y.-C. Wu, P.-F. Chen, Z.-H. Hu, C.-H. Chang, G.-C. Lee, and W.-C. Yu, “A mobile health monitoring system using rfid ring-type pulse sensor,” in *Dependable, Autonomic and Secure Computing, 2009. DASC ’09. Eighth IEEE International Conference on*, 2009, pp. 317–322.

[24] I. Korkmaz, C. Atay, and G. Kyparisis, “A mobile patient monitoring system using rfid,” in *Proceedings of the 14th WSEAS international conference on Computers: part of the 14th WSEAS CSCC multiconference - Volume II*, ser. ICCOMP’10. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2010, pp. 726–732.

[25] A. T. van Halteren, R. G. A. Bults, K. E. Wac, D. Konstantas, I. A. Widya, N. T. Dokovski, G. T. Koprinkov, V. M. Jones, and R. Herzog, “Mobile patient monitoring: The mobihealth system,” *The Journal on Information Technology in Healthcare*, vol. 2, no. 5, pp. 365–373, October 2004.

[26] F. Kargl, E. Lawrence, M. Fischer, and Y. Y. Lim, “Security, privacy and legal issues in pervasive ehealth monitoring systems,” in *Mobile Business, 2008. ICMB ’08. 7th International Conference on*, 2008, pp. 296–304.

[27] W. Leister, H. Abie, A.-K. Groven, T. Fretland, and I. Balasingham, “Threat assessment of wireless patient monitoring systems,” in *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on*, 2008, pp. 1–6.

[28] X. Lin, R. Lu, X. Shen, Y. Nemoto, and N. Kato, “Sage: a strong privacy-preserving scheme against global eavesdropping for ehealth systems,” *Selected Areas in Communications, IEEE Journal on*, vol. 27, no. 4, pp. 365–378, 2009.

[29] S. P. Ramli Rusyaizila, Zakaria Nasriah, “Privacy issues in pervasive healthcare monitoring system: A review,” *World Academy of Science, Engineering & Technology*, vol. 72, p. 741, 2011.

[30] A. Boonyarattaphan, Y. Bai, and S. Chung, “A security framework for e-health service authentication and e-health data transmission,” in *Communications and Information Technology, 2009. ISCIT 2009. 9th International Symposium on*, 2009, pp. 1213–1218.

[31] J. C. Sriram, M. Shin, T. Choudhury, and D. Kotz, “Activity-aware ecg-based patient authentication for remote health monitoring,” in *Proceedings of the 2009 international conference on Multimodal interfaces*, ser. ICMI-MLMI ’09. New York, NY, USA: ACM, 2009, pp. 297–304.

[32] M. Elkhodr, S. Shahrestani, and H. Cheung, “An approach to enhance the security of remote health monitoring systems,” in *Proceedings of the 4th international conference on Security of information and networks*, ser. SIN ’11. New York, NY, USA: ACM, 2011, pp. 205–208.

[33] M. Kamel, S. Fawzy, A. El-Bialy, and A. Kandil, “Secure remote patient monitoring system,” in *Biomedical Engineering (MECBME), 2011 1st Middle East Conference on*, 2011, pp. 339–342.

[34] K. Elmufti, D. Weerasinghe, M. Rajarajan, V. Rakocevic, and S. Khan, “Timestamp authentication protocol for remote monitoring in ehealth,” in *Pervasive Computing Technologies for Healthcare, 2008. PervasiveHealth 2008. Second International Conference on*, 2008, pp. 73–76.

[35] K. Rikitake, Y. Araki, Y. Kawahara, M. Minami, and H. Morikawa, “Ngn/ims-based ubiquitous health monitoring system,” in *Consumer Communications and Networking Conference, 2009. CCNC 2009. 6th IEEE*, 2009, pp. 1–2.

[36] K. Malhotra, S. Gardner, and R. Patz, “Implementation of elliptic-curve cryptography on mobile healthcare devices,” in *Networking, Sensing and Control, 2007 IEEE International Conference on*, 2007, pp. 239–244.

[37] E. A. Oladimeji, L. Chung, H. T. Jung, and J. Kim, “Managing security and privacy in ubiquitous ehealth information interchange,” in *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*, ser. ICUIMC ’11. New York, NY, USA: ACM, 2011, pp. 26:1–26:10.

[38] P. R. Croll and J. Croll, “Investigating risk exposure in e-health systems,” *International Journal of Medical Informatics*, vol. 76(5-6), pp. 460–465, 2006.

[39] H. E. S. T. Bnes E, Hasvold P, “Risk analysis of information security in a mobile instant messaging and presence system for healthcare,” *International Journal of Medical Informatics*, vol. 76(9), pp. 677–687, 2007.

[40] I. Maglogiannis, E. Zafiropoulos, A. Platis, and C. Lambrinoudakis, “Risk analysis of a patient monitoring system using bayesian network modeling,” *J. of Biomedical Informatics*, vol. 39, no. 6, pp. 637–647, Dec. 2006.

[41] X. Zhao and X. Bai, “The application of fmea method in the risk management of medical device during the lifecycle,” in *e-Business and Information System Security (EBISS), 2010 2nd International Conference on*, 2010, pp. 1–4.

[42] ENISA, “Being diabetic in 2011 - identifying emerging and future risks in remote health monitoring and treatment,” Technical Publication on ENISA website, 2009, last Accessed On: 13-Sept-2013. [Online]. Available: http://www.enisa.europa.eu/publications/archive/being-diabetic-2011/

[43] C. Liu, Y. Zhang, J. Zeng, L. Peng, and R. Chen, “Research on dynamical security risk assessment for the internet of things inspired by immunology,” in *Natural Computation (ICNC), 2012 Eighth International Conference on*, 2012, pp. 874–878.

[44] G. Paliwal and A. Kiwelekar, “A comparison of mobile patient monitoring systems,” in *Health Information Science*, ser. Lecture Notes in Computer Science, G. Huang, X. Liu, J. He, F. Klawonn, and G. Yao, Eds. Springer Berlin Heidelberg, 2013, vol. 7798, pp. 198–209.

[45] D. A. Tribble, “The health insurance portability and accountability act: security and privacy requirements,” *American Journal of Health-Systems Pharmacy*, vol. 58, pp. 763–770, 2001.

IP Multicast Receiver Mobility Using Multi-homing in a Multi-beam Satellite Network

Esua Kinyuy Jaff, Prashant Pillai, Yim Fun Hu
School of Engineering, Design and Technology,
University of Bradford
Bradford, United Kingdom
ekjaff@student.bradford.ac.uk, p.pillai@bradford.ac.uk, y.f.hu@bradford.ac.uk

***Abstract*—There are several merits of mobile communication devices having multiple network interfaces as compared to traditional devices with just one interface. Smart phones these days are a true example of a mobile multi-homed communication device with heterogeneous network interfaces. Several solutions are available for unicast applications to provide seamless handover using the multiple interfaces of a multi-homed device in terrestrial networks. However, very little has been done on similar support for IP multicast mobility support for mobile satellite terminals in a ubiquitous multi-beam satellite network. Most of the schemes proposed for handovers in multi-homed devices place a lot of emphasis on maintaining the multi-homed device identity especially when the second interface joins the communication session. This increases complexity in the whole system. The issue of maintaining the multi-homed device identity plus the additional signalling messages involve are neither necessary nor desired in an IP multicast communication handover in a multi-beam satellite scenario. This paper seeks to exploit the group communication features of IP multicast (i.e., the fact that anyone can join or leave a multicast group at any time and from any location) and the multiple interfaces of a mobile Return Channel Satellite Terminal (RCST) to support IP multicast communication during handover when a mobile multi-homed RCST changes its point of attachment to the network from one satellite gateway to another.**

Keywords-Multiple Interfaces; Handover; mobile Return Channel Satellite Terminal (mRCST); Multi-beam Satellite Network.

I. INTRODUCTION

Next generation satellite systems, nowadays, are characterised by the support for on-board processing (switching/routing) and multiple spot beams. These new features enable the satellite to make efficient use of its allocated resources and provide cost effective network services. IP multicasting is a technology in which the same data is sent to a group of interested recipients and the network replicating the data as required for delivery until a copy reaches all intended group members. In a multiple spot beam satellite network scenario, IP multicast can be used to communicate important service information like the weather conditions, on-going disaster zones and information, route updates, etc. in long haul flights, global maritime vessels and continental trains. Multicasting this information to all the interested parties rather than individually informing them (i.e., unicast) would save a lot of satellite bandwidth resources.

With an increasing mobile society like ours today, the need for mobility support for IP multicast especially in satellite networks with the potential to provide ubiquitous communications cannot be overemphasised.

Digital Video Broadcasting (DVB) [1] is an open standard published by European Telecommunication Standards Institute (ETSI) describing digital broadcasting using existing satellite (DVB-S), cable (DVB-C), and terrestrial (DVB-T) infrastructures. While originally DVB was designed primarily for audio and video broadcasting, the growth of the Internet and broadband data services has led to the development of the DVB networks to support the transport and delivery of IP based traffic. The Digital Video Broadcasting Return Channel Satellite (DVB-RCS) provides the mechanism to use a satellite as a send data on the return path via the satellite. The large geographic coverage and broadcast capabilities of the DVB-S/RCS network has the advantage of providing IP based services to areas where the deployment of terrestrial infrastructure is uneconomical or impossible.

Based on the possible network topologies, a DVB-S/RCS network support two types of IP multicast services, i.e., star and mesh IP multicast [2]. In star IP multicast, the multicast sources is assumed to be located on terrestrial network which sends the multicast data to the Regenerative Satellite Gateway (RSGW) which in turn forwards the multicast traffic to several RCSTs [3] in the satellite network. On the other hand, in Mesh IP multicast, the source and receivers are all RCSTs of the same interactive satellite network. Each RCST here may have one or more user terminals behind it.

This paper focuses on IP multicast receiver mobility which is the ability of a moving satellite terminal to continue receiver multicast traffic as it moves and changes it point of attachment within the satellite network from one satellite gateway (GW) to another. This is known as gateway handovers. This paper shall focus on the star IP multicast service in which a mobile receiver (i.e., an aircraft) with an on-going multicast session connected to a geosynchronous (GEO) satellite has to undergo a GW handover.

Due to the large round trip delay in GEO satellite networks all handover procedures in multi-beam satellite networks can cause serious link quality degradation or even

ISBN: 978-1-61208-313-1

disconnection of an on-going session. During handover, there is a time period when the mobile node cannot receive or send traffic because of the link switching delay. This period of time known as the handover latency constitutes the primary cause of packet loss during handovers. Longer round trip delays in DVB-S/RCS satellite networks imply longer handover latency meaning more packets loss.

Recently, mobile communication devices with multiple network interfaces (e.g., smart phones) are becoming more and more common. Currently, multi-homed mobile devices are mainly used for maintaining connectivity and achieving desired application quality of service. For example, when link quality on a given network interface drops below a certain threshold value, the multi-homed mobile device will initiate a handover to another network interface with better link quality. A common example of this is the handovers between 3G, HSPA and HSPA+ networks in new smartphones when travelling in a car from one city to another. This paper proposes a novel multi-homing based solution for achieving seamless mobility for IP multicast application in multi-beam satellite networks.

The rest of the paper is organized as follows. Section II presents the general gateway handover signalling sequence and some existing IP multicast receiver mobility solutions that may be adopted for satellite networks. In Section III, the proposed multi-homing based solution for IP multicast receiver mobility is described in detail. The performance evaluation of the proposed system is presented in Section IV. Finally, conclusions are discussed in Section V.

II. Gateway Handover Signalling and Multicast Mobility

A. *Gateway Handovers in a DVB-S/RCS network*

Figure 1 shows the signalling sequence at GW handover [3]. When the Network Control Centre (NCC) receives the synchronization (SYNC) burst from the mobile RCST (mRCST) containing the handover request, it will retrieve the target beam identity from its database and determine whether the beam belongs to a different GW. Once the NCC establishes that the target beam belongs to a different GW, a gateway handover is initiated. The NCC will then update its service information (SI) tables which include Terminal Burst Time Plan (TBTP), Super-frame Composition Table (SCT), Frame Composition Table (FCT) & Time-slot Composition Table (TCT). The NCC will send an SNMP Set-Request message that includes the updated SI tables and the routing update information (RUI) of the mRCST to the target GW to ensure that the target GW gets ready for connection with the mRCST. Upon reception of the SNMP Set-Request message, the target GW will allocate bandwidth resources for the mRCST according to the new burst time plan sent by the NCC. The SNMP Get-Response message is then sent by target GW to the NCC. This is followed by a SNMP Set-Request message from the NCC to the source GW, which includes the mRCST identity and the SI tables. Upon reception of the SNMP Set-Request message, the source GW will start buffering the Forward Link (FL) user traffic to be forwarded to the target GW during handover. The source GW then acknowledges the NCC by sending a SNMP Get-Response message. Once the SNMP Get-Response message is received from source GW, a gateway handover command is issued to the mRCST from NCC in a Mobility Control Descriptor carried in a Terminal Information Message Unicast (TIMu) using the old beam. The source GW now updates its route mapping table and released resources used by the mRCST. Once the mRCST receives the handover command, it synchronizes with the NCC and the target GW, retunes itself to the target beam and receives traffic from the target beam which comes through the target GW.

Hence, it can be seen that for a mRCST with one interface (i.e., one transceiver), there comes a time interval during the GW handover execution phase, when the forward link and/or return link user traffic is discontinued. This time is indicated in Figure 1 and is the time when the mRCST is switching its point of attachment to the network from source GW to target GW. Assuming here that the NCC is located in a different GW from the source and target GWs, handover latency for forward link (HL_{FL}) from Figure 1 is given by:

$$HL_{FL} = T_{pd}(FL) + T_{pd}(FL) + T_{D1} + T_{MAX(Tx\_,Rx\_tuning)} + T_{ACQ\_U} + T_{MSL} - T_{pd}(FL) \quad (1)$$

Simplifying gives

$$HL_{FL} = T_{pd}(FL) + T_{D1} + T_{MAX(Tx\_,Rx\_tuning)} + T_{ACQ\_U} + T_{MSL} \quad (2)$$

From [2], T_{MSL} is given by:

$$T_{MSL} = T_{pd(FL)} + T_{pd(RL)} + T_{D2} \quad (3)$$

Combining (2) and (3) and assuming that:

Tpd(FL) = Tpd(RL) = Tpd and TD1 = TD2 = TD3 = TD gives

$$HL_{FL} = 3T_{pd} + 2T_D + T_{MAX(Tx\_,Rx\_tuning)} + T_{ACQ\_U} \quad (4)$$

Similarly, return link handover latency (HL_{RL}) is given by:

$$HL_{RL} = 2T_{pd} + 2T_D + T_{MAX(Tx\_,Rx\_tuning)} + T_{ACQ\_U} \quad (5)$$

where,

T_{pd}(FL) is the forward link propagation delay
T_{pd}*(RL)* is the return link propagation delay
$T_{D1} = T_{D2} = T_{D3} = T_D$ is the processing delay
$T_{Max\ (Tx\_,\ Rx\_tuning)}$ is the Maximum time required for retuning the transmitter (Tx*)* and receiver (Rx) to new frequencies of the target beam;
T_{ACQu} is the acquisition uncertainty. This is the time taken by mRCST to issue the ACQ burst in new beam after complete retuning of Tx and Rx;
T_{MSL} is the minimum time interval from issuing a capacity request in mRCST and the mRCST dispatching traffic in the slots allocated in response to that request.
Based on Figure 1 and the above analysis, the FL handover latency is slightly greater than the RL handover latency.

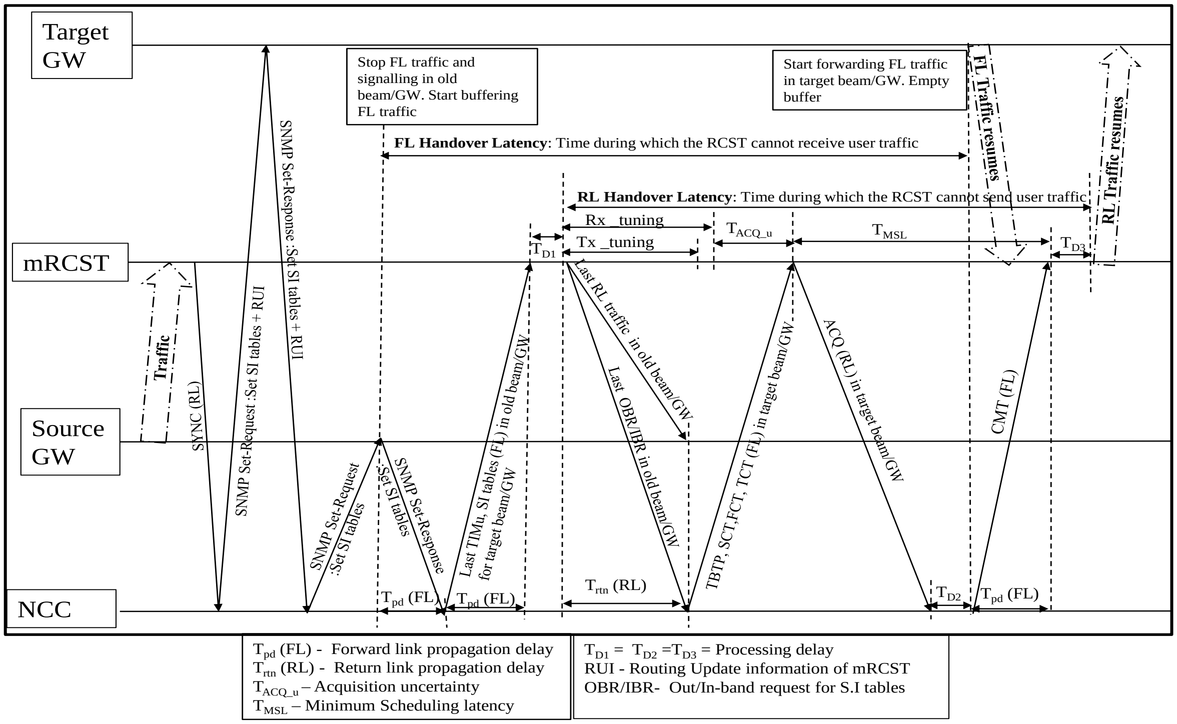

Target GW
Stop FL traffic and signalling in old beam/GW. Start buffering FL traffic
Start forwarding FL traffic in target beam/GW. Empty buffer
FL Traffic resumes
RL Traffic resumes
SNMP Set-Response :Set SI tables + RUI
FL Handover Latency: Time during which the RCST cannot receive user traffic
RL Handover Latency: Time during which the RCST cannot send user traffic
Rx _tuning
Tx _tuning
T_ACQ_u
T_MSL
T_D1
T_D3
mRCST
Traffic
SYNC (RL)
SNMP Set-Request :Set SI tables + RUI
Last RL traffic in old beam/GW
Last OBR/IBR in old beam/GW
Last TIMu, SI tables (FL) in old beam/GW for target beam/GW
TBTP, SCT,FCT, TCT (FL) in target beam/GW
ACQ (RL) in target beam/GW
CMT (FL)
Source GW
SNMP Set-Request :Set SI tables
SNMP Set-Response :Set SI tables
NCC
T_pd (FL)
T_pd (FL)
T_rtn (RL)
T_D2
T_pd (FL)
T_pd (FL) - Forward link propagation delay
T_rtn (RL) - Return link propagation delay
T_ACQ_u – Acquisition uncertainty
T_MSL – Minimum Scheduling latency
T_D1 = T_D2 =T_D3 = Processing delay
RUI - Routing Update information of mRCST
OBR/IBR- Out/In-band request for S.I tables

Fig .1. Gateway Handover Signalling Sequence

Since the FL and RL handovers are done in parallel, the handover latency of the one that takes longer to complete then becomes the overall handover latency. So, the FL handover latency here becomes the overall handover latency of the mRSCT.

Due to this link switching handover latency, seamless handover in multi-beam satellite networks cannot be achieved in a mRCST with only one transceiver. This handover latency has a huge impact on real-time applications. Buffering delay/jitter-sensitive application traffic during handover has no benefit since the off-time is rather high and therefore practically impossible to compensate for the delay introduced by buffering.

This handover latency is mainly associated with the link switching procedure in multi-beam DVB satellite networks i.e., the time when the forward link (FL)/return link (RL) user traffic is discontinued as the mRCST is releasing the resources in source beam and acquiring new set of resources in the target beam since the standard mRCST (one transmitter and one receiver) cannot establish connections on both beams simultaneously. This therefore implies that the handover latency is independent of the higher layer mobility management protocols used during GW handover so far as the mobile terminal has just one transceiver.

B. *IP Multicast mobility*

This section presents some existing solution for IP multicast mobility designed for terrestrial networks and the Internet that may be used in the satellite network. In general, Home Subscription (HS) and Remote Subscription (RS) based approaches have been proposed in [4] [5] [6] to support IP multicast receiver mobility in terrestrial networks. In the HS-based approach, the mobile node while away from home network establishes a bidirectional tunnel with its home agent (i.e., a multicast enabled router in the home network of the mobile node). Any multicast traffic received by home network for this mobile node is then tunnelled to the mobile node in the foreign network. HS based approach relies on mobile IP (MIP) [7] for its operation. While such HS based approach could be adapted for use in satellite networks, it inherits the drawbacks of MIP [8] like the triangular routing problem where, any traffic destined for the mobile receiver must pass through its home network. This will further increase the overall handover latency as additional signalling time is required to establish the bidirectional tunnel between home GW and target GW during the GW handover, as this can only be done when the terminal connects to the target GW and receives a new Care-of-Address. Traffic from the multicast source to the mobile receiver will incur additional propagation delay due to triangular routing problem. This triangular routing problem becomes even more acute if the target GW and the multicast source are located in the same terrestrial network as shown in Figure 2. In this case, after GW handover, the multicast traffic will be first sent from the source in target network to the home GW and then tunnelled back to target GW for onward transmission through satellite to the mobile multicast receiver (mRCST).

The RS-based approach, on the other hand, requires the mobile receiver to simply re-subscribe to all the multicast groups it was a member of in the home network after handover to a foreign network. Here, additional time is required for multicast group subscription and tree reconstruction to new location if the mobile receiver is the first member of the group in the new network.

These approaches if adapted for use in satellite networks to support receiver mobility will have no effect on the HL described above. This is because HL is the minimum time required for the mRCST with only one transceiver to release satellite resources in one beam and acquire new set of resources in the next beam and is independent of these mobility support mechanisms. Any such mobility support technique for a mRCST with only one transceiver can only increase the HL given in (2) by adding further signalling delay or multicast tree setup delay.

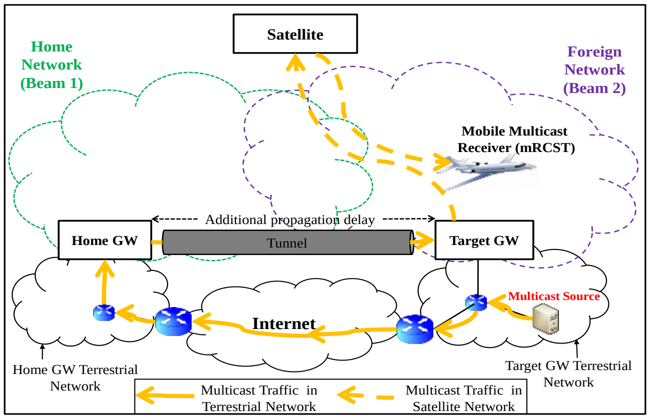

Satellite
Home Network (Beam 1)
Foreign Network (Beam 2)
Mobile Multicast Receiver (mRCST)
Additional propagation delay
Home GW
Tunnel
Target GW
Multicast Source
Internet
Home GW Terrestrial Network
Target GW Terrestrial Network
Multicast Traffic in Terrestrial Network
Multicast Traffic in Satellite Network

Fig .2. Triangular Routing Problem

A multi-homed LEO satellite connected with two ground stations to support network mobility in space was proposed in [9] and multi-homed end nodes were proposed in [10] to support terminal mobility. The two schemes rely solely on the Stream Control Transmission Protocol (SCTP) [11] to maintain the identity of the multi-homed devices when the second interface joins the communication during handover procedure. Mechanisms to support network and terminal mobility for multi-homed device in IPv6 networks were presented in [12] and [13]. Host Identity Protocol [14] was adopted for providing mobility for multi-homed mobile devices in [15] and [16]. The proposed mechanism in [15] suffers from long handover delay owing to duplicate address detection (DAD), location update, and other signalling overhead. The proposed mechanism in [16] introduces large complexities in access routers like the tracking and updating mobile host location, security signalling, assigning network prefix per host identifier and using the same network prefix within the same network domain to avoid DAD. All these multi-homed based schemes have been designed primarily for unicast communication, where the emphasis is laid on designing protocols/mechanisms to maintain the identity of the multi-homed node when its second network interfaces joins the communication during handover procedure. These system complexities plus additional signalling messages employed to maintain host identity during handover are neither necessary in an IP multicast communication nor desirable in a satellite network with scarce and expensive satellite resources.

III. PROPOSED MULTI-HOMING BASED IP MULTICAST RECEIVER MOBILITY

In order to reduce the link handover delay described in the previous section, the proposed method leverages on the group communications features of IP multicast and the fact that anyone can join or leave a multicast group at any time. Figure 3 shows the proposed internal architecture of a multi-homed mRCST for Satellite Interactive System.

Figure 3 contains new features/entities in addition to the standard RCST given in [1]. These include:

- An additional broadcast interface (IF1) (i.e., for receiving data via DVB-S) in the broadcast interface module with its corresponding additional interactive interface (IF1) (i.e., for sending data via DVB-RCS) in the interactive interface module, making the mRCST a multi-homed device.
- A database which holds information about the global map of the interactive satellite network (i.e., information about beams, their locations and frequency, gateways - location and IP addresses) as well as all active connections in the mRCST.
- A message chamber which can issue IGMP join report and leave messages during handover between IF0 and IF1
- The controller which manages the data base, the interfaces and has complete control over which interface the traffic leaves/enters the mRCST especially when the two are active (i.e., during handover)

It is assumed that the mRCST (on aircrafts, ships, trains etc.) knows its complete route map (all beams and GWs along its path) before start of journey. As shown in Figure 3, the multi-homed mRCST contains two pairs of satellite network interfaces, IF0 and IF1 in the broadcast interface module with their corresponding pairs in the interactive interface module. The interfaces in the broadcast interface module are used for receiving FL traffic and signalling while those in the interactive interface module are used to send RL traffic and signalling. If FL traffic is received through IF0 in broadcast interface module, then the reply (RL traffic) will be sent out through IF0 in the interactive interface module. The same holds for traffic received through IF1 in the broadcast interface module.

When the multi-homed mRCST, i.e., the aircraft in Figure 4, with an on-going multicast session through interface IF0 enters an overlapping area of two satellite beams belonging to different GWs, it will detect the presence of the new satellite beam.

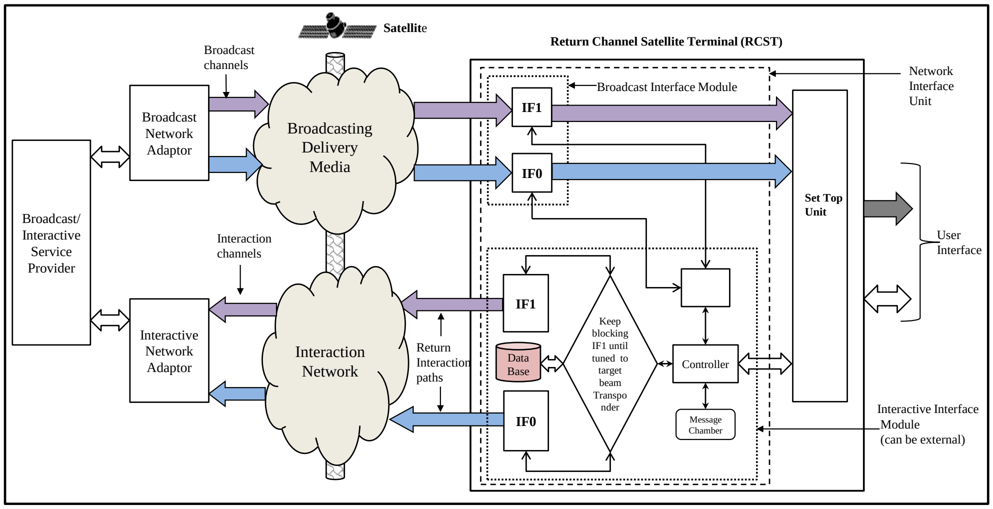

Satellite
Return Channel Satellite Terminal (RCST)
Broadcast channels
Broadcast Network Adaptor
Broadcasting Delivery Media
Broadcast/ Interactive Service Provider
Interaction channels
Interactive Network Adaptor
Interaction Network
Return Interaction paths
Broadcast Interface Module
Network Interface Unit
IF1
IF0
Set Top Unit
User Interface
Keep blocking IF1 until tuned to target beam Transponder
Data Base
Controller
Message Chamber
Interactive Interface Module (can be external)

Fig .3. Multi-homed mRCST for Satellite Interactive System

The controller will then consult the database within the mRCST to confirm whether the detected new beam is the target beam. If the detected new beam is the target beam, IF1 through instructions from the controller will then establish a connection with the target beam using normal logon procedure. This is closely followed by the message chamber issuing an IGMP join report through IF1 to the NCC to join all the multicast groups that the mRCST is a member of. Due to the fact that anybody can join or leave a multicast group at any time, the joining of the multicast session by the second interface IF1 does not need to be proven that the two interfaces (IF0 and IF1) belong to the same device. This therefore makes the handover hidden from the satellite network i.e., as far as the satellite network is concerned, the second interface (IF1) may just be another RCST that has logged on to the satellite network and established communication.

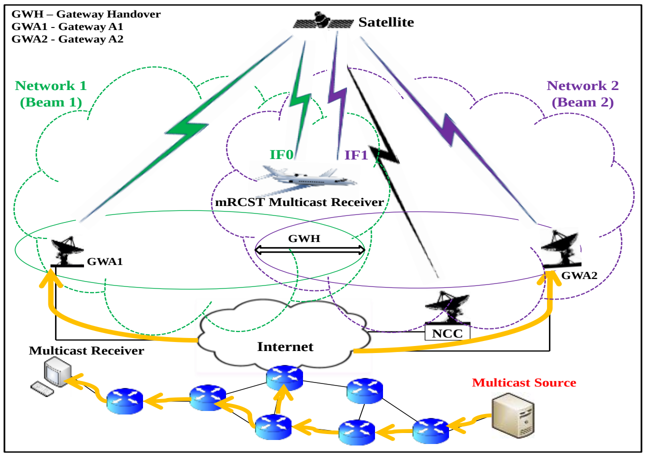


Fig .4. Gateway Handover for a multi-homing enabled mRCST in a Multi-beam Geo Satellite Network.

After this, the controller starts directing all other new communications or connections from the mRCST through IF1. Immediately IF1 starts receiving traffic from the on-going multicast session(s), the message chamber will issue an IGMP leave message through IF0. Eventually, all communications or connections from and to the mRCST are channelled through IF1 and once this happens then IF0 then enters a stand-by/log-off state. Considering the fact that in a GEO satellite network, the area of overlapping beams can stretch for many miles, it is possible to keep the old connection through old point of attachment (GWA1) alive until the new one via GWA2 is set up and all communications transferred to the new link. When the mRCST enters the next area of overlapping gateway beams, the same procedure is followed that will see all communications on mRCST transferred back to IF0 from IF1.

The advantages of this scheme are:

- it is simple to implement
- minimal handover latency – only due to link retuning time
- there are no packet losses at all due to handover as the handover is completely and truly seamless

IV. Comparison of GW Handover Latency

As stated in Section II, the GW handover latency of a standard mRCST with one interactive interface is given by (4). Assuming that the Super-frame duration is 500ms [17], satellite round trip delay is 250ms and using beam/gateway handover details given in [3], we can establish Table I. According to [3], the T_D in all satellite network devices (NCC, RCST and GW) could take 2 - 3 Super-frames and the $T_{Max\ (Tx\_,\ Rx\_tuning)}$, 1 - 2 seconds. This explains why T_D and $T_{Max\ (Tx\_,\ Rx\_tuning)}$ have two different sets of values in Table I.

Case 1: when the T_D and $T_{Max\ (Tx\_,\ Rx\_tuning)}$ take minimum values.

Case 2: when the T_D and $T_{Max\ (Tx\_,\ Rx\_tuning)}$ take maximum values.

TABLE I. TIME DELAYS

	Value			
	Time in Super frames		**Time in Seconds**	
	Case 1: Min value	**Case 2:** Max value	**Case 1:** Min value	**Case 2:** Max value
$\mathbf{T_D}$	2	3	1	1.5
$\mathbf{T_{pd}}$	-	-	0.125	0.125
$\mathbf{T_{Max(Tx-,Rx\_tuning)}}$	-	-	1	2

Using (4) and the values in Table I, the change in handover latency (HL) with respect to acquisition uncertainty (T_{ACQ_u}) for a mRCST with one interactive interface can be calculated and compare it with our proposed scheme which has two interactive interfaces and a HL of zero.

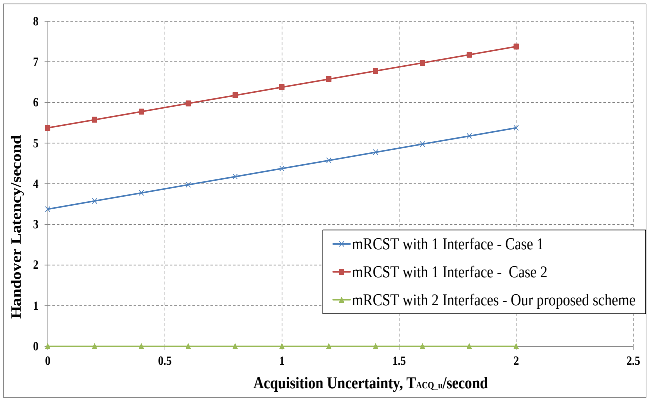


Fig .5. Comparing the Handover latency during GW handover in Geo Satellite Network

Figure 5 shows that the mRCST with two interactive interfaces during GW handover in satellite networks has the best possible handover latency (zero) for IP multicast communication.

V. Conclusion

Based on the DVB specifications, a mobile RCST with a single transceiver will always face a small period of service disruption during the handover phase. Higher layer mobility management protocols cannot remove this intrinsic delay. This paper describes in detail how a multi-homed mRCST can be used to support IP multicast receiver mobility during gateway handover in a global multi-beam satellite network. It proposes the internal architecture of such a multi-homed mRCST. The use of the proposed multi-homed terminal eliminates this handover latency for IP multicast communication over the satellite as it changes its point of attachment to the satellite network from one satellite gateway to another in a global multi-beam satellite network.

References

[1] ETSI EN 301 790, "Digital Video Broadcasting (DVB); Interaction channel for satellite distribution systems," vol. 1.5.1, May 2009.

[2] ETSI TS 102 429-2, "Satellite Earth Stations and Systems (SES); Broadband Satellite Multimedia (BSM); Regenerative Satellite Mesh - B (RSM-B); DVB-S/DVB-RCS family for regenerative satellites; Part 2: Satellite Link Control layer," vol. 1.1.1, October 2006.

[3] ETSI TR 102 768, "Digital Video Broadcasting (DVB); Interaction channel for Satellite Distribution Systems; Guidelines for the use of EN 301 790 in mobile scenarios," vol. 1.1.1, April 2009.

[4] I. Romdhani, M. Kellil, H. Lach, A. Bouabdallah, and H. Bettahar, "IP Mobile Multicast: Challenges and Solutions," IEEE Communications Survey and Tutorials, vol.6, First Quarter 2004, pp. 18-41.

[5] T. Nguyen, "IP Mobile Multicast: Problems and Solutions," Ph.D. Dissertation, EUROCOM, France March 2011.

[6] G. Xylomenos and G.C. Polyzos, "IP multicast for mobile hosts," IEEE Communications Magazine, vol. 35, January 1997, pp. 54–58.

[7] C. Perkins, "IP Mobility Support," RFC 2002, IEFT, October 1996.

[8] P. K. Chowdhury, A.S. Reaz, M. Atiquzzaman, and W. Ivancic, "Performance Analysis of SINEMO: Seamless IP-diversity based Network Mobility," in proceedings of IEEE International Conference on Communications, June 2007, pp. 6032 -6037.

[9] P. Chowdhury, M. Atiquzzaman, and W. Ivancic, "SINEMO: An IP-diversity based approach for network mobility in space," in Proceedings of IEEE Second International Conference on Space Mission Challenges for Information Technology (SMC-IT), 2006, pp.108-115.

[10] S. Fu, M. Atiuzzaman, L. Ma, and Y. Lee, "Signaling cost and performance of SIGMA: A seamless handover scheme for data networks," Journal of Wireless Communications and Mobile Computing, vol. 5, October 2005, pp. 825-845.

[11] R. Stewart, "Stream Control Transmission Protocol (SCTP)," RFC 4960, IETF, September 2007.

[12] M. S. Hossain and M Atiquzzaman, " A Network-based Seamless Handover Scheme for Multi-homed Devices," in Globecom Workshops of Fourth International Workshop on Mobility Management in the Networks of the Future World, December 2012, pp. 1042-1046.

[13] M. S. Rahman and M. Atiquzzaman, "SEMO6 - A Multihoming based seamless mobility management framework," IEEE Military Communication Conference (MILCOM), 2008, pp. 1-7.

[14] R. Moskowitz, P. Nikander, P. Jokela, and T. Henderson, "Host Identity Protocol," RFC 5201, IETF, April 2008.

[15] P. Nikander,T. Henderson, C. Vogt, and J. Arkko, "End-Host Mobility and Multihoming with the Host Identity Protocol," RFC 5206, IETF, April 2008.

[16] M. M. Muslam, H.A Chan, L.A. Magagula, and N. Ventura, "Network-Based Mobility and Host Identity Protocol," in IEEE Wireless Communications and Networking Conference (WCNC), 2012, pp. 2395-2400.

[17] O. Alphand, P. Berthou, and T. Gayraud. "SATIP6 : Satellite Testbed for Next Generation Protocols," June 2013,

http://researchwebshelf.com/uploads/166_P45.pdf.

Authentication with Keystroke Dynamics on Touchscreen Keypads - Effect of different N-Graph Combinations

Matthias Trojahn
Mobile-Devices & Auto-ID Technologies
Volkswagen AG
Wolfsburg, Germany
matthias.trojahn@volkswagen.de

Florian Arndt
Client Design & Management
Volkswagen AG
Wolfsburg, Germany
florian.arndt1@volkswagen.de

Frank Ortmeier
Computer Systems in Engineering
Otto-von-Guericke University
Magdeburg, Germany
frank.ortmeier@ovgu.de

***Abstract*—The security and access protection on mobile devices have become an important topic due to the increasing amount of personal and sensitive data stored on these devices. The traditional security techniques based on PIN (*Personal Identification Number*)-input are insufficient and do not correspond to the present password standards. An authentication with usage of keystroke behavior could increase the security and a lot of research has been published based on traditional PC keypads. Keystroke behavior on touchscreen keypads, as they are nowadays installed on smartphones, enables adding additional features for the authentication. For example, pressure, size or exact coordinates of keystroke can be used. The focus of this paper is that several time differences (e.g. digraph) are examined and checked for suitability for a keystroke authentication. For that, data of 152 subjects were classified. With additional features of the touchscreen, an error rate of false rejected persons of only 4.59% and false accepted persons of only 4.19% could be reached.**

Keywords*—*keystroke authentication; n-graph; mobile devices; capacitive display

I. Introduction

Smartphones are not only used to phone or write a SMS (*Short Message Service*), especially, with the introduction of the iPhone in the year 2007. This also increases the number of security relevant data and information which are stored on the smartphone or provided through applications. This could be private online banking data, passwords or confidential company documents, e.g., extracted from E-Mail attachments [1], [2]. In addition, personal data (e.g., GPS (*Global Positioning System*) data) have an increasingly important significance [3].

For this reason, the security on smartphones established itself as an important topic, especially the access protection [4], [5]. The challenge is to protect and to avoid economic damage for the enterprise data and assets [6]. However, traditional PIN authentication can easily be defeated and does not conform to present password standards [7]. The standard goes for a two-factor-authentication. In this case, in addition to the password, a possession of a person (e.g., bank card, token) [7] are required.

Greater security can be achieved by another (biometric) authentication factor, such as keystroke dynamics. Components are on the one hand an item that the person knows (password) and on the other an element that is the person himself/herself (keystroke) [8]. Other biometric methods are possible but are not discussed in this paper.

Keystroke has the advantage that no additional hardware is needed because the standard keyboard of the device can be used. However, special software is required for reading the typing behavior. This is, in terms of sustainability, an advantage over other authentication methods which need additional hardware, such as card reader for user ID's [9]. Furthermore, by the lack of additional hardware, the ease of use in keystroke dynamics as an authentication method is less affected [10]. For the optimal case, the user does not notice the procedure because the password entry and the keystroke is checked automatically by the software.

Through the use of a biometric modality additional personal information are stored. Especially, the methods for face recognition cannot be used every time because of religious or cultural reasons [11]. A further complication is that the storage of this data includes an agreement by the user (e.g., German Federal Privacy Act 4a [12]). Saving the typing behavior is seen by many people, however, as less critical, which is a further advantage of the process [1].

In this study, we will discuss keystroke dynamics on touchscreen devices and examine it by an experiment. It will be shown which features or feature combination is suitable for authentication. In addition, we added the pressure and size during typing to see whether it is improving the authentication. The pressure is also used in handwriting recognition [13]. For this purpose, different combinations of properties of typing habits are selected and examined for their suitability for authentication.

In this paper, we will first describe in Section 2 some basic backgrounds regarding biometric authentication and then the hypotheses. In Section 3, the experimental design will be explained. After this, the results are shown and discussed in Section 4. In the last Section, we will give a conclusion with some implications and future work.

II. Theoretical Background

In this section, we will describe some basics about biometric authentication. Then, we present our hypotheses.

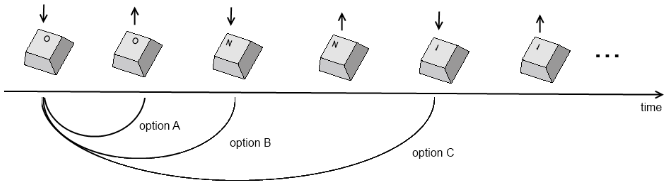


Fig. 1: Events between pressing different keys

A. Biometric Authentication

Keystroke dynamics is a biometric characteristic of a person which describes the rhythm how a person is typing on a keyboard [14]. It can be used like other biometric methods to identify a person. Basically, the rhythm is a characteristic which can be calculated by different aspects, but, in most cases, at least the time differences were used [15].

As seen in Figure 1, different time differences can be used for authentication. In general, two different types of events can be recorded: The first represents the length of time a key is pressed which is time between pressing and releasing a key (residence time) (option A). On the other hand, the time period can be calculated by the striking of keys (the release is here ignored). The time difference between two keystrokes, defined by the two press events, is also called movement time (option B). Both variants of the time period can be called digraph [16]. Furthermore, combinations of more than two keystrokes can be used. A time difference between n key events is called n-graph [15]. In addition to the use of digraphs in many publications, the combination of three key presses are also utilized-the trigraph (option C) [17]. In general, all the values for $n > 1$ are possible in order to determine the time differences. The higher the values the smaller the information which can be extracted by the input because an average over n events is used.

In this paper, the term digraph is used for a better understanding only for the movement time between two keystrokes. We use the residence time to denote the duration of time a key is pressed.

All these features are used to determine a person uniquely. Various classifiers are used to compare these characteristics. As classifiers, in most cases, a statistical classifier (such as distance measures) [18] or neural networks [19], [20] are used.

The authentication process will be measured using different error rates (mismatches). Two possible errors can be distinguished: First, the false acceptance rate (FAR), which indicates the proportion of falsely accepted people:

$$FAR = \frac{number\ of\ false\ acceptances}{number\ of\ impostor\ identification\ attempts} \tag{1}$$

On the other side is the false rejection rate (FRR), which represents how many users are rejected by the system although they are the right person:

$$FRR = \frac{number\ of\ false\ rejections}{number\ of\ enrollee\ identification\ attempts} \tag{2}$$

The last value is the Equal Error Rate (EER) which represents the point where FAR and FRR are the same [21].

The keystroke dynamics on conventional keyboards have been part of numerous scientific papers. The residence time [22], [23], [24] as well as the movement time [14], [25], [26] have been examined as a feature for typing behavior. Besides these features, the number of input errors can be used as well [2]. The time characteristics and the error rate are the most widely used features that can be extracted by computer keyboards [27].

Even the typing behavior for authentication on mobile phone keypads has been investigated [1], [2]. Here, devices were used which have 12 hardware buttons.

In the survey paper of Banerjee [28] a good overview was given about different experiments and their results. Here, only a set of these studies will be presented.

Basically, all publications are related to the keyboard of a computer or to a mobile device with 12 keys. The error rates are influenced by the number of subjects, the classifier and the features. The first study about keystroke dynamics on computer keyboards were done by Umpress et al. (FAR: 11.7%, FRR: 5.8%) [18] and Joyce (FAR: 0.3%, FRR: 16.4% with 33 subjects) [29] and used only digraphs and a statistical classifier. Later on Ord and Furnell [30] used a neuronal network for 14 subjects and get a result of 9.9% for the FAR and 30.0% for the FRR.

Also on the mobile device (12-key layout) good results were published by Clarke et al. [31] with an EER of 15.2% and Zahid et al. [2] had a FAR of 11% and a FRR of 9.22%. Both experiments used neuronal network and had 25 respectively 30 subjects.

One of the first studies with a touchscreen device was done by Saevanee [32] who used the pressure of the fingertip. With the ten subjects he achieved an accuracy rate of 99%.

In addition to the pressure feature, the size of the fingertip was used by Trojahn and Ortmeier [33]. They analyzed the typing behavior of 35 subjects (FAR: 9.53%, FRR: 5.88%).

B. Hypotheses

The existing publications are mainly dealing with the computer or 12-key mobile phone keyboard. On touchscreen keyboards which are now installed in nearly every smartphone, besides the well-known features, other possibilities for typing behavior can be read. Examples for this are the pressure or the size of the fingertip during typing. These can be used in combination with the time values for authentication [34].

For this, we want to prove that an authentication with a touchscreen keyboard is possible. We identified therefore the following hypotheses:

H1: If an authentication is done with a touchscreen keyboard using n-graph the same error rates can be achieved compared with the existing keystroke dynamics studies.

H2: If the model is extended with additional features (pressure and size of the finger) the error rates can be reduced.

III. Experimental Design

The subjects were asked to enter a predefined, 17-digit passphrase ten times in a row on a smartphone. A Samsung Galaxy Nexus has been used as a test device. A soft keyboard was implemented to deactivate uppercase and the alignment changes. In addition, an application was designed to read the keystroke. Other descriptive data were also queried about the

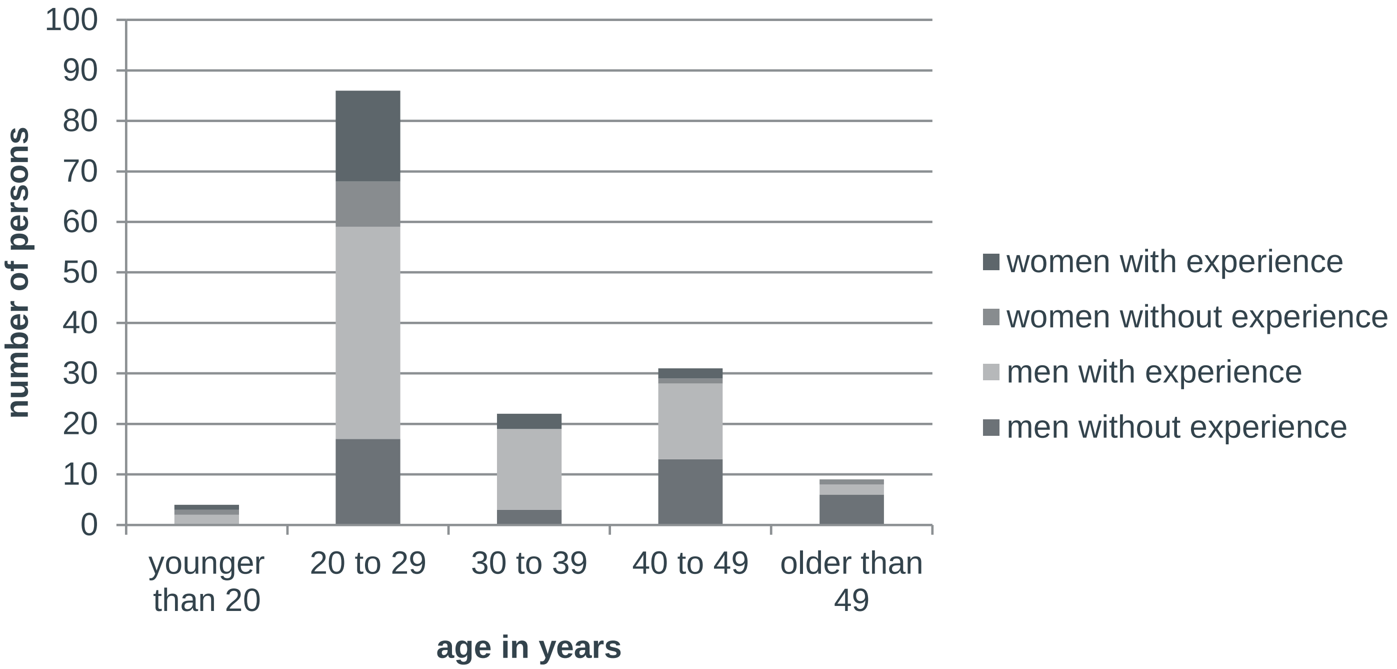


Fig. 2: Experience structure in combination with age and sex

self-written application and stored on the mobile device in a file.

In the experiment, 101 out of the 152 participants already had prior experience with a touchscreen keyboard. Out of these 101 persons, 62% had experience with the Android operating system. On average, the subjects use their smartphone 1.7 hours per day. The experiment lasted 15 to 20 minutes per subject. More detailed information of the data set can be found in Figure 2.

In order to avoid the user familiarity with the device affect the results of the test, the subjects had time to practice with the equipment and get used to the keyboard. In addition, they had to enter a test text before the real test started. All ten repetitions of the password had to be entered correctly. Otherwise, this entry attempt was rejected and the user had to enter the same password a second time. Afterwards, the data set was divided into test and evaluation data. Every third input of a test person had been assigned to the evaluation data and the remaining part to the training data. This means that per subject seven inputs are used for the training data and three were used as evaluation data. This mutual selection was continued for further inputs and was done to reduce the fault which is created the rising learning curve.

The recorded raw data include all interactions (events) with the keyboard: Press and release buttons and movements on the screen. For each event, the date and the code of the pressed key are stored as well.

In the present study, different combinations of button presses and different n-graph (digraph and trigraph) for their suitability for authentication are examined. First, the individual features are calculated and in the later course fused to represent the ability of combinations (no weights between the features).

Since at each authentication attempt, the tip pattern shows up slight differences and entry errors are made, it is important to smooth the differences. Therefore, a classification method (statistical classifier using K-Means) was used with brute force to filter the best solutions. Two filters were implemented in front of the classifier. At first, the data were divided into test and evaluation data (ratio 7:3). After this, for each feature the two largest values were extracted and from the rest an average value was calculated for the model. The individual evaluation data were compared to this average. The value has to be within a specified tolerance of the average. At the end, it was decided based on the individual values and a predefined threshold value, whether it is a person or not.

IV. Results and Discussion

For the first evaluation of the recorded data, the single features have been extracted. In addition, the residence time of the keystroke, the digraph and trigraph were calculated. The following figures show the extracted data for five randomly selected subjects for better representation (analogue [35], [31]). Figure 3 shows the data of the digraph and Figure 4 represents the data of the retention time.

The amount of data which can be extracted depends on the number of letters and on which feature is selected. For the residence time there can be extracted one more than for the digraph. That means for a 17 letter key phrase 16 values can be extracted.

In Figure 3 it can be seen that for the randomly selected subjects the average duration is in most cases more constant over the time. The same applies for the whole record. However, there are considerable differences among the people, even in the general speed and then between single digraphs of one subject. The differences between individuals can be explained on the basis of experience. So the fifth person, for example, has a lot of experience while subject number four has no experience. This explains also the high values for each digraph

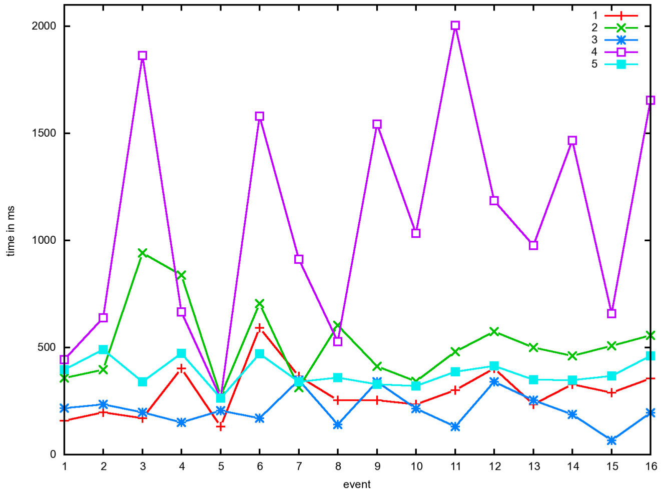


Fig. 3: Digraph for five different users

for the fourth person. The slight movement time for all subjects at fifth time event is explained by the fact that the password has the same letter twice in a row. In this situation the person does not need to search next letter. The described features of the typing behavior of the five randomly chosen test subjects can be observed in all subjects of the experiment in similar ways.

The rhythm of the residence time (Figure 4), however, is less constant between individuals and also between different attempts by one person. Furthermore, the residence time tends to be less than the movement time. A person needs more time to press the next key than to hold a key. Also the differences in experience of a touchscreen are less noticeable at the residence.

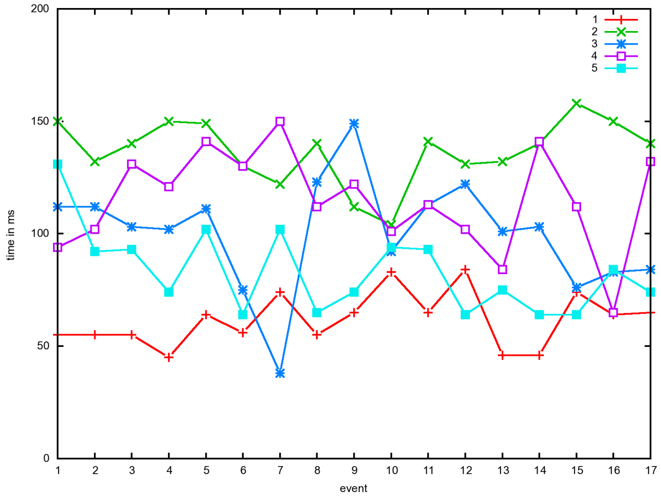


Fig. 4: Residence time for five different subjects

Table I shows the results of the classification as average values of all 152 subjects with the statistical process with several different combinations of the extracted features.

To simplify the comparison, the results shown in Table I are generated by comparing the sum of the error rates for each feature selection. This approach was already selected to determine the best error ratio for the individual characteristics and allows obtaining the respective minima. A comparison of the first two lines indicates that the assumption that the residence time is more suitable than the digraph is correct. With the feature digraph, for example, one of nine people is incorrectly accepted and only every 15th attempt to authenticate someone is falsely rejected. The feature of the trigraph is even less appropriate and gives worse results than the feature over two keystrokes.

TABLE I: Results for different feature combinations

FAR (in %)	FRR (in %)	Selection of features
8.03	12.3	residence time
12.66	11.64	digraph (two different keys)
13.63	33.33	trigraph (three different keys)
9.28	6.72	residence time + digraph
10.01	10.98	residence time + trigraph
10.02	13.77	digraph + trigraph
6.61	8.03	residence time + digraph + trigraph

Better results can be achieved with combinations of features. Then the sum of the lowest error rate is achieved with the combination of residence time digraph. This combination was, therefore, continued to be used in order to merge with add-on features.

Very good results have been achieved with the additional features of the pressure, the size of the key presses, the exact coordinates while pressing and releasing the finger. A combination of residence time and digraph and additional features (e.g., pressure and size) can provided a FAR of 4.19% and a FRR of 4.59%. This means that an authentication on the basis of these characteristics, only every 24th person is falsely accepted and only every 22th attempt is falsely rejected. In comparison to the general password approach where each attempt is successful it is an improvement. Altogether this means, the traditional feature (duration, digraph and trigraph) can be improved with the new features.

V. Conclusion

This section will give a summary and will describe some limitations as well as some future work.

A. *Summary and Implications*

Due to the increasing number of sensitive, personal information on mobile devices, the security and access control settings on these devices are becoming an increasingly important issue. A higher security is, however, in accordance with normal additional hardware less user friendly, for example, by stringent password standards (more letter or alphanumeric).

Based on the study, it was shown that an authentication using the keystroke is possible on touchscreens and enhances the security. Furthermore, the study has shown an attack on the authentication. Each subject used the same password (so it was known by everyone) in this situation only the behavior was important. It showed the same impact as they were already achieved in previous scientific publications with conventional keyboards (see Table 1). By adding further features of the typing behavior the error ratio can be reduced more.

In most combinations of features the calculated FRR was smaller than the FAR. In particular, in the best case when using the pressure and the size of the keystrokes, in addition, the

FRR was particularly low. This is desirable in mobile devices, since the ease of use is desired [36] to ensure acceptance of the procedure. If the correct user is rejected too often incorrectly, the method is not usable.

B. Limitation and Future Work

Like most studies, some limitations exist in our study that should be mentioned at this point.

The experiment was conducted in a controlled environment, which reduces many outside influences in the observation. These include stress and negative emotions to the subject, during the authentication process on their smartphone, which may affect the keystroke dynamics [37]. Even movements and orientation of the phone can affect the typing and should be observed in further studies [38]. This includes whether the person is sitting, lying or standing.

The presented method for authentication of mobile devices can be applied to all devices with a touchscreen. Thereby, it is possible to use the method for mobile devices such as the smart grid environment [39], [40]. Examples of this would be that a technician who is equipped with a mobile device (stores sensitive data) secures his device against unauthorized access in case of theft. For this, a two-factor authentication using the keystroke would be a useful and user-friendly solution that can be used without any additional hardware and thus lower costs.

Future research could aim to achieve a further reducing of the error rates by including the factors that the user no longer perceives the additional authentication. In addition, it could be tested how much influence a rhythm during typing has [41], like using own pauses or music melodies during typing. This could greatly increase the usability and acceptance of the process.

REFERENCES

[1] A. Buchoux and N. L. Clarke, “Deployment of keystroke analysis on a smartphone,” in *Proceedings of the 6th Australian Information Security & Management Conference*, 2008.

[2] S. Zahid, M. Shahzad, S. A. Khayam, and M. Farooq, “Keystroke-based user identification on smart phones,” in *Proceedings of the 12th International Symposium on Recent Advances in Intrusion Detection*, ser. RAID ’09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 224–243.

[3] C. Bettini, X. S. Wang, and S. Jajodia, “Protecting privacy against location-based personal identification,” in *In 2nd VLDB Workshop SDM*, 2005, pp. 185–199.

[4] X. Ni, Z. Yang, X. Bai, A. Champion, and D. Xuan, “Diffuser: Differentiated user access control on smartphones,” in *Mobile Adhoc and Sensor Systems, 2009. MASS ’09. IEEE 6th International Conference*, 2009, pp. 1012–1017.

[5] I. Muslukhov, “Survey: Data protection in smartphones against physical threats,” 2012.

[6] P. M. Milligan and D. Hutcheson, “Business risks and security assessment for mobile devices,” in *Proceedings of the 8th Conference on 8th WSEAS Int. Conference on Mathematics and Computers in Business and Economics - Volume 8*, ser. MCBE’07. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2007, pp. 189–193.

[7] T.-Y. Chang, C.-J. Tsai, and J.-H. Lin, “A graphical-based password keystroke dynamic authentication system for touch screen handheld mobile devices,” in *Journal of Systems and Software*, vol. 85, no. 5, 2012, pp. 1157–1165.

[8] C. Vielhauer, *Biometric User Authentication for IT Security: From Fundamentals to Handwriting*, ser. Advances in information security. Springer-Verlag, 2006.

[9] D. Zissis and D. Lekkas, “Addressing cloud computing security issues,” in *Future Generation Computer Systems*, vol. 28. Elsevier Science Publishers B. V, 2000, pp. 583–592.

[10] S. Sonkamble, R. Thool, and B. Sonkamble, “Survey of biometric recognition systems and their applications,” in *Journal of Theoretical and Applied Information Technology*, 2010.

[11] A. Esposito, “Debunking some myths about biometric authentication,” in *arXiv preprint arXiv:1203.0333*, 2012.

[12] S. Simitis, “Bundesdatenschutzgesetz: Bdsg,” Bundesbeauftragte für den Datenschutz und die Informationsfreiheit, Baden-Baden, Tech. Rep., 2011.

[13] A. Makrushin, T. Scheidat, and C. Vielhauer, “Handwriting biometrics: feature selection based improvements in authentication and hash generation accuracy,” in *Proceedings of the COST 2101 European conference on Biometrics and ID management*, ser. BioID’11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 37–48.

[14] F. Monrose and A. D. Rubin, “Authentication via keystroke dynamics,” in *Proceedings of the 4th ACM conference on Computer and communications security*, ser. CCS ’97. New York, NY, USA: ACM, 1997, pp. 48–56.

[15] R. Moskovitch, C. Feher, A. Messerman, N. Kirschnick, T. Mustafic, A. Camtepe, B. Löhlein, U. Heister, S. Möller, L. Rokach, and Y. Elovici, “Identity theft, computers and behavioral biometrics,” in *Proceedings of the 2009 IEEE international conference on Intelligence and security informatics*, ser. ISI’09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 155–160.

[16] C. Epp, M. Lippold, and R. L. Mandryk, “Identifying emotional states using keystroke dynamics,” in *Proceedings of the 2011 annual conference on Human factors in computing systems*, ser. CHI ’11. New York, NY, USA: ACM, 2011, pp. 715–724.

[17] M. Choraś and P. Mroczkowski, “Keystroke dynamics for biometrics identification,” in *Proceedings of the 8th international conference on Adaptive and Natural Computing Algorithms, Part II*, ser. ICANNGA ’07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 424–431.

[18] D. Umphress and G. Williams, “Identity verification through keyboard characteristics,” in *International Journal of Man-Machine Studies*, vol. 23, 1985, pp. 263–273.

[19] F. A. Alsulaiman, J. Cha, and A. Saddik, “User identification based on handwritten signatures with haptic information,” in *Proceedings of the 6th international conference on Haptics: Perception, Devices and Scenarios*, ser. EuroHaptics ’08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 114–121.

[20] M. Faundez-zanuy, “Study of a committee of neural networks for biometric hand-geometry recognition,” in *Neural Networks*, 2005, pp. 1180 – 1187.

[21] S. Karatzouni and N. L. Clarke, “Keystroke analysis for thumb-based keyboards on mobile devices,” in *New Approaches for Security, Privacy and Trust in Complex Environments, Proceedings of the IFIP TC-11 22nd International Information Security Conference (SEC 2007), 14-16 May 2007, Sandton, South Africa*, ser. IFIP, H. S. Venter, M. M. Eloff, L. Labuschagne, J. H. P. Eloff, and R. v. Solms, Eds., vol. 232. Springer, 2007, pp. 253–263.

[22] S. Cho, C. Han, D. H. Han, and H.-I. Kim, “Web-based keystroke dynamics identity verification using neural network,” in *Journal of Organizational Computing and Electronic Commerce*, vol. 10, no. 4, 2000, pp. 295–307.

[23] D.-T. Lin, “Computer-access authentication with neural network based keystroke identity verification,” in *Neural Networks,1997., International Conference on*, vol. 1, 1997, pp. 174 –178.

[24] M. Obaidat and B. Sadoun, “Verification of computer users using keystroke dynamics,” in *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 27, no. 2, 1997, pp. 261–269.

[25] M. Obaidat and D. Macchiarolo, “An online neural network system for computer access security,” in *Industrial Electronics, IEEE Transactions on*, vol. 40, no. 2, 1993, pp. 235–242.

[26] S. Haider, A. Abbas, and A. Zaidi, “A multi-technique approach for user identification through keystroke dynamics,” in *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, vol. 2, 2000, pp. 1336 –1341.

[27] D. Shanmugapriya and G. Padmavathi, “A survey of biometric keystroke

dynamics: Approaches, security and challenges," in *International Journal of Computer Science and Information Security*, vol. 5, no. 1, 2009.

[28] S. Banerjee and D. Woodard, "Biometric authentication and identification using keystroke dynamics: A survey," in *Journal of Pattern Recognition Research*, vol. 7, 2012, pp. 116–139.

[29] R. Joyce and G. Gupta, "Identity authentication based on keystroke latencies," in *Commun. ACM*, vol. 33. New York, NY, USA: ACM, 1990, pp. 168–176.

[30] T. Ord and S. Furnell, "User authentication for keypad-based devices using keystroke analysis," in *Proc. 2nd Int'l Network Conf. (INC 2000)*, 2000, pp. 263–272.

[31] N. L. Clarke and S. M. Furnell, "Authenticating mobile phone users using keystroke analysis," in *Int. J. Inf. Secur*, vol. 6. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 1–14.

[32] H. Saevanee and P. Bhattarakosol, "Authenticating user using keystroke dynamics and finger pressure," in *Consumer Communications and Networking Conference, 2009. 6th IEEE*, 2009, pp. 1–2.

[33] M. Trojahn and F. Ortmeier, "Biometric authentication through a virtual keyboard for smartphones," in *International Journal of Computer Science & Information Technology (IJCSIT)*, 2012.

[34] A. Ross and A. K. Jain, "Information fusion in biometrics," in *Pattern Recognition Letters*, vol. 24, 2003, pp. 2115–2125.

[35] E. Lau, X. Liu, C. Xiao, and X. Yu, "Enhanced user authentication through keystroke biometrics," in *Computer and Network Security*, 2004.

[36] A. Peacock, X. Ke, and M. Wilkerson, "Typing patterns: A key to user identification," in *IEEE Security and Privacy*, vol. 2, no. 5, 2004, pp. 40–47.

[37] D. N. Glaser, B. C. Tatum, D. M. Nebeker, R. C. Sorenson, and J. R. Aiello, "Workload and social support: Effects on performance and stress," in *Human Performance*, vol. 12, no. 2, 1999, pp. 155–176.

[38] Z. Syed, S. Banerjee, Qi Cheng, and B. Cukic, "Effects of user habituation in keystroke dynamics on password security policy," in *High-Assurance Systems Engineering (HASE), 2011 IEEE 13th International Symposium on*, 2011, pp. 352–359.

[39] H. Khurana, M. Hadley, Ning Lu, and D. Frincke, "Smart-grid security issues," in *Security & Privacy, IEEE*, vol. 8, no. 1, 2010, pp. 81–85.

[40] A. Metke and R. Ekl, "Security technology for smart grid networks," in *IEEE Transactions on Smart Grid*, vol. 1, no. 1, 2010, pp. 99–107.

[41] S.-s. Hwang, S. Cho, and S. Park, "Keystroke dynamics-based authentication for mobile devices," in *Computers & Security*, vol. 28, no. 1–2, 2009, pp. 85–93.

Revisiting Mobility, Devices and Business Models

A user-centric perspective

Yan Cimon
CIRRELT
Université Laval
Quebec City, Canada
e-mail: yan.cimon@fsa.ulaval.ca

***Abstract*— Given the array of devices now available to users, how can one make sense of the evolution of user-side mobility? The purpose of this paper is to revisit mobility from a user-centric perspective. We use a mixed methods approach based on action research and case studies. We find that mobility's pervasiveness leads to changes in habits that have a profound effect on devices and business models. We conclude by looking into possible directions for new advances in light of the present research and recent trends in the IT industry.**

Keywords-User-centric mobility; devices; business model; user-centric design.

I. INTRODUCTION

Given the array of devices now available to users, how can one make sense of the evolution of user-side mobility? Mobile devices are now ubiquitous in everyday life and they are seldom noticed anymore [1]. Furthermore, in recent years, consumers have faced an increasing array of devices that are very eclectic as exemplified by the vastly different characteristics and capabilities of tablets, handsets, e-readers and others. At the same time, the characteristics and enablers of mobility are present in ever-increasing settings. These may be related 1) to how the user experiences connectedness, or 2) to the device itself. This is especially important, yet often overlooked, because it has been known for a while that information and perception are intertwined [2].

Whatever the design process, whether from groups of designers to the consumer-market or from co-creation by consumers and designers, a user-centric approach is necessary as consumers are ever more demanding when it comes to mobile devices or mobility in general [3]. This has seemingly led to a paradox where one expressed need for mobility yields to a fragmented market of devices that allow mobility. This entails a simultaneous convergence and divergence in the types of devices that are available. At one end of the spectrum, users acquire different devices for different uses, while, at the other end, other users seek use only one sort of device, for all their needs. Along the same lines, users exhibit various learning patterns [4] and some (in)ability to adapt to change when switching devices. Some are tinkerers and like to exploit and configure any possibility afforded by their device while a significant segment could be termed a plug-and-play crowd happy with general default settings. This explains why understanding mobility, devices and design from a user-centric side and examining business models remain relevant for users and businesses alike.

This paper is structured as follows. First, some related work is examined. Second, in the problem statement section, the general issue being researched is explained. Third, a user-centric view of mobility is put forth. On one hand, the user-side of mobility is examined. On the other hand, issues pertaining to attributes of mobile devices are discussed. Fourth, an overview of the methods used is provided. Fifth, some contemporary business models are discussed in light of the evolution of mobility through mobile devices. Finally, some implications for this work and a way forward for research are presented.

II. RELATED WORK

Related work gives partial insight on how mobility, devices and business models intersect from a user-centric perspective. First, past research on mobility has examined issues related to architecture [5], [6] and awareness management in networks [7]. Furthermore, cloud computing is presented as an option to deal with mobility issues [8].

Second, a strand of literature on devices examines application mobility for cross-network roaming through multiple devices [9] as well as in heterogeneous network environments [10]. Other recent work on devices shows an increased interest for semantic solutions that makes the device a privileged nexus between users and services [11].

Third, while research on business models per se historically focused on value creation [12] business models that may be harnessed and their design [13]. Other work ventured toward mobile payment adoption [14] or demand-related variables [15]. However, cultural factors also affect mobility [16] and users put a high premium on the quality of their experience as evidenced by research on mobile TV[17].

III. PROBLEM STATEMENT

As previous research rarely examined mobility-enabling technology, devices, and business models from a user-centric perspective, the present research makes a contribution by looking at these factors simultaneously. Indeed, making sense of the new trends in devices, platforms and available content, software and applications from the consumer side is increasingly difficult given their diversity. It is not clear yet whether there is a growing convergence or a competitive co-existence of available devices centered on a variety of user needs. While other studies have focused on specific user behaviour issues from a very technical and quantitative perspective [18], this research contributes a different user-centric way of tackling some challenges associated with mobility while looking at devices, users and some trends as well as business models that may be harnessed.

IV. USER-CENTRIC VIEW MOBILITY AND CONNECTEDNESS

Mobility is of an increasing value for users living a connected life. They also wish to remain connected for longer parts of the day. One factor behind this desire is partly explained by the "Fear of Missing Out" (FOMO), or the fear of missing something important to them – or their social circle – should they not be connected [19]. Another factor may be linked to the emotional response related to higher levels of connectedness. Beyond the sense of belonging or security that is attached to connectedness [20], users also find comfort in the fact that being connected via a device is a reassurance. Thus, it is possible to map the sense of belonging on a spectrum that puts peace of mind at the lowest end and active engagement at the highest end. A third factor to consider is that mobile devices are ideal to fulfill certain types of tasks or for personal distraction during micro breaks [21], whether on the move or not.

A. User-related considerations

1) Some psychological considerations: Many psychological factors may be associated with mobility. The feeling of convenience that comes with increased mobility may bring a higher level of psychological comfort, even in users whose lifestyle does not require a high level of mobility. As such, mobile devices may be construed as extensions of the self or as influencing it [22]. Different devices broadcast different messages. As Apple's product appeals more to emotion and a desire to be part of a certain "in" crowd, MS Surface tries to cater to "cool" business users that seek to interface their device in a familiar Microsoft Office environment, Samsung counts on the pervasiveness of Android and some "fun" characteristics associated to its products. Finally, there is a level of status that is associated with the type of device one uses and its level of customization.

2) Anchoring: Another consideration is the need for an anchor that users have. Traditionally, this anchor was in a physical space, such as a home for example. This provided a sense of place in the world and an emotional anchor that too made this sense of belonging important [23]. As users lead increasingly dynamic lives, where geography is losing ground to information technologies, the traditional need for physical/geographic anchoring has morphed into one for digital anchoring as it remains now the "constant" in many users lives. A direct implication is that one may expect users to develop an emotional attachment to their devices and that this will impact the physical characteristics that they will want in a mobile device as this will become a reflection of their identity.

B. The device side

Many conditions exist to enable mobility in the consumer's mind when it comes to devices.

1) Physical attributes: First of all, the physical attributes of a device matter much for adoption and for its intended usage patterns. The first factor that matters here is purely physical. The devices weight and their ease of manipulation matter. They are the crux of physical characteristics (Palm CEO famously carried a wooden replica of a device in his pocket to find an optimal design [24]). There is a trend toward more embeddedness of mobility in objects that are beyond the realm of classical mobile devices, as evidenced in the automotive industry (Hyundai BlueLink, MyFordTouch, etc.), but this embeddedness brings challenges in terms of obsolescence and updating needs. As in the car industry where embedded electronics become obsolete faster than the engine or mechanical parts.

2) Connectedness: Second, connectedness matters a lot too. 3G and 4G connectivity on a wide range of carriers are important to ensure mobility over large areas as evidenced in the USA by the wars between Verizon, AT&T and Sprint over coverage and speed. Furthermore, more and more consumers expect fast and easy WiFi connectedness in tablets and phones (which is fairly standard nowadays) but also in a range of other devices.

3) Software: Third, the issue of software that is not platform dependent, many would say "cloud ready", is a very important desirable characteristic. Information access and manipulation are cornerstones of the perceived value of mobility. The ability to manage structured information and unstructured information is very important but often overlooked. For instance, software that allows for unstructured or fuzzy queries close to natural language queries is also a desirable characteristic to find specific information on the device. It is important that queries beyond simple keywords be supported. Furthermore, social possibilities afforded by the device, i.e., easy connection to social media [25] and social networks are another element that needs to be factored in (Facebook, LinkedIn, Pinterest, Reddit, etc.)

V. METHODS

This research uses a mixed methods approach to serve its exploratory nature. The combination of methodologies is felt to allow better insights into the complexity and richness of the subject matter in order to move beyond purely technical challenges.

The first method used is that of action research [26] of an inductive nature. Over a full year, mobile devices were used in three different countries (Canada, USA and Switzerland) on major commercial platforms (BlackBerry, Android, and iOS) linked to major national providers (Bell, Verizon, and Swisscom).

The second method used is case study research [27]. The cases were selected for the breath of insight they could provide and their diversity (for-profit, non-profit, maker of devices and software). The cases were examined through secondary data sources (newspaper articles, video interviews by third parties, etc) that would shed light on both the user experience and devices.

Important elements like validity, does the research measure what it claims to be measuring, and reliability, does it measure it in a consistent matter, were taken into account [28]. Care was taken to control for biases [29] by adopting a comparative perspective. There are no conflicts of interest to declare with regards to the cases discussed in this paper.

VI. Perspectives on Mobility's Evolution and Some Contemporary Business Models

A. *Consumer-side mobility*

Consumer-side mobility has evolved in cycles. In the 1990's, mobility from the user side essentially came in the form of mobile phones, or mobile handsets and bulky laptops with mobile radios being phased out. In the 2000's, it became a defining characteristic of many devices (PDAs, mp3 players, etc.) that saw their functions converge into "smart" phones and eventually became a noted feature of a range of devices (tablets, e-readers, etc). That trend is morphing into embedded mobility (in cars, public spaces, in airplanes, etc), with the next step likely being improved wearability of devices (Google glasses or intelligent clothing). While wearable computing has been around for a long time, the next generation may prove better adopted as users become used to well-designed I/O apparatus that will make this sort of mobility both sensible and natural.

B. *Business models*

These changes in consumer-side mobility have a great impact on business models. For wireless carriers, it means more devices are connected to their networks and thus a rise in their traditional businesses of carrying voice and data. But this also implies new value streams that come from a different way/process of interacting with users: different fee schedules to fit different uses and devices and different ways of modulating the contracting agreements. Device makers can now make money on the physical devices themselves, but also derive advantages from sharing platforms/OS. Applications (apps) and software developers, also have new possibilities to maximize downloads and adjust payment possibilities by moving from free to fee-based models. Last, but not least, ancillary material and peripheral makers also have more devices to cater to and are able to better take advantage of consumers' need for personalizing their devices.

C. *Cases*

Three cases were examined in light of the research question: Major League Baseball, OneBusAway, and BlackBerry.

1) Major League Baseball: A first case is that of Major League Baseball (MLB). Traditionally, the league made a lot of revenue from TV broadcasting rights. But as consumers migrate from – or increasingly use other platforms than – TV, the league created MLB Advanced Media (BAM) to distribute content on a variety of devices. This means that BAM is able to create value by leveraging apps, providing a lot of content that allows for a great interactive experience (videos, etc.) beyond the traditional statistics baseball fans love [30], [31]. In doing so, they also take time to better understand users to better cater to their interests. For example, their data showed them that different devices were associated with different user behaviours and thus different needs: e.g., users with mobile phones did not have the same usage patterns as those with tablets or laptops [32].

2) OneBusAway: A second case is that of OneBusAway [33], a non-profit that developed the eponymous app that provides real-time transit information. Transit schedules are often complicated for users to understand, especially in large cities that have a denser transit network. A wide range of users, from daily commuters to tourists, need to plan their journeys. With the strong penetration of mobile devices in major American metropolitan areas, it made sense to develop an easy-to-use app that would assist users in planning their trips in a real-time manner. The app was developed for the Puget Sound area and now covers Atlanta, New York and Tampa.

3) BlackBerry: A third and last case is that of BlackBerry [34]. Better known for its handsets, it also ventured in the tablet market, but users did not follow. After market share losses and devices that were not appealing to younger generations, it remained popular with business users. It reinvented its handsets including touchscreens and getting rid of their trackpad. It proposed a model with a keyboard (Q10), another one without (Z10) and a third at a lower price point (Q5) [35]. It fielded a new OS that allows multitasking, provided a secure workspace that can operate with other OS' than BB10. It provided a better app ecosystem for example adding Skype and the new BlackBerry Messenger for real-time communications. It did try to cater better to user needs and habits. The market will tell if it is valued by consumers [36] underlying the fact that reality is the ultimate test.

D. *Premilinary findings*

The cases that were examined seem to confirm two general trends. First, there appears to be a heavy and cross-platform adoption of intuitive user-centric mobility by consumers. Second, firms do alter significantly or create entire business models based on this shift instead of just paying lip service to this trend. These models create value beyond simple micropayments.

VII. Conclusions and Implications

These cases are revealing because they are simultaneously enabled by and enablers of mobility. This self-reinforcing mechanism has many implications for business, academics and future research efforts.

A. *Implications for business*

Better configured business models taking a user-centric perspective are bound to maximize value creation. Fee structures and modulations along each segment of the value chain will enable better decision-making on hardware, on software, and on ancillary services and apps from the ecosystem surrounding each device. User-centricity also matters a great deal when it comes to critical decision making [37], especially when mobile devices are involved.

B. *Implications for academics*

Academics need to help industry better bridge the gap between the device, the user and their perceptions, the ecosystem that surrounds them, and business models. This research constitutes a step in this direction. Furthermore, since devices – and the software that they run – are produced in complex networks of firms that collaborate, then working from a common user-centric perspective may in the end reduce the conflicts arising from these firms' asymmetries [38] and lead to a more coherent, yet highly value added, ecosystem of devices, applications and more.

C. *Future research*

In conclusion, developing a user-centric view of mobility that simultaneously takes into account users and devices is useful to understand how mobility and business models that underpin it may deliver more value for specialized applications like the Physical Internet [39].

Thus, a user-centric perspective in coming research will be paramount to finding new ways of harnessing important contemporary trends such as 1) Bring Your Own Devices (BYOD) [40] to work environments, 2) An increasing pervasiveness of mobility and information combined [41], and 3) challenges brought by the Internet of things [42].

These will force a dramatic rethinking of business models around value creation from intangibles [43] that are focused on user habit "dynamics", i.e., how user habits and preferences change over time. Another relevant dimension will be a careful examination of the interaction between mobility and contents access, especially since a very real possibility of commoditization of "smart" devices may occur. Finally, the interaction of mobility and potentially disruptive technologies like 3D printing and high precision distributed manufacturing may also prove to be a game changer.

Acknowledgments

The author would like to thank the *Fonds de recherche québécois – société et culture* (FRQSC), the *Soutien à la Recherche* fund from *Université Laval* for funding that enabled portions of this research. Some of this research was carried out while the author held the Fulbright Visiting Chair in Innovation at the University of Washington (Seattle) and visited McGill and EPFL.

References

[1] V. Oksman and P. Rautiainen, "Perhaps it is a Body Part: How the Mobile Phone Became an Organic Part of the Everyday Lives of Finnish Children and Teenagers," Machines that become us: The social context of communication technology, 2003, pp. 293-308.

[2] J. E. Cutting, "Perception and Information," Annual Review of Psychology, vol. 38, 1987, pp. 61-90.

[3] Y. Cimon, F. Z. Barrane, and P. Diane, "Meeting the Challenge of Global Mobile Phone Usability: Design and practices," in Proceedings - MOBILITY 2011, Barcelona, 2011, pp. 123-126.

[4] R. Wang, R. Wiesemes, and C. Gibbons, "Developing digital fluency through ubiquitous mobile devices: Findings from a small-scale study," Computers & Education, vol. 58, 2012, pp. 570-578.

[5] Q. Wei and Z. Jin, "Service discovery for internet of things: a context-awareness perspective," Proceedings of the Fourth Asia-Pacific Symposium on Internetware, 2012, pp. 1-6.

[6] S. Malek, G. Edwards, Y. Brun, H. Tajalli, J. Garcia, *et al.*, "An architecture-driven software mobility framework," Journal of Systems and Software, vol. 83, 2010, pp. 972-989.

[7] J. W. Mwangoka, P. Marques, and J. Rodriguez, "Cognitive mobility management in heterogeneous networks," Proceedings of the 8th ACM international workshop on Mobility management and wireless access, 2010, pp. 37-44.

[8] N. Fernando, S. W. Loke, and W. Rahayu, "Mobile cloud computing: A survey," Future Generation Computer Systems, vol. 29, 2013, pp. 84-106.

[9] A. Ahlund, K. Mitra, D. Johansson, C. Ahlund, and A. Zaslavsky, "Context-aware application mobility support in pervasive computing environments," Proceedings of the 6th International Conference on Mobile Technology, Application & Systems, 2009, pp. 1-4.

[10] J. Flinn, T. J. Giuli, B. Higgins, B. Noble, A. Reda, *et al.*, "The case for intentional networking," Proceedings of the 10th workshop on Mobile Computing Systems and Applications, 2009, pp. 1-6.

[11] A. Toninelli, A. Corradi, and R. Montanari, "Semantic-based discovery to support mobile context-aware service access," Computer Communications, vol. 31, 2008, pp. 935-949.

[12] R. Amit and C. Zott, "Creating Value Through Business Model Innovation," MIT Sloan Management Review, vol. 53, Spring2012 2012, pp. 41-49.

[13] A. Osterwalder and Y. Pigneur, "Designing Business Models and Similar Strategic Objects: The Contribution of IS," Journal of the Association for Information Systems, vol. 14, 2013, pp. 237-244.

[14] C. Kim, M. Mirusmonov, and I. Lee, "An empirical examination of factors influencing the intention to use mobile payment," Computers in Human Behavior, vol. 26, 2010, pp. 310-322.

[15] J. Iden and L. B. Methlie, "The drivers of services on next-generation networks," Telematics and Informatics, vol. 29, 2012, pp. 137-155.

[16] J. Blom, J. Chipchase, and J. Lehikoinen, "Contextual and cultural challenges for user mobility research," Commun. ACM, vol. 48, 2005, pp. 37-41.

[17] S. Buchinger, S. Kriglstein, S. Brandt, and H. Hlavacs, "A survey on user studies and technical aspects of mobile multimedia applications," Entertainment Computing, vol. 2, 2011, pp. 175-190.

[18] M. Vojnovic, "On Mobile User Behaviour Patterns," Microsoft Research - Technical Report MSR-TR-2008-08, 2008, p. 5.

[19] J. J. Kandell, "Internet addiction on campus: The vulnerability of college students," CyberPsychology & Behavior, vol. 1, 1998, pp. 11-17.

[20] R. Wei and V.-H. Lo, "Staying connected while on the move Cell phone use and social connectedness," New Media & Society, vol. 8, 2006, pp. 53-72.

[21] Y. Cui and V. Roto, "How people use the web on mobile devices," in Proceedings of the 17th international conference on World Wide Web, 2008, pp. 905-914.

[22] L. A. Jackson, A. von Eye, H. E. Fitzgerald, Y. Zhao, and E. A. Witt, "Self-concept, self-esteem, gender, race and information technology use," Computers in Human Behavior, vol. 26, 2010, pp. 323-328.

[23] A. Williams, K. Anderson, and P. Dourish, "Anchored mobilities: mobile technology and transnational migration," in Proceedings of the 7th ACM conference on Designing interactive systems, 2008, pp. 323-332.

[24] K. J. Vicente, The Human Factor: Revolutionizing the Way People Live with Technology: Taylor & Francis, 2013.

[25] Y. Cui and M. Honkala, "A novel mobile device user interface with integrated social networking services," International Journal of Human-Computer Studies, vol. 71, 2013, pp. 919-932.

[26] M. Brydon-Miller, D. Greenwood, and P. Maguire, "Why action research?," Action research, vol. 1, 2003, pp. 9-28.

[27] R. K. Yin, Case Study Research: Design and Methods: SAGE Publications, 2009.

[28] E. G. Carmines and R. A. Zeller, Reliability and Validity Assessment: SAGE Publications, 1979.

[29] N. K. Denzin and Y. S. Lincoln, The SAGE handbook of qualitative research: Sage, 2011.

[30] M. Ozanian. (2013, May 26th) Baseball's Next Home Run: Chatting Cage. Forbes.com. Available: http://www.forbes.com/sites/mikeozanian/2013/05/26/baseballs-next-home-run-chatting-cage/ [retrieved: June, 2013]

[31] C. Salter. (2012, March 19th) MLB Advanced Media's Bob Bowman is playing digital hardball. And he's winning. Fast Company. Available: http://www.fastcompany.com/1822802/mlb-advanced-medias-bob-bowman-playing-digital-hardball-and-hes-winning [retrieved: May, 2013]

[32] "Charlie Rose: A conversation with Bob Bowman," ed. NYC: PBS, Jul 4th, 2013.

[33] OneBusAway. Available: http://onebusaway.org/ [retrieved: March, 2013]

[34] BlackBerry. Available: http://global.blackberry.com/sites.html [retrieved: June, 2013]

[35] R. Cheng. (2013, May 18th) How BlackBerry is fixing its once 'broken' brand. CNET. Available: http://news.cnet.com/8301-1035_3-57585099-94/how-blackberry-is-fixing-its-once-broken-brand/ [retrieved: June, 2013]

[36] H. Miller and M. McMahon. (2013, Jul 1st) BlackBerry Shares Plunge After Touch-Screen Model Flops. Bloomberg. Available: http://www.bloomberg.com/news/2013-07-01/blackberry-shares-plunge-after-touch-screen-model-flops.html [retrieved: July, 2013]

[37] S. Veronneau and Y. Cimon, "Maintaining robust decision capabilities: An integrative human-systems approach," Decision Support Systems, vol. 43, 2007, pp. 127-140.

[38] Y. Cimon, "Knowledge-related asymmetries in strategic alliances," Journal of Knowledge Management, vol. 8, 2004, pp. 17-30.

[39] B. Montreuil, J.-F. Rougès, Y. Cimon, and D. Poulin, "The Physical Internet and Business Model Innovation," Technology Innovation Management Review, 2012, pp. 32-37.

[40] K. W. Miller, J. Voas, and G. F. Hurlburt, "BYOD: Security and privacy considerations," IT Professional, vol. 14, 2012, pp. 53-55.

[41] D. Saha and A. Mukherjee, "Pervasive computing: a paradigm for the 21st century," Computer, vol. 36, 2003, pp. 25-31.

[42] M. Zorzi, A. Gluhak, S. Lange, and A. Bassi, "From today's intranet of things to a future internet of things: a wireless-and mobility-related view," Wireless Communications, IEEE, vol. 17, 2010, pp. 44-51.

[43] A. Lapointe and Y. Cimon, "Leveraging intangibles: how firms can create lasting value," Journal of Business Strategy, vol. 30, 2009, pp. 40 - 48.