Collaborative Location Aware Mobile Services

Konstantas, Dimitri (ed.); Seigneur, Jean-Marc (ed.)

UNIVERSITÉ DE GENÈVE

Collaborative Location Aware Mobile Services

Edited by Dimitri Konstantas & Jean-Marc Seigneur

Advanced Systems Group
Centre Universitaire d'Informatique
Department of Information Systems
Faculty of Social and Economic Sciences
2008

University of Geneva - Faculty of Social and Economic Sciences
Department of Information Systems
Centre Universitaire d'Informatique *Phone:* +41 22 379.01.02
Battelle - bâtiment A *Fax:* +41 22 379.02.33
7, route de Drize
CH-1227 Carouge *WWW:* http://asg.unige.ch/
SWITZERLAND

Advanced Systems Group (names and emails):

Dimitri Konstantas (Dimitri.Konstantas@unige.ch)
Jean-Marc Seigneur (Jean-Marc.Seigneur@unige.ch)
Lemonia Ragia (Lemonia.Ragia@unige.ch)
Michel Deriaz (Michel.Deriaz@unige.ch)
Katarzyna Wac (Katarzyna.Wac@unige.ch)
Alfredo Villalba (Alfredo.Villalba@unige.ch)
Hikari Watanabe (Hikari.Watanabe@unige.ch)
Dejan Munjin (Dejan.Munjin@unige.ch)
Xavier Titi (Xavier.Titi@unige.ch)

Co-authors from other institutes:
Tewfiq El Maliki – Ecole d'ingénieurs de Genève HES-SO, CH
Richard Bults, Hong Chen, Bert-Jan van Beijnum, Aart van Halteren
Lennart Isaksson, Pravin Pawar, Mortaza Bargh, Arjan Peddemors and Hermier Hermens – University of Twente, Enschede, NL
Pierre Maret – INSA, Lyon, FR
Elias Kalapanidas and Costas Davarakis Systema Technologies, Athens, GR
Hannes Kaufmann ,Vienna University of Technology, Wien, AT
Fernando Fernandez Aranda, Hospital Universitario de Bellvitge, Barcelona, ES
Tony Lam, NetUnion SARL, Lausanne, CH
Todor Ganchev, University of Patras, Rion-Patras, GR
Giovanna Di Marzo Serugendo - School of Computer Science and Information Systems, Birkbeck College, London, UK

The financial assistance of the following grants from the Swiss Federal Government and the European Union is gratefully acknowledged:
OFFT CTI-Project no. 8643.2 PFES-ES *Geo-Tags: Trusted Geo-Spatial Tags*
SER Action COST IC0703, SER project No. C08.0025 *Monitoring and Analysis of wireless networks for a QoS-Prediction Service design and development*
FP7-ICT-2007-2-224024 *PERIMETER* project
FP7-ICT-215839 *PlayMancher* Project

ISBN 978-2-88903-001-9 Editions SES - Universite de Geneve

Imprimé par l'Atelier d'Impression de l'Université de Genève
October 2008

Table of Contents

Preface

Mobile devices and services are gaining every day more and more users with new services and applications appearing every day. The high success of new powerful mobile hardware able to support innovative services, like the iPhone and its competitors Blackberry and HTC, has shown that the users are anxiously waiting and are ready to adopt new (mobile) services and applications. Today mobile services are more complementary of the fixed network services, allowing the users to access traditional fixed network based services, while being mobile. However now that the users have been accustomed in mobile services and content access and in response to their needs, new services and new challenges start appearing that do not have an equivalent fixed wire service, ranging from location based trust services to middleware for network handover. In addition mobile users are recreating the very successful operation of internet, where services are based on a collaborative model for the creation and exchange of (mobile) content.

The main research effort of our group is based on the vision of the future needs of collaborative mobile services. In this report we present 10 papers and one position paper, representing the main work areas of the group. Our work is mostly based on the notion of virtual tags as the basic collaboration unit of mobile users.

The first paper, *The Uncertainty of the Truth*, provides investigates the issues related to the trust level one must have into mobile tags, in the absence of any reference regarding the author of the tag, and how a collaborative service can be implemented on this basis.

The second paper, *Mobile Location Based Services for Trusted Information in Disaster Manageme*nt, presents a prototype service for disaster management in road traffic control, discussing the related issues and problems.

In the third paper, *User-centric Mobile Identity Management Services*, we discuss the issues related to mobile identity management, providing a survey how the requirements have evolved for mobile user-centric identity management and their associated technologies.

In the forth paper, *LSPEnv: Location-based Service Provider for Environmental Data*, we present an approach for forecasting environmental data for location based services and we propose a system for making predictions for spatial-temporal variables using the Bayesian Network method as a machine learning.

In the fifth paper, *Context-Aware Middleware Architecture for Vertical Handover Support to Multi-homed Nomadic Mobile Services*, we propose a context-aware middleware architecture supporting vertical handover for the Nomadic Mobile Service providers hosted on the handheld mobile devices, based on a context-aware computing approach.

In the sixth paper, *Collaborative QoS-information Sharing for Mobile Service Users: A Web 2.0 Business Model proposal*, we discuss the business related issues and problems, and present different business scenarios for the commercial implementation and exploitation of a mobile network QoS prediction service.

In the seventh paper, *Power- and Delay-Aware Mobile Application-Data Flow Adaptation: the MobiHealth System Case Study*, we report on a case study of a cardiac telemonitoring application delivered by the MobiHealth system. Our results show the trade-off between the delay and battery savings achieved by various network inteface activation strategies in combination with application-data flow adaptation

The next two papers present the continuation of our research in the concept of Hovering Information. In the eight paper, *Hovering Information - Self-Organising Information that Finds its Own Storage*, we discusses results of simulations performed for two algorithms aiming to ensure the availability of a piece of hovering information at its anchor area.

In the ninth paper, *Hovering Information - Infrastructure-Free Self-Organising Location-Aware Information Dissemination Service*, we discuss issues related to the scalability of Hovering information and presents the results on a series of simulations involving multiple pieces of hovering information.

The last two papers present an overview of two 7th framework European projects in which we participate. In the tenth paper, *PlayMancer : A European Serious Gaming 3D Environment*, we present the project PlayMancer, targeting in the design and implementation of a platform facilitating the development of multiplayer, network serious 3D games and its validation in two different medical cases. The last paper, User Experience and Emotion-AwareBusiness Network Service Selection, gives a short of the PERIMETER project, targeting the establishment of a new paradigm of user-centricity for advanced networking, putting the user at the centre rather than the telecom operator.

Dimitri Konstantas
October 2008

The Uncertainty of the Truth[1]

Michel Deriaz

Abstract. How to trust without knowing the truth? This is probably the key question that arises while designing applications using virtual tags. A virtual tag is a geo-referenced note that is visible for all the people that are in a specific place. But what if you see a tag about an event or an object that is not here? How to know if you are facing a spam attack, or if the tag is simply outdated? And, how to update the trust values of the author and the other people that confirmed the tag, since you do not know if they are honest? To answer these questions, we designed and implemented FoxyTag, a free and collaborative system which consist in posting virtual tags over speed cameras in order to warn the other drivers. We used it to test our new generic trust engine and got very promising results.

1 Introduction

Spatial messaging, also called digital graffiti, air graffiti, or splash messaging, allows a user to publish a geo-referenced note so that any other user reaching the same place gets the message. For example, let us consider the community of the Mt-Blanc mountain guides. The members would like to inform their colleagues about dangers in specific places. One guide publishes a geo-referenced message that informs about a high risk of avalanches, and any other guide that goes to the same place will get the warning, and comment it if necessary. Spatial messaging is a kind of blog in which editors and readers share the same physical place.

There are many reasons to believe that spatial messaging will become a widespread concept in a nearby future. Today, people use the connection capabilities of their mobile phone mostly in one way, to download information. But in the same way that people passed from television to Internet, the next generation of users will probably become more "active" and create new content with their mobile phones. We already observe this tendency today for specific cases, like sharing pictures or videos recorded by mobile phones and published on some websites. If we remember how fast the computer power and the communication capabilities of these little devices increase, we can easily paint a glorious future for mobile technology.

We will see in the related work section that lots of non critical applications are already running on mobile technology. We insist here on the "non critical" aspect;

[1] The work presented in this paper is supported by the Swiss Federal Office for Professional Education and Technology (OFFT) under the CTI-Project no. 8643.2 PFES-ES *Geo-Tags: Trusted Geo-Spatial Tags*.

This paper has been presented and published in the proceedings of the Sixth Annual Conference on Privacy, Security and Trust, PST2008, October 1-3, 2008, Delta Fredericton, Fredericton, New Brunswick, Canada.

it clearly implies that there is today no third party that proposes any serious application using virtual tags.

To our view the reason is simple: We cannot trust their tags. We do not talk about POI (Points Of Interest), data that is usually provided by a unique source and copied in the devices. We talk here about virtual tags, pieces of information that can be posted by unknown users and modified by other unknown users. These virtual tags are posted in a collaborative way, like it is done in the Google Earth Community [1] where every user can post any geo-referenced information. But, we observe then that we cannot trust this information. Security tools are not sufficient; even if you can be sure about the identity of an author, it is useless if you do not know him and therefore cannot trust the content of his message.

And trusting virtual tags is not that easy. We will see that applying "conventional" trust algorithms doesn't work. One reason is because of what we call the "Uncertainty of the Truth". For instance, a user that sees a tag that warns about a danger of avalanche in mid-summer doesn't know if it is spam (in which case he must decrease the trust value of the tag's author) or if the tag is simply outdated.

2 Related Work

At least to our knowledge, we are the first to study the trust aspects in spatial messaging. Actually, even if we type only "spatial messaging" in Google [2], the first results point directly to our former papers. We find also some people that use our definition to describe it, like for instance in the alvafilm website [3]. So if we add the trust component to spatial messaging, we reduce even more the chances of finding some parallel work.

Since we couldn't find any similar work, we divided this section in three parts. The first part describes other work done for spatial messaging. The second part gives a state of the art in the trust domain. And finally, since we used a speed cameras warning system in order to test our models, the third part gives a list of other warning systems.

2.1 Spatial Messaging

Before starting this section, we would like to precise the difference between spatial messaging and LBS (Location Based Services). In short, LBS is a kind of spatial messaging in which the user can only get data, and not post it. Lots of LBS applications for augmented cities (tourists get information, in their mother tongue, about their current place) or augmented museums (visitors get information about what they are looking at) have already been implemented. We are clearly interested in spatial messaging in general, where users also post information.

2.1.1 E-Graffiti.

E-Graffiti [4] is a spatial messaging application that allows a user to read and post geo-localized notes. These notes can be either public or private, meaning that only the set of people defined by the author are able to read the note.

E-Graffiti has been designed to study the social impacts on spatial messaging. 57 undergraduate students were given a laptop with E-Graffiti for a semester. All their activity has been logged and studied. And the results are far from encouraging. At the end of the semester, it came out that a user logged into the system only 7.6 times in average (std dev: 12.6), and that actually most of the users stuck to initial test messages. Another disappointment was that most of the posted notes were not related to their position. For example, a number of people posted notes to advertise a website. The system was designed so that the user could only get messages available at his current position, but it was possible to post a new message at any place from anywhere.

Technically, the position of the user is determined by the wireless access point to which the device is connected. The precision is therefore limited to the building in which the user is.

2.1.2 GeoNotes

GeoNotes [5] has more functionalities than E-Graffiti. While posting a note, the user can choose how he is going to sign it (for privacy reason the user can write any text he wants as a signature), decide whether people are allowed to comment it, and decide whether anyone can remove this message. For the readers, the graphical interface of the application provides some interesting functionalities like showing all the neighboring messages or sort them according to different criteria. Inspired by the E-Graffiti evaluation, GeoNotes discarded the remote authoring of tags as well as the possibility to "direct" notes to certain users.

The main interest of the GeoNotes authors seems to be the navigation problems in the virtual messages space. How to find a specific note? How to select only relevant messages? One answer of these questions consists in giving to the readers the possibility of ranking the notes. Each user maintains also a friends list, which can be used as a filter. But the trust and security aspects have not been taken into account. It is easy to usurp someone's identity and post false notes. An analysis of a GeoNotes log made during a real-use study showed that 6% of the messages have been signed using someone else's identity.

2.1.3 ActiveCampus Explorer

ActiveCampus Explorer [6] goes a step further by displaying also where other users are. Every user holds a PDA and its location is determined by comparing the signal strength of different wireless access points. Thus, the system knows the position of all its users, and communicates this information to the all of them that are close together. Like E-Graffiti and GeoNotes, it is also possible to tag objects.

2.1.4 Socialight

Socialight [7] allows a user to post some data to a specific place, intended for himself, for his friends, or for everybody. Meta-data containing keywords and geographical coordinates are attached to the posted data, in order to facilitate searches. Tags are called "Stickyshadows" and can be viewed with some specific mobiles phones (and equipped with a positioning system) via the Socialight Mobile application, or by browsing the Socialight website. A nice feature they provide consists in showing Stickyshadows on maps.

2.1.5 Context Watcher

Context Watcher [8] is a mobile phone application written in Python for Nokia Series 60 based on the MobiLife framework [9]. The first version of this application already uses the notion of confirmed buddy for security and trust purposes. They have a part that they called trust engine in their architecture but a closer look at it shows that it is actually only an access control system. Policies and profiles are used to decide who can access what data and under what condition, but there is no trust mechanism that informs how reliable a requested information is.

2.1.6 Summary

These projects don’t seem to be successful. E-Graffiti and GeoNotes have been abandoned shortly after their launch. Socialight is still active, but there are seldom new posts. We believe that the lack of success is related to the lack of interest... in publishing notes just for publishing notes! Spatial messaging would probably have more chance to emerge if we focus on specific communities, with real problems that could be solved by this concept, rather than imposing the system to students without giving them any good reason to use it. But then we need a trust mechanism to exclude malevolent users. In GeoNotes people may stay anonymous, but we saw that users then usurped others’ identities; it is therefore not possible to trust a message. In E-Graffiti users reveal their real identity, but it useless to know that a message has been posted by a certain "John" if you do not know John.

Commercial systems usually implement all the conventional security tools (username, password), but there is no trust engine that informs about the reliability of a given message. It means that it is always a human that plays the role of the trust engine and that excludes what he thinks are malevolent users.

However, in widely deployed systems (like for instance our FoxyTag [10] application that informs about speed cameras in all Europe) where there is only very little human interaction, only a trust engine can ensure a high quality of the data.

2.2 Trust

Trust is a very active research domain. It started by providing solutions for centralized systems (for instance the reputation system in eBay where seller and buyer can rate one another after a transaction), and then quickly switched to peer-to-peer systems. Peers rate each other and the combination of all the values informs about the reputation of the peer. The challenge here is where to store trust values, as there is no central server. Among the proposed solutions, we mention here a few of them:

In EigenTrust [11] each peer has a set of mother peers responsible for storing its trust value, and therefore each peer acts also as a mother peer for others. It resists to an attack even when up to 70% of the peers are colluding in order to subvert the system. Peers are anonymous.

An interesting system that is similar to EigenTrust, but in which peers store their own trust value locally, is called Elicitation-Storage [12]. The Elicitation-Storage protocol is used to protect cryptographically the trust value. The requester gets the IP address of the former requesters and checks with them the authenticity of their vote.

The Secure project [13] aimed to describe in a formal way what trust is, staying as close as possible to the human notion of trust. The motivation for the project was that the number of entities in Internet systems is becoming very large. Consequently, it was important to develop security models that allow nodes to measure the risk involved in interacting with other nodes that they have not met before. The secure project implementation has been tested with a mail application: A proxy between the peer and his mailbox was analyzing the behavior of the user (for instance if he moved a message in his spam folder) and updated the trust values according to it. The reputation system allowed the different peers to share their information in order to exclude faster the spammers.

Kinateder and Rothermel [14] present a peer-to-peer system that provides trust and recommendations about different categories of topics. Similar to sites like Epinion.com or the rating system that we find in eBay, but peer-to-peer.

The TrustMe protocol [15] builds trust in peer-to-peer networks. The trust value of a specific peer is anonymously stored on another peer. Communications are encrypted using sets of private/public keys. The drawback is that all peers have to connect to a bootstrap server when they join and when they leave the network (in order to transmit the hosted trusted values to another peer).

Anwitaman Datta, Manfred Hauswirth and Karl Aberer present in [16] how P-Grid can be used to implement a distributed PKI (Public Key Infrastructure), enabling c2c (customer to customer) services like eBay but without any centralized system. Unlike PGP that uses the web of trust approach to access a particular public key, this system uses a statistical method; many peers are queried, and the information is rejected if a quorum a peers cannot be obtained.

Very interesting and promising decentralized solutions like the EigenTrust algorithm made the community to forget one aspect that is only seldom taken into account: time. In practice time is important. Someone you trusted a long time ago is perhaps not trusty anymore. Even people with a very high reputation can become malevolent afterwards. Since in human communities the trust is very time dependent, we believe that this component should also be included in trust engines and particularly in the spatial messaging context where posted information can simply become obsolete after a while.

Guha [17] built a generic trust engine allowing people to rate the content and the former ratings. He recognized however that in case of highly dynamic systems (like in spatial messaging where tags can appear and disappear very quickly), "Understanding the time-dependent properties of such systems and exploiting these properties is another potentially useful line of inquiry." Most existing trust metrics update their trust values only after a specific action, like a direct interaction or the reception of a recommendation. The few trust engines that take the time component into consideration simply suggest that the trust value decreases with the time.

Mezzetti's trust metric [18] consists in multiplying the trust value at time *t* by a constant between 0 and 1. In Bayesian-based trust metrics [19, 20], the trust value converges to its initial value over time. All these models work in situations where the changes occur slowly, but are challenged in short-lived cases. Unlike the spatial messaging community that seems to be less and less active, the trust community seems to grow and commercial applications are more and more interested in their work. We find for instance some attempts to add trust in Wikipedia articles, like it is presented in a paper from Pierpaolo Dondio, Stephen Barrett, Stefan Weber and Jean-Marc Seigneur [21]. However, we haven't found yet any work on trust in the spatial messaging domain.

2.3 Speed Cameras Warning Systems

As the number of speed cameras increases on European roads, we find more and more services that help the drivers avoiding expensive pictures. We will talk neither about illegal means (for the majority of European countries), like the radar detectors provided by RadarBusters [22], nor about non-technical means like phone centrals providing vocal information. We will concentrate here only on information systems that inform drivers about speed camera positions, which is completely legal according to the law of most European countries.

2.3.1 Mogoroad

Mogoroad [23] is a well-known system in Switzerland to announce traffic perturbations, police controls, and of course fixed and mobile speed cameras. It works on most mobile phones. They collect their information from different partners, like radio stations and newspapers, as well as from their own users that can either signal an event by phone or through an application running on mobile phones. There is no trust engine to validate the quality of the data. According to their CEO, Roberto Marra, it is the experience of the employees that collect the data that is used to differentiate useful and correct information from spam. In practice this works quite well since the covered area is small. However, such a system could not easily be extended to work worldwide while providing the same quality of information. The cost of this service is (in 2008) about 110 € per year.

2.3.2 SmartSpeed

SmartSpeed [24] is an application running on Windows Mobile that informs the driver about dangerous zones, traffic jams, and speed cameras. Working with all NMEA compatible Bluetooth GPS, the program compares the current position with the "events" to come and informs the user through a voice synthesizer. Maps and "events" files can be downloaded in advance, and a GPRS connection allows the user to get recent information. An interesting functionality allows any user to send a new event to the server, which will in turn inform all the users. A typical use consists in signaling mobile speed cameras to other drivers. Even if presented differently, it is clearly a way of doing spatial messaging.

The light version a SmartSpeed is relatively cheap (30 € including free updates for one year) if you possess already a smartphone and a Bluetooth GPS. However, messages sent by other users to signal mobile speed cameras are not verified and are available only for one hour. And users are not really motivated to post such messages since they have nothing to gain in signaling a new "event". SmartSpeed seems more adapted to signal fixed speed cameras than mobile ones.

2.3.3 Coyote

Coyote [25] is an independent system sold as a little box containing a GPS. When the driver approaches a speed camera, Coyote informs him orally about the remaining distance to this camera. To signal a new speed camera (or a new position for a mobile one), the user can simply press once the button on the top of the box. To signal a speed camera on the opposite direction, the user presses twice the button. This information is then sent to the server thanks to an included GPRS card, where a human operator verifies (previous messages of that user, comparison with other users, using another speed camera information service...) the plausibility of the information before broadcasting it to all users.

Despite it is very simple to use, Coyote remains an expensive system (699 € for 2 years with unlimited use and including communication fees) that not everybody can afford. And if there are too few users, then the chance that you are the first that discover a speed camera (by being flashed!) is high...

2.3.4 InfoRad

Autonomous and easy to use, InfoRad [26] beeps when the driver enters a "risky area". All the risky areas, materialized with a speed camera, are stored in the on-board database. It works thus only with fixed speed cameras and it is not possible to signal a new one to other drivers. It allows however a user to add its own risky areas for personal use. Their website provides time to time updates of risky areas. The device with an unlimited access to their database costs about 200 €.

3 Trusting Virtual Tags

Lots of work has already been done in the trust context (see 2 - "Related Work"), and the question that arises is why not just using well-known trust models and apply them to virtual tags? The answer is simply that it will not work. Indeed, traditional trust models are mainly designed with file sharing or auction applications in mind. In this case, people are rating each other and when user *A* wants to download a file (or buy an item) from user *B*, he questions the system in order to determine how trustworthy user *B* is. Currently, commercial systems (like eBay) are using very basic centralized systems, and the academics are suggesting solutions to transform such systems into peer-to-peer architectures.

But spatial messaging is noticeably different from file sharing or auctioning and needs therefore a new trust model. The key difference is that in spatial messaging it is difficult to increase its own trust without making a significant contribution. For instance, to post a new tag that will be confirmed by others (in order to create a trust link), a user will have to be physically there (to make the observation that deserves a tag). In a similar way, the user that deletes an outdated tag makes also a significant contribution. So, even if a user wants to increase his trust value in order to harm the system later, his former contribution will compensate his future bad behavior. And this is an interesting difference that will be used in this work in order to construct our trust engine. It is, at least to our knowledge, a novelty in the trust domain and can be considered as the key point of this work. In "traditional" trust systems, it is always possible to easily increase one's own trust value in order to subvert the system later. For instance, it is easy to sell honestly a few goods in eBay in order to increase one's trust value. It is also easy to provide a few good files in a file-sharing system and then use the resulting good reputation to send Trojan horses. But in spatial messaging, a user can increase his trust value only in return of a significant contribution. We will also see in 3.2 - "Updating trust values" how we can make it impossible for a user to switch regularly between good and bad behavior in order to keep a minimum trust value, and how to avoid that a user that behaved correctly for a long time and became malevolent afterwards uses its long-term good reputation to harm the system.

3.1 The Uncertainty of the Truth

In traditional computational trust, we usually agree over a set of axioms and hypothesis. For instance, the "truth" is a notion that is common to all. A corrupted file is seen as corrupted by everybody. In spatial messaging however, the truth is context dependent. The truth becomes a subjective and temporal notion. Something that is true for one user is not necessarily true for the others. Something that is true at a certain time is not necessarily true later. We call this new notion the "uncertainty of the truth". If user *A* posts a tag saying "Dangerous path", user *B* only knows that user *A* finds this path dangerous. But *A* is perhaps just a tourist and the path is in no way dangerous for user *B*, which can be a confirmed mountain guide. Or this path was maybe dangerous because of the snow, which melt away by the time.

To our view, trust is not only a tool that can be used to exclude malevolent users from a given system. Trust is also a way of creating relationships between users that behave in a similar way. Like in real life, each user has its own definition of what the truth is. The aim is therefore to create trust relationships between people that share the same definition.

3.2 Updating Trust Values

A traditional way to store and update a trust value consists in counting the number of positive outcomes P, the number of negative outcomes N, and to define the current trust value as $T = P / (P + N)$. It is a simple model that fits very well to file sharing applications where a good file is simply considered as a positive outcome and a corrupted file as a negative one. In spatial messaging however, defining a positive and a negative outcome is more complicated. And since we have to deal with what we called previously the "uncertainty of the truth", we need to define a model that is specific for spatial messaging.

A model that can be used in case people are honest is one that uses data mining techniques in order to determine how reliable a given tag is, in a given situation for a given person. Data mining consists in picking up relevant information in large data sets. A good definition can be found at [27]. Basically, when you rate a tag, you increase the trust links with all the people that reviewed it in the same way, and decrease the trust links with all the people that rated it differently. While requesting tags, data mining algorithms are then able to determine how "close" you are with each reviewer according to the situations where you previously interacted with these people, and take this into account to determine how pertinent this tag is to you.

This model is however challenged when malevolent users take part in the system. For instance, an attack would consist in rating automatically and positively all new tags so that the next reviewers increase the malevolent user's trust value. And then this user will use his high value to post "reliable" false tags. A solution to this consists in increasing only the trust value of the author of a tag, since posting randomly interesting tags (if they are not "interesting", nobody will rate them positively) is almost impossible.

In applications where it is possible to scan all the tags, and rate them automatically, it seems easy to cheat the system. It is difficult in some cases to differentiate a normal behavior from a malevolent one. For instance, if you see a tag warning about a specific danger and you do not see this danger, you do not know if the author is a spammer (and you need to decrease his trust value) or if the danger simply disappeared (and then you should not decrease his trust value). We need to determine how much a trust value must be decreased when we rate negatively a tag, so that an honest user is not too much penalized, but so that a spammer can be excluded from the system in a reasonable delay. It means that even if the system is generic, it needs a high comprehension of the application domain in order to determine what are the right rules and parameters. For instance a rule will define how much we must decrease the trust value of someone that doesn't vote like us and a parameter will define what the minimum trust value is.

Like in the human world, trust varies not in the same way when it increases than when it decreases. Trust takes time be built, but can be destroyed very fast. And this non-linear way of handling trust is certainly necessary to protect ourselves. If you lent 10 times 2 € to someone that always paid you back, you will probably stop to trust him before 10 times when he stops refunding you. The reason is even more accentuated in a digital world where people can act in an automatic way, thus very fast. If we use our former $P / (P + N)$ example, it is easy for a user to behave correctly (most probably in an automatic way) for a certain time, and then use its high trust value to subvert the system. A first idea consists in representing a trust value as a single value. A good behavior increases it, a bad behavior decreases it. But the maximal value is limited. It means that even if someone behaves very well for years, his trust value is not that high, and can quickly become negative in case of a big bad behavior, or a succession of a few bad behaviors. Another important point is that trust increases in a linear way but decreases exponentially. An exponential function varies very slowly at the beginning and then increases endlessly. Like in the human model, we accept to forgive seldom and small misbehaviors, but we break our trust relationships if you we face a big misbehavior or a succession a small misbehaviors.

4 A Generic Trust Engine for Virtual Tags

4.1. Overview

The main idea of our generic trust engine, called GenTE, is to remember only important or recent information, like it is done in human communities. The virtual tags (called vTags or simply tag) and the users keep a history of their last or important transactions.

To know whether a tag must be shown to the user, the trust engine checks the *n* last reviews done by trustworthy users. A user is trustworthy if his combined trust value, computed as a mix of the trustor's opinion (based on former direct interactions) and the opinions of the trustor's friends (who ask their own friends, and so on until a certain level), is above a certain threshold. A trustor calls "friend" every user with who he has a good trust relationship, or better said, each user with a local trust value higher than 0.

When a user rates a tag, he updates the trust values of the author and the former reviewers according to rules and parameters that depend on the application. In certain cases, a review can be done on both directions. For instance an author can update the trust value of every reviewer that gives a positive rating, since they seem to share the same opinion about the tag.

4.2 A vTag in GenTE

A vTag contains different information, like its position and its content, as well as a history. The history is a two-column table containing pairs of user ID and corresponding vote. An example is given in figure 1.

ID	Vote
8	0
3	1
4	1
2	1

Figure 1. The history of a vTag

The lines are ordered in an inverse chronological order, meaning that the last user that voted for this tag is user 8. The vote can be either a "1", if the user confirms the tag, or a "0" if the user denies it. So we see that users 2, 4 and 3 agreed with the content of the tag, but later user 8 disagreed with it. The reasons can be either because user 8 is a malevolent user that wants to delete the tag, or, more probably, that the tag is outdated and needs therefore to be removed. If a user that is already in the history votes again for this tag, then his line is moved at the top of the table and the corresponding vote is updated.

When a user requests tags in a given area, the trust engine checks the vote of the two last friends (remember that a friend is someone in which we have a local trust value higher than 0) and if at least one of them voted "1", the tag is sent to the user.

It means that even if someone denied the tag by mistake, the tag is still returned to people that are asking for it. This choice implies that we suppose that the price of a false positive (a tag that should not be sent is sent) if lower than the price of a false negative (a tag that should be sent is not sent), which seems to be the case in all the practical applications we thought about.

When the two last users denied the tag (they voted "0"), the tag gets a request-to-delete order. It means that the tag remains for the same amount of time than elapsed since its creation before being deleted by the trust engine. A tag that has been created a long time ago needs therefore more time to be deleted than a recent one. However, to avoid that an "old" tag needs too much time to be deleted we have a maximum delay. And, to avoid that malevolent users scan the network and deny the tags as soon as they appear, we added also a minimum delay. Since then, each new tag is at least present for a certain amount of time (the minimum delay), so even if malevolent users deny these tags, honest users will have time to confirm them (which will cancel the request-to-delete order) and by the same time decrease the trust value of the malevolent deniers.

4.3 A user in GenTE

A user is represented by an ID and a trust table. The trust table is a three-column table containing pairs of user ID and corresponding trust values. We differentiate the AT trust (author trust) which indicates how reliable a given user is to post or to confirm an existing tag and the DT trust (denier trust) which indicates how reliable a given user is to deny tags that are outdated or false. An example is given in figure 2. We see that this user has in his trust table two friends (users 3 and 7), one user he doesn't trust (user 8), and one user in who he has the same trust as for an unknown one (user 13).

ID	AT Trust	DT Trust
3	5	4
7	2	3
8	-3	-4
13	0	0

Figure 2. The trust table of a user

After modifying the trust value of a user, the corresponding line is placed on top of the list, so that there are sorted in an inverse chronological order. Each trust value is simply an integer in the range [t_{min}, t_{max}] so that $t_{min} < 0 < t_{max}$. GenTE allows specifying rules to describe how a trust value must be changed according to a given situation. A typical case is to have a linear way to increase a value (for instance adding *n* when you agree with a tag) and an exponential way to decrease a value (for instance multiplying by *m* a negative trust value). And if $-t_{min}$ is much bigger than t_{max} (for instance t_{min} =-50 and t_{max} =5), then we imitate the human way of handling trust [28]: Trust takes time to be built, we forgive some small misbehaviors (exponential functions moves slowly at the beginning), but when we loose trust in someone (one big disappointment or lots of small disappointments) then it becomes very difficult to rebuild a good trust relationship. We avoid that malevolent users switch between good behaviors (in order to increase their trust value) and bad behaviors (in order to subvert the system).

It is important that our system forgives small mistakes in cases where the truth is unknown. Imagine that a user sees a tag, but the tagged object does not exist anymore. He will disagree with the author of the tag as well as with all the people that agreed. He will therefore decrease their trust values since they are perhaps spammers. But, most likely, the object simply disappeared in the meantime and they are not spammers. Our model is built to forget easily such mistakes, as long as they do not happen too often, but to decrease quickly the trust values of malevolent users. The combined trust value of a user is relative and is computed by the following function:

combined_trust = q * myOpinion + (1-q) * friendsOpinions , q=[0..1]

It is a recursive function where *myOpinion* is the local trust value and *friendsOpinions* is the average opinion of the *n* first friends (where local trust > 0). These friends apply the same function, so they return a mix between their own opinion and the average opinion of their own friends. And so on until we reached the specified depth. This way of processing is fast (all the values are centralized) and gives a good idea of the global reputation of a user. Typically, if we choose *n*=10 (number of friends) and a depth level of 3, then we have already the opinion of $10^0 + 10^1 + 10^2 + 10^3 = 1111$ reliable people including ourselves, with more importance given to close friends. The higher is *q*, the more the user gives importance to his own value. In situations where people are susceptible of making mistakes, this value is usually quite small.

4.4 Trust updates

When a user votes for a tag, he puts his ID and his vote at the first line of the tag's history. This newly updated history is then analyzed by the trust engine, and the trust values of the users (that are in the history) are update according to their votes. For instance, if a user votes "1" and the two previous voters voted "0", the confirmer will decrease the trust value of the deniers. And perhaps increase the trust value of the author. The trust engine proposes a default behavior for each situation that can be adapted by the application developer in order to better meet the requirements of his application.

4.5 Rules

The rules that define the trustworthiness of a tag for a given user, as well as the rules that define how the trust values must be updated, are written by the application developer. To test our trust engine, we chose a speed camera warning system and wrote the following rules for a tag request:

History	Rules
Ø (empty)	if I trust the author, return true; return false;

1	if I trust the author, return true; if I trust the confirmer, return true; return false;
1-1	return true;
0	if I trust the author, return true; return false;
0-0	if I trust booth deniers, return false; if I trust the author, return true; return false;
1-0	if I trust the author, return true; if I trust the confirmer, return true; if I trust the denier, return false; return true;
0-1	if I trust the author, return true; if I trust the confirmer, return true; if I trust the denier, return false; return true;

We chose for this case that the size of the history is 2. We therefore keep, for each tag, the author ID as well as the two last votes. For instance, the notation 0-1 means that the last user denied the tag (he voted "0") and the last but one user confirmed it (he voted "1"). If we need to be more precise, we use also the notation 0(U_2)-1(U_1) meaning that user U_1 confirmed the tag, followed by user U_2 who denied it.

These rules decide whether a given tag must be returned to the requester. We execute the rules one by one until a condition make us to execute a "return true", in which case we return the tag, or a "return false", in which case we do not return the tag.

We then defined also how the trust values must be updated. The next two ables show the current history and shows how the trust tables of the author, the current user and the people in the history are updated according to the current vote ("1" or "0").

To show how we modify the trust values in each case, we define two functions. The first updates the AT trust value and is written like: UAT(U_1, U_2, *a*, *b*, *c*, *d*). It means that U_1 updates the local trust he has in U_2 as following: If the current trust of the trustee is equal or greater than 0, it multiplies the current trust by *a* and adds *b*, and if the trust of the trustee is negative, then it multiplies the current value by *c* and adds *d*. In a similar way, we define UDT(U_1, U_2, *a*, *b*, *c*, *d*) to update the DT trust. Finally we add also two functions, UAT(U_1, U_2, *a*, *b*, *c*, *d*, *C*) and UDT(U_1, U_2, *a*, *b*, *c*, *d*, *C*), where *C* is a specific condition that must be true in order to update the trust.

For instance, if the current user U_c votes 1 and the history is empty, then this user will increase the author's trust value if the condition *C* is met. In our case, *C* returns true only if there are at maximum *N* voters that already voted for this tag.

History	Rules if vote = $1(U_c)$
Ø (empty)	$UAT(U_c, U_a, 1, 5, 1, 5, C)$
$1(U_1)$	$UAT(U_c, U_a, 1, 5, 1, 5, C)$
$1(U_2)\text{-}1(U_1)$	$UAT(U_c, U_a, 1, 5, 1, 5, C)$
$0(U_1)$	$UAT(U_c, U_a, 1, 5, 1, 5, C)$ $UDT(U_c, U_1, 1, -1, 1.3, -1)$
$0(U_2)\text{-}0(U_1)$	$UAT(U_c, U_a, 1, 5, 1, 5, C)$ $UDT(U_c, U_1, 1, -3, 2, -3)$ $UDT(U_c, U_2, 1, -3, 2, -3)$
$1(U_2)\text{-}0(U_1)$	$UAT(U_c, U_a, 1, 5, 1, 5, C)$ $UDT(U_c, U_1, 1, -1, 1.3, -1)$
$0(U_2)\text{-}1(U_1)$	$UAT(U_c, U_a, 1, 5, 1, 5, C)$ $UDT(U_c, U_2, 1, -1, 1.3, -1)$

History	Rules if vote = $0(U_c)$
Ø (empty)	$UAT(U_c, U_a, 1, -1, 1.3, -1)$
$1(U_1)$	$UAT(U_c, U_a, 1, -1, 1.3, -1)$ $UAT(U_c, U_1, 1, -1, 1.3, -1)$
$1(U_2)\text{-}1(U_1)$	$UAT(U_c, U_a, 1, -1, 1.3, -1)$ $UAT(U_c, U_1, 1, -1, 1.3, -1)$ $UAT(U_c, U_2, 1, -1, 1.3, -1)$
$0(U_1)$	$UAT(U_c, U_a, 1, -1, 1.3, -1)$ $UDT(U_c, U_1, 1, 5, 1, 5)$ $UDT(U_1, U_c, 1, 5, 1, 5)$
$0(U_2)\text{-}0(U_1)$	$UAT(U_c, U_a, 1, -1, 1.3, -1)$
$1(U_2)\text{-}0(U_1)$	$UAT(U_c, U_a, 1, -1, 1.3, -1)$ $UAT(U_c, U_2, 1, -1, 1.3, -1)$
$0(U_2)\text{-}1(U_1)$	$UAT(U_c, U_a, 1, -1, 1.3, -1)$ $UAT(U_c, U_1, 1, -1, 1.3, -1)$ $UDT(U_c, U_2, 1, 5, 1, 5)$ $UDT(U_2, U_c, 1, 5, 1, 5)$

4.6 Additional rules

Rule 1: If you are in the first place of the history and you vote the same as previously, do nothing (no trust update and no modification of the history).

Without this rule a single user could delete a tag (by voting twice "0"). However, it is important to note here that this rule mentions explicitly that the two votes are the same. If you vote differently, the trust tables and the history are updated normally. We could thing that if someone votes differently, it was a mistake the first time and we can simply remove the former vote in the history and replace it by the new one. However, this behavior opens the door to a structured attack: The hacker finds a tag whose history is $0(U_2)$-$1(U_1)$, and then simply votes alternatively "0" and "1". He first votes "0", so he increases his DT trust with U_2. Then he votes 1, which would erase his last vote, and then he votes again 0, which will again increase his DT trust with U_2. And so on. In short, this would allow anyone to get the maximum DT trust value.

Rule 2: If you are in the first place in the history and voted "0", then the tag is not returned.

This rule avoids that users are disturbed by an object that disappeared. For instance, if a user tagged an object, then you need two different users to give a request-to-delete order to this tag. But if you are the only one that votes for this tag, you will never be able to delete it, and the tag will always be returned to you.

Rule 3: If an author denies his tag, and if the history is either empty or contains a single "0", then the tag is removed immediately.

If you post a tag and nobody sees it, or if you post a tag by mistake and want to remove it, this rule avoids keeping a wrong useless tag. We see also that if the only person that voted for this tag denied it ("0"), then it is a good idea to remove the tag immediately. However we do not remove the tag if the history equals 0-0. The reason is because a malevolent user can set up a structured attack in order to increase his AT trust: He authors a new tag, wait for a while so that people confirming the tag increase his trust value, and then with the help of a friend denies the tag (0-0) and then revokes it. Since the tag disappears, he can post a new one at the same place and again benefit from the trust increases given by the *N* first users that will confirm the new tag.

4.7 Validation process

We chose a speed camera tagging application to validate our trust engine. The first reason is because the topic is quite complex and interesting. Speed cameras can appear and disappear at any time, and it is not always possible to know if a false alarm is due to spammers or if it is actually the speed camera that just disappeared. The second reason is that it was very easy to find volunteers to test our system. We set up a simulator that allowed us to test different scenarios (spammers, users that try to delete all the tags...) as well as a widely deployed application used to confirm the results of the simulator. This application is FoxyTag [10], a worldwide free and collaborative system to signal speed cameras. The idea of FoxyTag consists in posting tags over speed cameras in order to warn the other drivers. Users are also motivated to confirm existing speed cameras; by doing so, they create trust links with the author and the other users that confirmed the camera, allowing them to get more reliable information in the future. More information about FoxyTag can be found on the website of the project [10].

5 Simulator

Our simulator randomly positions speed cameras on a road and simulates user's cars navigating according to given scenario parameters. An additional user, whose behavior can also be completely specified, logs its observations and returns the number of true positives (alarm: yes, camera: yes), false positives (alarm: yes, camera: no), true negatives (alarm: no, camera: no) and false negatives (alarm: no, camera: yes).

We model our road as a single way on a highway. Exits are numbered between 1 and n. Between two exits there is only one speed camera, numbered between 1 and n-1. So the camera c1 is between exits e1 and e2, the camera c2 is between exits e2 and e3, and so on. Figure 3 shows a road model.

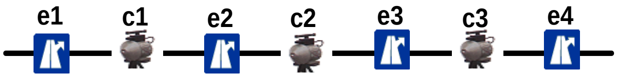


Figure 3. The road model

This model seems to be very simplistic. It is however sufficient to validate our trust metrics. Of course, we do not take into account some contextual information, like shadow areas (tunnels, urban canyons...) or what happens when the user posts a tag for the user driving in the opposite direction. These are more technical issues that need to be validated in the field and it is what we actually did with a real device in a real car. Since we can define the behavior of every user (where they enter and exit, how reliable they are by signaling speed cameras...) as well as the behavior of each speed camera (frequency of turning on, for how long...), we can precisely define which user drives in which area and how many speed cameras he is meant to cross on average. Our simulator accepts an input file that looks like this:

```
cam;1-4;8;15,10
cam;5-5;24;2,0
cam;5-5;240;3,30
usr;1-10;1-5;24;95;90
usr;1-1;3-5;240;80;75
usr;11-15;1-10;1;10;10
usr;11-11;1-10;0;20;25
col;5-7;1-11;6;10;100
spm;20-23;1-10;1
scn;100;2;run(24);pas(1,10);act(1,10,50,60)
```

In the first line, "cam;1-4;8;15,10" means that cameras 1 to 4 have one chance out of 8 to become active within an hour, and when one becomes active then it stays active for 15 minutes. After it stays inactive (paused) for at least 10 minutes. Note that these cameras will on average become active less than 3 times a day, since they cannot switch to active while there are already active or paused.

Precisely, these cameras will become active every 8+(15+10)/60 = 8.42 hours on average.

The next two lines define two different behaviors for camera 5.

In the fourth line, "usr;1-10;1-5;24;95;90" means that users 1 to 10 entry the highway at 1 and exits it at 5, that they run once a day and that they vote 95% of the time correctly when they signal the presence of a speed camera, and 90% of the time correctly when they cancel a camera.

In the collusion line, "col;5-7;1-11;6;10;100", we deduce that users 5 to 7 are colluding by entering all at the same time on entry 1, exiting on exit 11, and voting (all similarly) about all 6 hours with 10% of true positives and 100% of true negatives.

In the spam line, "spm;20-23;1-10;1", we deduce that users 20 to 23 spam by entering all at the same time on entry 1, exiting on exit 10, and voting 1 about every hour at every speed camera place.

The scenario, "scn;100;2;..." contains 100 big loops and 2 small loops. The scenario itself will be executed twice, then the trust engine is initialized, and then we re-execute the scenario twice. And so on (100 times).

run(t) means that the system will run for t hours (simulation time). Each minute, the go method of each camera and each user is called, allowing them to act according to their specified behaviors.

pas(e1, e2) means that our test user will passively drive once from exit e1 to exit e2. Passively means that he does not vote. His observations are logged and printed.

act(e1, e2, tp, tn) means that our test user will actively drive once from exit e1 to exit e2 and has tp (True Positive) chances (in %) to vote correctly if he sees a speed camera, and tn (True Negative) chances (in %) to vote correctly when he tries to cancel a speed camera that does not exist (anymore). His observations are logged and printed.

Everything after a // is a comment and is ignored by the parser.

6 Results

We compare here our GenTE trust engine with one called BasicTE, which simply adds a tag when a user posts such a request and remove it when a user denies it (there is in fact no trust engine). This permits to the reader to appreciate the efficiency of the GenTE trust engine. We tested it once with fixed speed cameras (Gen_F), and once with mobile speed cameras (Gen_M). The only difference is that in Gen_M the tags are automatically removed after 6 hours.

Scenario 1

```
cam;1-10;0;9999999;0
usr;1-100;1-11;24;100;100
usr;101-105;1-11;1;0;100
scn;100;100;run(24);act(1,11,100,100)
```

Scn 1	tp	fp	tn	Fn

Basic	43030	0	0	56970
Gen_F	99948	0	0	52
Gen_M	92022	0	0	7978

Scenario 1 tests our trust engine when malevolent users try to remove all the tags. We have 10 speed cameras that are always turned on (they are fixed speed cameras), a hundred users that behave always correctly and five users that systematically try to cancel all speed cameras they cross. Each hacker runs on average 24 times more often than an honest user. In the results table we compare the Basic and the GenTE trust engines. We used also the following abbreviations: "tp" means true positives (alarm: yes, camera: yes), "fp" means false positives (alarm: yes, camera: no), "tn" means true negatives (alarm: no, camera: no) and "fn" means false negatives (alarm: no, camera: yes).

With the BasicTE trust engine, we see that there are more false negatives (alarm: no, camera: yes) than true positives (alarm: yes, camera: yes). This is normal since the malevolent users are driving more than the honest ones. But our GenTE trust engine eliminates quite well these malevolent users, since less than 0.06% (52 / 99948) of the speed cameras where not tagged when we mentioned them as fixed ones (Gen_F).

```
Scenario 2
cam;1-10;9999999;0;0
usr;1-100;1-11;24;100;100
spm;101-105;1-11;1
scn;100;100;run(24);act(1,11,100,100)
```

Scn 2	tp	fp	tn	Fn
Basic	0	20820	79180	0
Gen_F	0	925	99075	0
Gen_M	0	840	99160	0

Scenario 2 tests how the trust engine reacts against a spam attack. This time the cameras are always turned off and the malevolent users vote "1" for each speed camera position. Again, we observe a significant improvement with our new trust engine.

```
Scenario 3
cam;1-10;48;360;720
usr;1-100;1-11;24;100;100
scn;100;100;run(24);act(1,11,100,100)
```

Scn 3	tp	fp	tn	Fn
Basic	8705	143	90767	385
Gen_F	8759	748	90146	347
Gen_M	8787	245	90619	349

In scenario 3 we have 10 speed cameras that are turned on every 66 hours (48 + (360 + 720) / 60) for 6 hours, and 100 users that vote always correctly. We have of course more false positives since we need two users to remove a tag (against only one in BasicTE). But if we tag the cameras as mobile ones (Gen_M), we observe an interesting improvement for the number of false positives.

```
Scenario 4
cam;1-10;48;360;720
usr;1-100;1-11;24;95;95
scn;100;100;run(24);act(1,11,95,95)
```

Scn 4	tp	fp	tn	Fn
Basic	8423	294	90472	811
Gen_F	8806	802	89990	402
Gen_M	8488	277	90856	379

In scenario 4 the users are voting incorrectly 5% of the time. This figure is clearly overrated (according to the tests realized with FoxyTag where this number is less than 1% in practice), but it let us to prove that our trust engine is tolerant with unintentional incorrect votes made by honest users.

```
Scenario 5
cam;1-10;48;360;720
usr;1-100;1-11;24;100;100
usr;101-105;1-11;1;0;100
scn;100;100;run(24);act(1,11,100,100)
```

Scn 5	tp	fp	tn	Fn
Basic	3845	76	90801	5278
Gen_F	8765	719	90102	414
Gen_M	8761	262	90591	386

In scenario 5 we added 5 deniers that try to remove all the tags they cross. The honest users are behaving correctly 100% of the time. We have clearly more false positives than for the BasicTE trust engine. This is normal since the deniers removed all the tags, whether there is a camera or not. If we compare the results with the ones from scenario 4 (for Gen_M), we see that our trust engine eliminates efficiently deniers.

```
Scenario 6
cam;1-10;48;360;720
usr;1-100;1-11;24;95;95
usr;101-105;1-11;1;0;100
scn;100;100;run(24);act(1,11,95,95)
```

Scn 6	tp	fp	tn	Fn
Basic	3612	60	91000	5328

Gen_F	8637	795	90109	459
Gen_M	8679	267	90604	450

In scenario 6 the users vote incorrectly 5% of the time. Unfortunately, we observe for Gen_M that the number of false negatives increases (compared to scenario 5). It seems that 5% of incorrect votes is a critical limit for this scenario.

```
Scenario 7
cam;1-10;48;360;720
usr;1-100;1-11;24;100;100
spm;101-105;1-11;1
scn;100;100;run(24);act(1,11,100,100)
```

Scn 7	tp	fp	tn	Fn
Basic	8781	17824	73124	271
Gen_F	8073	3073	87754	1100
Gen_M	8420	1345	89435	800

In scenario 7 we replaced the deniers by a spammer team, who votes "1" at every speed camera position. The other users are voting correctly 100% of the time. We observe quite bad numbers for GenTE. We first thought of a weakness in our trust engine, but further investigations concluded that it is actually the simulator that presents a weakness. The problem is that the positions of the cameras are always the same (which is not the case in reality), and that sometimes, by chance, a spammer really signal a new speed camera, which generously increases its trust value. In reality this would not be a problem, since signaling randomly a real speed camera at the right place is almost impossible.

```
Scenario 8
cam;1-10;48;360;720
usr;1-100;1-11;24;95;95
spm;101-105;1-11;1
scn;100;100;run(24);act(1,11,95,95)
```

Scn 8	tp	fp	tn	Fn
Basic	8595	18699	72115	591
Gen_F	7878	3471	87498	1153
Gen_M	8085	1403	89695	817

In scenario 8 the honest users are voting incorrectly 5% of the time. We face the same weakness as in scenario 7. We got therefore a bit worse results, since the honest users are less reliable.

7 Conclusion

This paper presented a generic trust engine to manage virtual tags. We saw that we couldn't simply use existing trust algorithms, since virtual tags have some particularities that need to be handled in a specific way. For instance we faced what we called the "uncertainty of the truth" problem, or how to rate a user if we cannot be sure if he is honest or not. We saw that this situation can happen in presence of an outdated tag. A user that sees a tag about an object or an event that is not present is either victim of a spam attack, in which case he should decrease the trust value of the tag's author, or he simply sees a tag that is outdated, in which case the author shouldn't be too much penalized.

We designed and implemented a trust engine called GenTE, which is able to exclude malevolent users but which is sufficiently tolerant with honest users, even if they do sometimes little mistakes. Since these mistakes are inevitable in spatial messaging (due to the uncertainty of the truth issue but also due to environmental ones, like a tag over a partially hidden object), GenTE is able to forgive small misbehaviors so that frequent users are not penalized.

We personalized GenTE through rules and parameters in order to adapt it for a speed cameras warning system called FoxyTag. We chose FoxyTag to test GenTE because the speed camera topic is quite complex (cameras can appear and disappear at any time, some are partially hidden...), and because it was easy to find volunteers to test our application. We got very promising results.

8 References

[1] Google Earth Community website, visited the 5th of May 2008: http://bbs.keyhole.com/

[2] Google website, visited the 5th of May 2008: http://www.google.com

[3] Tiny tiny blog, visited the 5th of May 2008: ttp://www.alvafilm.ch/blog/tinytiny/?cat=3

[4] Burrell, Jenna, Gay, Geri K. (2002): E-graffiti: evaluating real-world use of a context-aware system. In Interacting with Computers, 14 (4) p. 301-312

[5] Persson, P., Espinoza, F., Fagerberg, P., Sandin, A., and Cöster, R. GeoNotes: A Location-based Information System for Public Spaces, in Höök, Benyon, and Munro (eds.) Readings in Social Navigation of Information Space, Springer (2000)

[6] William G. Griswold, Patricia Shanahan, Steven W. Brown, Robert S. Boyer, Matt Ratto, R. Benjamin Shapiro, Tan Minh Truong: ActiveCampus: Experiments in Community-Oriented Ubiquitous Computing. IEEE Computer 37(10): 73-81 (2004)

[7] N. Mezzetti, "A Socially Inspired Reputation Model", in Proceedings of EuroPKI, 2004.

[8] R. Guha, "Open Rating Systems", 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web, 2004.

[9] S. Buchegger and J.-Y. Le Boudec, "A Robust Reputation System for P2P and Mobile Ad-hoc Networks", in Proceedings of the Second Workshop on the Economics of Peer-to-Peer Systems, 2004.

[10] FoxyTag website, visited the 5th of May: http://www.foxytag.com

[11] Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The Eigen-Trust Algorithm for Reputation Management in P2P Networks. 2003.

[12] Prashant Dewan. Peer-to-Peer Reputations. Proceedings of the 18th International Parallel and Distributed Processing Symposium (IPDPS'04) IEEE.

[13] V. Cahill, et al. Using Trust for Secure Collaboration in Uncertain Environments. IEEE Pervasive Computing Magazine, July-September 2003.

[14] Michael Kinateder, Kurt Rothermel. Architecture and Algorithms for a Distributed Reputation System. 2003.

[15] Aameek Singh, Ling Liu. TrustMe: Anonymous Management of Trust Relationships in Decentralized P2P Systems. Proceedings of the Third International Conference on Peer-to-Peer Computing (P2P'03). IEEE.

[16] Anwitaman Datta, Manfred Hauswirth, Karl Aberer. Beyond "web of trust": Enabling P2P E-commerce. Proceedings of the IEEE International Conference on E-Commerce (CEC'03).

[17] R. Guha, "Open Rating Systems", 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web, 2004.

[18] N. Mezzetti, "A Socially Inspired Reputation Model", in Proceedings of EuroPKI, 2004.

[19] S. Buchegger and J.-Y. Le Boudec, "A Robust Reputation System for P2P and Mobile Ad-hoc Networks", in Proceedings of the Second Workshop on the Economics of Peer-to-Peer Systems, 2004.

[20] D. Quercia, S. Hailes, and L. Capra, "B-trust: Bayesian Trust Framework for Pervasive Computing", in Proceedings of the 4th International Conference on Trust Management (iTrust), LNCS, Springer, 2006.

[21] Pierpaolo Dondio and Stephen Barrett and Stefan Weber and Jean Marc Seigneur, Extracting Trust from Domain Analysis, a Study Case on the Wikipedia Project 3rd International Conference on Autonomic and Trusted Computing (ATC 2006) LNCS 4158, Wuhan, China, 2006, L.T. Yang et al., 4158, Lecture Notes in Computer Science, pp. 362--373, sep, Springer-Verlag

[22] Radar buster website, visited the 5th of May: http://www.radarbusters.com/

[23] Mogoroad website, visited the 5th of May: http://www.mogoroad.ch

[24] Smart speed website, visited the 5th of May: http://www.smartspeed.fr/

[25] Coyote website, visited the 5th of May: http://www.moncoyote.com/

[26] Inforad website, visited the 5th of May: http://www.gpsinforad.co.uk/

[27] Data mining according to Wikipedia website, visited the 5th of May 2008: http://en.wikipedia.org/wiki/Data_mining

[28] Book: "Trust Rules: How to Tell the Good Guys from the Bad Guys in Work and Life (Hardcover)", by Linda K. Stroh, Praeger Publishers (August 30, 2007), 184 pages, ISBN: 978-0275998646

Mobile Location Based Services for Trusted Information in Disaster Management[1]

Lemonia Ragia, Michel Deriaz and Jean-Marc Seigneur

Abstract. The goal of the present paper is to provide location based services for disaster management. The application involves services related to the safety of the people due to an unexpected event. The current prototype is implemented for a specific issue of disaster management which is road traffic control. The users can ask requests on cell phones or via Internet to the system and get an answer in a display or in textual form. The data are in a central database and every user can input data via virtual tags. The system is based on spatial messages which can be sent from any user to any other in a certain distance. In this way all the users and not a separate source provide the necessary information for a dangerous situation. To avoid any contamination problems we use trust security to check the input to the system and a trust engine model to provide information with a considerable reliability.

1 Introduction

The wireless technology becomes important in our daily life because it provides a lot of services. The World Wide Web gives the opportunity to people to connect mobile phones or portable devises to Internet. Universal Mobile Telecommunication System (UMTS) with the new smart phones enable more services. The number of people that use the Web and the wireless technology is increasing rapidly.

User location was difficult to find out but with the usage of Global Positioning System (GPS) new possibilities are open. The integrated technology of GPS devices gives the location of the people quickly and with accuracy. That means that we can have location based services (LBS) which connect, in principle, the geographic location with user requests.

There are several approaches that show personalized LBS services for different applications: in the area of tourism [14], [1], [16] or navigation [10]. There are also some approaches for LBS which discuss the connection to databases [7], [6].

Disaster management is an important topic for local authorities, governments and disaster managers because they try to manage efficiently all the information provided mainly from people on the field to provide directions to the public. LBS for disaster management is extremely useful for the citizens since they can have great benefits having the right information in the appropriate time. In the scientific area of disaster management there are approaches which simulate a pre-disaster phase [9] or demonstrate an open source software especially for natural hazards [4]. Application for LBS for disaster management can be found on the area of

[1] A revised version of this work is published in the Proceedings of the IEEE International Conference on Information Systems Development

health care [13]. There are different aspects for services for disaster management. We can classify them in the following categories:

- Services for Natural hazards

This service provides information about the natural physical phenomena which can happen any time. Earthquake, flood, cyclones, fire etc., belong to this category. These information use *historical data* and try to make prediction for local authorities or other responsible offices to share the information and advise people how to avoid such a situation and protect themselves.

- Safety related services

In this category the information is related to the safety of the people in unexpected events. Man-made disasters such as car accidents or a plane crash are included. It provides information for a dangerous situation and it uses *real time data*. These *real time data* are related to this event and can be provided by any user.

In our approach we deal with the safety related services. An important issue in disaster management is for instance the traffic control. This service gives information about a safe and free travel and helps the users to avoid any kind of unexpected difficult occasion. It does not include the normal traffic jams during rush hours but it is related to unexpected events happening in special conditions. It takes into account a big area of infrastructure and it is updated by the users living through the event.

The mobile LBS application in our system is based on *spatial messages*. A *spatial message* is a message which refers to a specific geographic location. It allows a mobile user to publish a geo-referenced note so that any other user close and affected can get the message. Let us consider a community of car drivers. For example, an accident can happen or there is a fire next to the road. The car drivers would like to communicate about such event related dangers in specific places.

Spatial messaging has been already used. We could site for instance E-Graffiti [2]. E-Graffiti is a spatial messaging application that allows a user to read and post geo-localized notes. These notes can be either public or private, meaning that only the set of people defined by the author are able to read the note. E-Graffiti has been designed to study the social impact on spatial messaging.

Another interesting example is GeoNotes [12]. GeoNotes has more functionalities than E-Graffiti. While posting a note, the user can choose how he is going to sign it (for privacy reasons the user can write any text he wants as a signature), decide whether people are allowed to comment on it, and decide whether anyone can remove this message. For the readers, the graphical interface of the application provides some interesting functionalities like showing all the neighbouring messages or sort them according to different criteria. Inspired by the E-Graffiti evaluation, GeoNotes discarded the remote authoring of tags as well as the possibility to "direct" notes to certain users.

In our system the mobile location based services include the connection to a central database and in principle every user can send data to the central database using *virtual tags*. The *virtual tags* include any spatial messages which are related to a Geo referenced context related to disaster information. An important issue in our system is the use of a trust engine which gives information with considerable reliability to the users. We develop a framework that provides, among other things, a set of generic trust engines and a tool box providing geo-related tools.

This framework, called LBSDisMan (Location Based Services for Disaster Management), should provide APIs (Application Programming Interfaces) in order to ease future development of applications using *virtual tags*. The results can be presented in a cell phone or any other internet appliance.

In the next section we present the system architecture and give some details for the server and the client part. Section 3 gives an example of a real application and section 4 discusses results and further work to improve the system.

2 System Architecture

In order to share the spatial data among the users, we use a centralized architecture. The data is organized in small units that we call *virtual tags*. Each tag contains geo-related information, that means its position, and a content that is written in HTML.

The server part of our framework is represented in figure 1. The application designer starts by choosing the trust engine according to the kind of tag he is dealing with, then customize it with code (if needed) and parameters, and finally defines how the tags have to be stored (memory, flat files, database). For the storage, template classes should be provided in order to ease the development but still let the possibility for the developer to implement his own specifications.

The trust engines are generic and easily extensible. Each trust engine proposes a set of parameters in order to adapt its behaviour according to a given application, and all the trust computations are made in a standard and formalized way. This means that an application designer is able to adapt a trust engine by adding, modifying or removing the rules used to compute a trust value. Roughly speaking, the designer of a new application will have to code "how much a specific behaviour in a specific context costs in terms of trust value". He will therefore only have to code behaviours directly related to its application, leaving the framework doing all the job of maintaining and managing the trust information. This should guaranty that our trust engines can be adapted to any situations, and therefore really be generic.

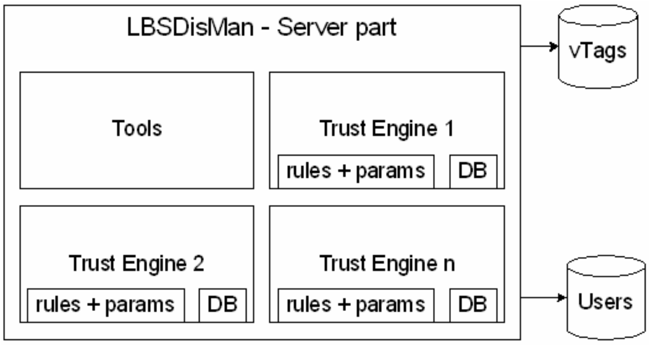


Figure 1**:** LBSDisMan server part of the framework

The Tools box is used by the trust engines and can also be accessed by the application. It contains mostly geographical related tools, like methods allowing conversions or methods handling tags of different formats.

All accesses to the database (vTags contains the *virtual tags* and Users contains the ID of the users) are done via the trust engines. It can of be any storage solution, including no permanent storage (information is kept in memory), flat files or a SQL standard database. Throughout this document, we will use the term "database" or its abbreviation "DB" to mention any storage system, and use the term "SQL database" or "SQL DB" if we talk about a "traditional" relational database using the SQL language to interact with.

Each trust engine provides a box allowing personalization it through rules and parameters, as well as a DB box responsible to store the tags in a permanent way. The latter should provide classes that can be adapted for the main storage architectures but also provide a generic solution that can be extended by a developer willing to implement its own storage architecture.

The trust engine should be accessed via three main primitives:

setTag. This primitive simply creates a new tag. No trust mechanism is used.

- **getTags**. Returns a list of tags. The requester specifies which filter he wants to apply to the result. For instance, a user can ask to get all the tags in a certain radius, with updated trust values for the author and the reviewers, and let the application decide what to do. But he can also ask to get only the tags that are above a certain trust level and ignore the others. Or he can apply a personal filter and not use the trust mechanism at all, like asking all the tags that are authored or reviewed by a user.

reviewTag. Reviewing a tag means to rate it, optionally to add a comment, and then update the trust tables of the reviewer, the author and the former reviewers. The way the trust tables are updated is defined through the rules and the parameters. The framework splits all the behaviours so that the application developer can simply write the rules according to the needs of its application.

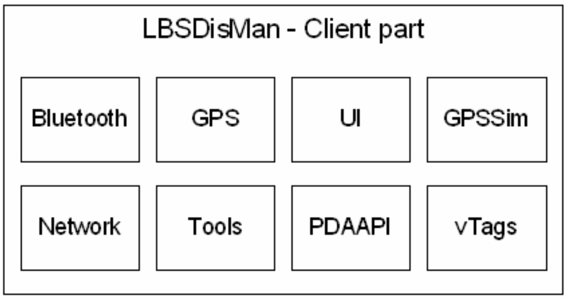


Figure 2: LBSDisMan client part of the framework

The LBSDisMan framework provides also an API for the client part. This API provides geo-related tools, tools to manage virtual tags, and also some general tools that will be needed by spatial messaging applications like sending information over the Internet from a mobile device, storing information on the local device, or accessing to an external or internal GPS (or another positioning device). A graphical representation of the client part is given in figure 2.

2.1 Security in the System

In a secured spatial messaging system, a user can be sure that the message he is reading is really written by the mentioned author, that nobody has modified the content of the original message, and that all other available messages at this place are available. More precisely, a secured spatial messaging system has to respect the "traditional" security services that are [3]:

- Confidentiality. Protection of the information against divulgations.
- Integrity: Protection of the information against modifications.
- Availability: Information is always available.
- Entity authentication: The author can be identified.
- Data origin authentication: Information can be linked to its author.
- Non-repudiation: The author cannot repudiate a message.
- Non-duplication: Protection against copying the information.
- Anonymity: The real-life identity of the users must be preserved.

Our aim is to focus on specific security services, the ones that are required for spatial messaging (in addition to the "traditional" ones). These are centered on the *pseudonym concept* [8]. What we would like is a system in which an author can be identified, but at the same time we would like to prevent any link with his real-life identity. A new user is therefore able to get a pseudonym in an anonymous way, but only one. If the person can obtain an unlimited number of pseudonyms, then the system can be victim of a Sybil attack [5]. The user must also be able to change its pseudonym. Again, this must be done in an anonymous manner and it must be impossible to link a former pseudonym with the new one.

A secured spatial massaging system must therefore respect, in addition to the "traditional" security services, the following "specific" ones:

- A user has only one pseudonym at a time.
- A user must be able to change its pseudonym.
- It is impossible to link a pseudonym to a real-life identity.
- It is impossible to link two pseudonyms of the same real-life identity (an old one with a new one).

Each pseudonym is unique, it is impossible that two different real-life identities share the same pseudonym. This is even true over time; if a user changes its pseudonym, the old one is locked and can never be used again. If in our application, we have a small community of users, we could choose to base the security of the database, its access and the users information via the use of a Public Key Infrastructure (PKI).

2.2 Trust in the System

The previous section discussed the security aspects of spatial messaging. A reader can be sure that a given message is really posted by its signer and that the content has not been modified since. But even if the reader can be sure about the author's identity, it is useless if they do not know each other. This section discusses how to add trust information on spatial messages so that the reader can evaluate the reliability of a message.

Trust is a very complex concept. Even if it is part of everyday life, different people give also different definitions of what trust is. This observation is even strongly accentuated when we try to explain how to build a trust relation between machines, or between humans and machines. One reason is that most models are only designed and specialized for peer-to-peer files sharing systems. For example, these models do not take time into account. In spatial messaging time is very important. For example a message indicating a high risk of avalanches posted yesterday has to be taken more seriously than the same message posted six months ago.

Spatial messaging needs a specific trust model that takes time into account, as discussed previously, and that is sufficiently flexible to be adapted to different situations. For example, in a mountain guide example, we suppose that the community of users is quite small and that a Web-Of-Trust trust model [14] will be sufficient. If user A trusts user B at 0.8 (out of 1), and user B trusts user C at 0.5, then user C rating (in user A 's eyes) will only count for 0.8 * 0.5 = 0.4. This does not mean that user A 's trust in user C is only 0.4. It is only the number by which user C 's rating will be multiplied.

However this model does not work for large communities. In this case we need to know the global reputation of the author. We could of course provide two different models depending on the size of the community. There is also a third trust model, the one that informs about the reliability of the message itself, without taking care of the author's reputation. Even a very reputable author can make a mistake and publish wrong information. Or, even more likely, a message signed by a reputable editor can contain outdated information.

We use a trust model which is actually the model that will combine the former ones. Its role is to answer the "How to trust the different trust models" question. The three previous models will give us three different trust values, and the fourth model's role is to determine, according the current situation, how much weight to give to each value. In this way we obtain a trust engine that is generic and can be easily applied to any situation.

3 Implemented Prototype

We have developed a system for Mobile Location Based Services for Disaster Management according to the architecture outlined in the previous section. We applied it in a specific topic of disaster management which is road traffic control. We used a central database including traffic data and additional data related to unexpected events and disaster phenomena. For the geometry we follow the standards of Open Geospatial Consortium [11] using their geometrical attributes.

The system incorporates spatial queries including requests regarding the content, the geographical position, the address and the time. The content describes all the information about this specific theme, e.g. “give all the fires in a distance of 50 km of the place with coordinates X and Y”. The geographical position is based on the longitude and latitude of a location which can be taken by a navigation system or

GPS, e.g. a spatial query can be “car stops at a position with X and Y coordinates, is there any problem in the highway”. An address refers to a street (name, number or code), to postal code, to a name of a city, e.g. “give all information about traffic jam in the highway number 5 in Switzerland” and the date can be a day, a month, week, year, part of day, hour, minute, e.g. “show all the accidents positions during April 2006”. Or when car driver is on a forest road can ask “is any fire in a specific part of a road”. The content will be also chosen by the service and there will be a lot of possibilities to make different kind of queries.

The system allows the user to enter data via virtual tags using a location. Then spatial messages can be sent to other users. In this case a user can ask for a user profile. The trust engine provides a trust value which is between [0,1]. The zero means totally unreliable, 0.5 neutral and one highest reliability. We developed an application running on mobile phones that helps the user to find the closest exit from the centre of a city (fig. 3). Something unexpected happened and the traffic stops for some time. A Bluetooth GPS connected to the mobile phone gives the current position of the user, and the GPRS protocol is used by the mobile in order to connect to the server that hosts the data. The user receives spatial messages in his cell phone “there are flames in a building” with high trust value.

In this example the user using a GPS system provides his/her coordinates to the system and ask the query “which are the next exits from this specific location in the centre of Geneva”. The green point with a circle shows the position of the user. Then the system shows the exit 1, 2, 3 with blue colour (fig. 3). Due to the security part of the system and after analysing the data the system gives only one solution to the user which is the number 3 in this case and highlighted in green on the mobile phone display. The visualization of the results can be displayed in a mobile electronic device like a cell phone (fig. 3).

The implementation of our system includes spatial queries in SOAP protocol and in XML language and answers can be shown via XML or SOAP. The system uses suitable methods for selecting, storing and detecting user profiles according to their location.

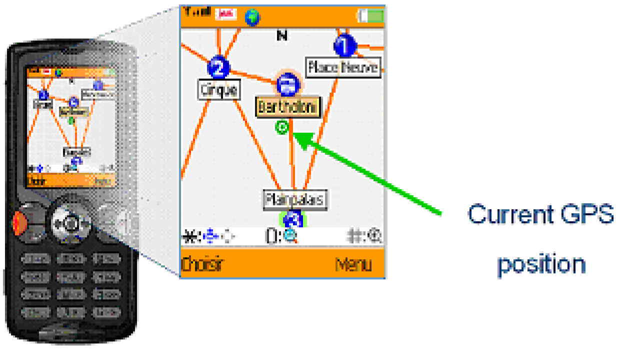


Figure 3: Visualization of the information

4 Conclusions and Future

We present a system for mobile location based services for disaster management and its application for traffic control. The system uses spatial messages to share geo referenced information to the users. It incorporates a central database and every user is allowed to feed data in the database. The users can use the services to ask queries at a given spatial location and receive the messages real time in a smart phone or other Internet device. Our system integrates trust engines and its security is taken also into account. In this way we improve the quality and reliability of the services. We implemented a disaster management scenario using real examples and we used the cell phone display to show the results of the spatial messages.

Currently we work on the implementation of the designed services to improve the results. We would like to use a bigger scenario with more real data. We envision a system applied in other applications of disaster management. From the database perspective we will work more in database integration and try to use the system with real data provided by other sources. In addition, we investigate the model of the trust engine as a general framework for open applications.

References

[1] Antikainen, H., Rusanen, J., Vartiainen, S., Myllyaho, M., Karvonen, J., Oivo, M., Similä, J. & Laine, K., 2006: Location-based Services as a Tool for Developing Tourism in Marginal Regions. Nordia Geographical Publications, 35: 2, pp. 39-50.

[2] Burrell, Jenna, Gay, Geri K. 2002: E-graffiti: evaluating real-world use of a context-aware system. In Interacting with Computers, 14 (4) pp. 301-312.

[3] [Charton E., 2005: *Hacker's Guide, Edition DeLuxe*. Campus Press.

[4] Currion P., Silva de C., and Walle Van De B., 2007: Open Source Software for Disaster Management. *Communications of the ACM*, Vol. 50, Issue 3, pp. 61-65.

[5] Douceur J.R., 2002: The sybil attack. In Proc. of the IPTPS02 Workshop, Cambridge, MA USA, March.

[6] Gruber B., Winter S., 2002: Location Based Services using a Database Federation. In: Ruiz, M.; Gould, M.; Ramon, J. (Eds.), 5th AGILE Conference. Universitat de les Illes Balears, Palma, Spain, pp. 243-252.

[7] Jensen C.J., Christiensen A.F., Pedersen T. B., Pfoser D., Saltenis S. and Tryfona N., 2001: Location based services – A Database Perspective. In J.T Bjorke and H. Tveite (Eds). Proc. Of 8th Scandinavian Research Conference on Geographical Information Science, pp. 59-68.

[8] Lubinski A., 1998: Security Issues in Mobile Databases Access. In Proceedings IFIP WG 11.3 12th International Conference on Database Security

[9] Meissner A., Luckenbach T., Risse T., Kirste T., and Kirchner H., 2002: Design Challenges for an Integrated Disaster Management Communication and Information System. In the IEEE DIREN `02, The First IEEE Workshop on Disaster Recovery Networks.

[10] Müller J., 2006. Location based services Indoor Navigation. Presentation. ifgi.uni-muenster.de/~muellerj/lbs06/vortraege/8-IndoorNavigation.ppt

[11] Open Geospatial Consortium, http://www.opengeospatial.org/

[12] Persson, P., Espinoza, F., Fagerberg, P., Sandin, A., and Cöster, R., 2000: GeoNotes: A Location-based Information System for Public Spaces, in Höök, Benyon, and Munro (eds.) Readings in Social Navigation of Information Space, Springer.

[13] Rahman A. A. and Zlatanova S., 2006: Pre-Hospital Location Based Services (LBS) for emergency management In: E. Fendel, M. Rumor (Eds.); Proceedings of UDMS'06 Aalborg, pp. 11.49-11.57

[14] Zimmerman P., 1994: PGP User's Guide, The MIT Press.

[15] Zipf A., Malaka R., 2001: Developing Location Based Services for Tourism the service providers. In: P. Sheldon, K. Wöber, D. Fesenmaier (Eds.), Information and Communication Technologies in Tourism, Proceedings of ENTER 2001, 8th International Conference. Montreal, Springer Computer Science, Wien, NewYork, pp. 83–92.

[16] Zipf A., 2002: User Adaptive Maps for Location Based Services (LBS) for Tourism. In Proc. Conference for Information and Communication Technologies in Travel & Tourism (ENTER). Springer-Verlag.

User-centric Mobile Identity Management Services[1]

Tewfiq El Maliki and Jean-Marc Seigneur

Abstract. Digital identity is the ground necessary to guarantee that the Internet infrastructure is strong enough to meet basic expectations such as security and privacy. Anywhere anytime mobile computing is becoming true. In this ambient intelligent world, the choice of the identity management mechanisms will have a large impact on social, cultural, business and political aspects: privacy is a human need and the all of society would suffer from the de-mise of privacy; people have hectic life and cannot spend their whole time administering their digital identities. The choice of identity mechanisms will change the social, cultural, business and political environment. Furthermore, the identity management is also a promising topic for modern society. Recent technological advance in user identity management has highlighted the paradigm of federated identity management and user-centric identity management as improved alternatives. The first one empowers the management of identity and the second the users to actively manage their identity information and profiles. It also allows providers to deal easily with privacy aspects regarding user expectations. This problem has been tackled with some trends and emerging solutions. Firstly, we provide an overview of identity management from identity 1.0 to identity 2.0 with emphasis on user centric approaches. Also we survey how have evolved the requirements for user-centric identity management and their associated technologies with emphasis on the federated approaches and user-centricity. Secondly, we will focus on related standards XRI and LID issued from Yadis project, and platforms mainly ID-WSF, OpenID, InfoCard, Sxip and Higgins. At the end, we treat the identity management in the field of mobility and focus on the future of mobile identity management.

Keywords: identity management, security, mobility, context-awareness, user-centricity

1 Introduction

Anytime, anywhere mobile computing is becoming easier, more attractive and even cost-effective: the mobile devices carried by the roaming users offer more and more computing power and functionalities including sensing and providing location-awareness [1]. A lot of computing devices are also deployed in the environments where the users evolve; for example, intelligent home appliances or RFID-enabled fabrics. In this ambient intelligent world, the choices of identity mechanisms will have a large impact on social, cultural, business and political aspects. Moreover, Internet of things will generate more complicated privacy problems [2]. Identity has become a burden on the online world. When it is stolen it engenders a massive fraud, principally in online services which generate a lack of confidence in doing business for providers and frustration for users.

[1] A revised version of this work is published as a book chapter of the "Computer and information security handbook" by Elsevier [to appear].

Therefore, the whole of society would suffer from the demise of privacy which is a real human need. As people have hectic live and cannot spend their time administering their digital identities, we need consistent identity management platforms and technologies enabling usability and scalability among others [3]. In this paper, we survey how the requirements have evolved for mobile user-centric identity management and their associated technologies.

The chapter is organized as follows. First, we present the evolution of identity management requirements. Section 3 surveys how the different most advanced identity management technologies fulfill present day requirements. Section 4 discusses how mobility can be achieved in the field of identity management in an ambient intelligent/ubiquitous computing world.

2 Evolution of Identity Management Requirements

In this section, we first define what we mean by a digital identity. In subsection 2.2, we summarize all the different requirements and detail the most important ones in the following subsections, namely, privacy, usability and mobility.

2.1 Digital Identity Definition

A digital identity is a representation of an entity in a specific context [18]. For a long time, a digital identity was considered as the equivalent of our real life identity which indicates some of our attributes:

- ✓ who we are, Name, Citizenship, Birthday;
- ✓ what we like, our favorite Reading, Food, Clothes, etc;
- ✓ what our reputation is, whether we are honest, without any problems, etc.

A digital identity was seen as an extended identity card or passport containing almost the same information.

However, recent work [4] has argued that the link between the real-world identity and a digital identity is not always mandatory. For example, on e-Bay what matters is to know whether the seller's digital identity reputation has been remarkable and that the seller can prove that she controls that digital identity. It is less important to know that her real-world national identity is from the Bermuda Islands, where suing anybody is rather unlikely to succeed. It should be underlined that in a major identity management initiative [5], a digital identity is defined as “the distinguishing character or personality of an individual. An identity consists of traits, attributes, and preferences upon which one may receive personalized services. Such services could exist online, on mobile devices at work, or in many other places”, that is, without mentioning a mandatory link to the real-world identity behind the digital identity.

The combination of virtual world with ubiquitous connectivity has changed the physical constraints to entirely new set of requirements as the associated security issues such phishing, spam, and identity theft has emerged. They are aggravated by the mobility of the user, the temporary and anonymity of cyber relationships. We are going toward new truly virtual world with always the implication of human. Therefore, we are facing the problem of determining the identity of our interlocutor and the accuracy of his/her claims. Simply using strong authentication will not resolve all these security issues.

Digital identity management is a key issue that will ensure not only the service and functionality expectations but also security and privacy.

2.2. Identity Management Overview

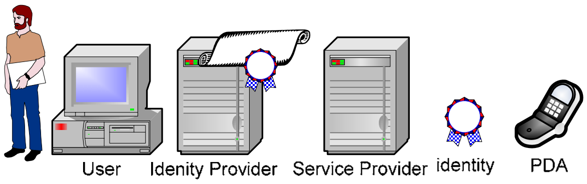


Fig. 1 : Identity legend

A model of identity can been as follows [6]:

- User who wants to access to a service
- Identity Provider (IdP): is the issuer of user identity
- Service Provider (SP): is the relay party imposing identity check
- Identity (Id) : is a set user's attributes
- Personal Authentication Device (PDA) : Device holding various identifiers and credentials and could be used for mobility

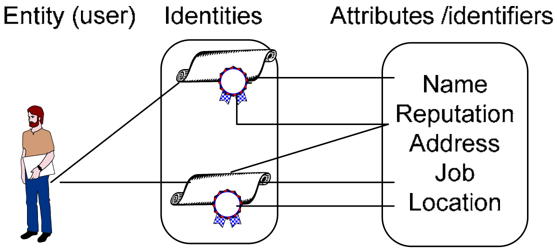


Fig. 2 : relationship between identities, identifiers and entity

The relationship between entities, identities and identifiers are shown in Fig.2 which illustrates that an entity, such as a user, may have multiple identities, and each identity may consist of multiple attributes that can be unique or non-unique identifiers.

Identity management refers to "the process of representing, using, maintaining, deprovisioning and authenticating entities as digital identities in computer networks".

Authentication is the process of verifying claims about holding specific identities. A failure at this stage will threaten the validity in the entire system. The technology is constantly finding stronger authentication using claims based on:

- Something you know: password, PIN
- Something you have: one-time-password
- Something you are: your voice, face, fingerprint (Biometrics)
- Your position
- Some combination of the four.

The BT report [3] has highlighted some interesting points to meet the challenges of identity theft and fraud:

- Developing risk calculation and assessment methods
- Monitoring user behavior to calculate risk
- Building trust and value with the user or consumer
- Engaging the cooperation of the user or consumer with transparency and without complexity or shifting the liability to consumer
- Taking a staged approach to authentication deployment and process challenges, using more advanced technologies

Digital identity should mange three connected vertexes: usability, cost and risk as illustrated in fig 3.

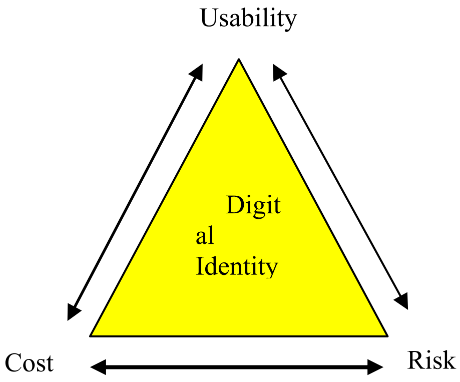


Fig. 3: Digital identity environment to manage

The user should be aware of the risk he/she facing if his/her device/software's security is compromised. The usability is the second aspect that should be guaranty to the user unless he/she will find the system difficult which could be a source of security problem. Indeed, a lot of users when they are flooded by passwords write them down and hide them in a secrete place under their keyboard. Furthermore, the difficulty to deploy and manage a large number of identities discourages the use of identity management system. The cost of a system should be well studied and balanced related to risk and usability. Many systems such as one-Time-Password token are not widely used because they are too costly for a widespread deployment for large institutions. Traditionally identity management was seen as service provider centric as it was designed to fulfill the requirements of service provider, such as cost effectiveness and scalability. The users were neglected in many aspects because they were forced to memorize difficult or too many passwords.

Identity management systems are elaborated to deal with the following core facets [7]:

- ✓ Reducing identity theft: The problem of identity theft is becoming a major one, mainly in the online environment. The providers need more efficient system to tackle this problem.
- ✓ Management: The amount of digital identities per person will increase, so the users need convenient support to manage these identities and the corresponding authentication.
- ✓ Reachability : The management of reachability allows user to handle their contacts to prevent misuse of their address (spam) or unsolicited phone calls
- ✓ Authenticity: Ensuring authenticity with authentication, integrity and non-repudiation mechanisms can prevent from identity theft.
- ✓ Anonymity and pseudonymity: providing anonymity prevent from tracking or identifying the users of a service.
- ✓ Organization personal data management: A quick method to create, modify a delete work accounts is needed, especially in big organizations.

Without improved usability of identity management [7], for example, weak passwords used by users on many Web sites, the number of successful attacks will remain high. To facilitate interacting with unknown entities, simple recognition rather than authentication of a real-world identity has been proposed, which usually involves manual enrollment steps in the real-world [4]. Usability is indeed enhanced, if there is no manual task needed. There might be a weaker level of security but that level may be sufficient for some actions, such as, logging to a mobile game platform. Single Sign-On (SSO) is the name given to the requirements of eliminating multiple password issues and dangerous password. When we use multiple user Id's and passwords just to use the emails systems and file servers at work, we feel the inconvenience that comes from having multiple identities. The second problem is the scattering of identity data which causes problems for the integration of IT systems. Moreover, it simplifies the end-user experience and enhances security via identity-based access technology.

Microsoft first largest identity management system was Passport Network. It was a very large and widespread Microsoft Internet service to be an identity provider for the MSN and Microsoft properties, and to be an identity provider for the Internet. However, with Passport, Microsoft was suspected by many persons of intending to have an absolute control over the identity information of Internet users and thus exploiting them for its own interests. Passport failed to become the Internet identity management tool. Since then, Microsoft has clearly understood that an identity management solution cannot succeed unless some basic rules are respected [8]. That's why Microsoft's Identity Architect, Kim Cameron, has stated the seven laws of identity. His motivation was purely practical in determining the prerequisites of successful identity management system. He formulated the essential principles to maintain privacy and security.

1. User control and consent over the handling of their data
2. Minimal disclosure of data, and for specified purpose.
3. Information should only be disclosed to people who have a justifiable need for it.
4. The system must provide identifiers for both bilateral relationships between parties, and for incoming unsolicited communications.
5. It must support diverse operators and technologies.
6. It must be perceived as highly reliable and predictable.
7. There must be a consistent user experience across multiple identity systems and using multiple technologies.

Most systems do not fulfill several of these tests particularly they are deficient in fine-tuning the access control over identity to minimize disclosure of data.

The formulated Cameron's principles are very clear but they are not enough explicit to compare finely identity management systems. That's why we will define explicitly the identity requirements.

2.3 Privacy Requirement

The privacy is a central issue due to the fact that the official authorities of almost all countries has a legal strict policies related to identity. It is often treated in the case of identity management because the management deals with personal information and data. Therefore, it is important to give a definition. Alan F. Westin defines privacy as "*the claim of individuals, groups and institutions to determine for themselves, when, how and to what extent information about them is communicated to others*" [2]. However, we will use Cooley's broader definition of privacy [24]: "the right to be let alone", because it also emphasizes the problems related to disturbing the user's attention, for example, by email spam.

2.4 User Centricity

The evolution of identity management system is toward the simplification of user experience and reinforcing authentication. It is well known that a poor usability implies the weakness of authentication. Mainly federated management has responded to some of these requirements by facilitating the use and the managing of identifiers and credentials in the boundary of a federated domain. Nevertheless, it is improbable that only one federated domain will subsist. Moreover, different levels of sensitivity and risks of different services will need different kinds of credentials. It is obvious that we should give users support and atomization of the identity management on the user's side.

A new paradigm must be introduced to solve the problems of usability, scalability and universal SSO. Therefore, a user-oriented paradigm has emerged which is called user centric identity management. The word user controlled management [25] is the first used to explain user-centric management model. Recent federated identity management systems keep strong end-user controls over how identity information is disseminated amongst members of the federation. This new paradigm gives the user full control over his/her identity by notifying him the information collected and by guarantying his/her consent for any type of manipulation over collected information. A user control and consent is also defined as the first law in Cameron's Laws of Identity [8]. A user-centric identity management system supports the user's control and considers user-centric architecture and usability aspects.

There is no uniform definition but this one is "User-centric identity management is understood to mean digital identity infrastructure where an individual end-user has substantially independent control over the dissemination and use of their identifier(s) and personally-identifiable information (PII)."[21]

We can also give this definition of user centricity: "In user-centric identity management the user has the full control over hi/hers identity and consistent user experience during all transaction when accessing his/her services"

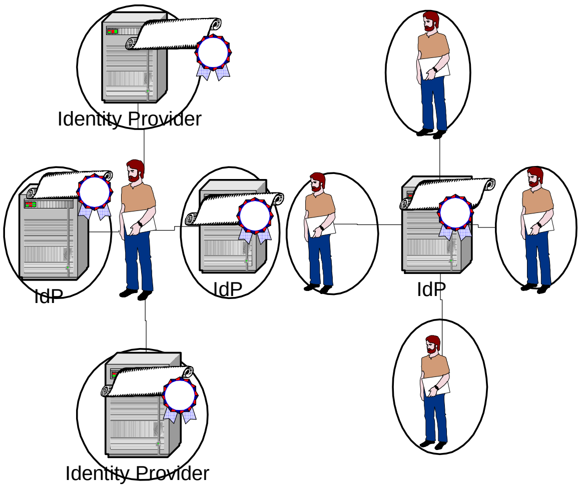


Fig. 4. IdP centric and User-centric models

In other terms it means that it allows the user to keep at least some or total control over his/her personal data.

One of the principles of user-centric identity is the idea that the user of a Web service should have full control over his/her identity information. A lot of technology discussion and solution has been focusing on service provider and rarely on user's perspectives. User-centric identity paradigm is a real evolution because it moves information technology architecture forward users with the following advantages:

1. Empowering the total control of users over their privacy;
2. Usability, as users are using the same identity for each identity transaction
3. Giving a consistent user's experience thanks to uniformity of identity interface
4. Limiting identity attacks i.e. Phishing

5. Limiting reachability/disturbances, such as spam
6. Reviewing policies on both sides when necessary, identity providers and service providers (Web sites);
7. Huge scalability advantages as the Identity Provider does not have to get any prior knowledge about the Service Provider
8. Assuring secure conditions when exchanging data
9. Decoupling digital identity from applications
10. Pluralism of Operators and Technologies

User-centricity approach allows user to gain access anonymously as he detains the full control on his/her identity. Of course, full anonymity [26] and unlinkability may lead to increased misuse by anonymous users. Then, Pseudonymity is alternative which is more suitable to the e-commerce environment. In this regard, anonymity must be guaranty at the application and at network levels. Some frameworks have been proposed to ensure user-centric anonymity using the concepts of One-task Authorization key and Binding Signature [26].

2.5 Usability Requirement

The security is also compromise with the proliferation of the user's password and even by their weakness. Indeed, some users note their passwords on scratch pads, because their memorization poses problem. The recent FFIEC guidance on authentication in online banking reports that "Account fraud and identity theft are frequently the result of single factor (e.g., Id/password) authentication exploitation" [27]. From then on, the security must be user oriented as he/her is the effective person concerned with it and a lot of recent attacks take advantage from the lack of awareness of users attacks (i.e. spoofing, pharming and phishing) [28]. Without strong control and improved usability [29] of identity management some attacks will be always possible. To facilitate interacting with unknown entities, simple recognition rather than authentication of a real-world identity, which usually involves manual enrollment steps in the real-world, has been proposed [4]. Usability is indeed enhanced if there is no manual task needed. There might be a weaker level of security reached but that level may be sufficient for some actions, such as, logging to a mobile game platform.

Single Sign-On (SSO) is the name given to the requirements of eliminating multiple password issues and dangerous password. When we use multiple user Id's and passwords just to use the emails systems and file servers at work, we feel the pain that comes from having multiple identities. The second problem is the scattering of identity data which causes problem for the integration of IT systems. Moreover, it simplifies the end-user experience and enhances security via identity-based access technology.

Therefore, we offer these features:

- flexible authentication,
- directory independence
- session and password management
- seamless

3. The Requirements Fulfilled by Current Identity Management Technologies

This section provides an overview of identity management solutions from identity 1.0 to identity 2.0 and how they address the requirements introduced in Section 2. We will focus on related standards XRI and LID issued from Yadis project and platforms mainly ID-WSF, OpenID, Higgins, InfoCard and Sxip. At the end, we treat the identity management in the field of mobility.

3.1 Evolution of Identity Management

This section provides an overview of almost all identity management 1.0. First of all, we describe the silo model, then different kind of centralized model and the federated identity management.

3.1.1 Identity Management 1.0

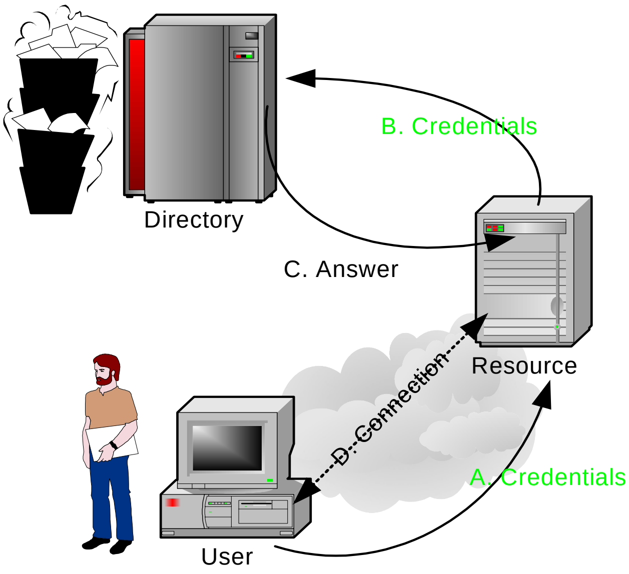


Fig. 5. Identity 1.0 principle

In the real world I use my identity card to prove who I am. How about the online world?

The first digital identity appeared when the user was associated with the pair (username, password) or any other shared secret. This method is used for authentication when connecting to an account or a directory. It proves your identity if you follow the guidelines strictly otherwise there is no proof. In fact, it is a single authority using opaque trust decision without any credentials (cryptographic proofs) choice or portability. In the context of Web access, the user must enroll for every non-related service, generally with different user interfaces and follows diverse policies and protocols. Thus, the user has a non-consistent experience and deals with different identity copies. In addition, some problems related to privacy have also emerged. Indeed, our privacy was potentially invaded by sites. It is clear that sites have a privacy policy, but there is no control from the user on his/her identity. What are the conditions for using these data? How can we improve our privacy? And to what granularity we allow them to use it? The same problem is revealed when having access to resources. The more resources, the more management we have. It is an asymmetric trust. And the policy decision maybe opaque. It allows access with an opaque trust decision and a single centralized authority without a credentials choice. It is a silo model [9] because it is neither portable nor scalable. This is Identity 1.0.

The identity management appeared with these problems in the 1980s. The fist identity management system was the Rec. X.500, developed by ITU [10], covering directory services like Directory Access Protocol (DAP). ISO was also associated to the development of the standard. Like a lot of ITU standards, this one was very heavy and complex. A light version appeared in the 1990s for DAP. It was LDAP which was standardized by the IETF and widespread and adopted by Netscape. Microsoft has invented an equivalent Active Directory, and for users, they introduced Passport. It is also the ITU which standardized X.509 for identities related to certificates. It is the format currently recognized. It is a small file, generated by an authority of certification.

If there is a loss or a usurpation of the certificate, it can always be revoked by the authority of certification.

This is for single user and what about business corporations who have automated their procedures and have a proliferation of applications with de-provisionning but still been in a domain-centric model. What about resources shared between domains?

3.1.2 Silo model

The main identity management system deployed in currently in the world of Internet is called silo model. Indeed, the identity provider and service provider are mixed up and they share the same space. The identity management environment is put in place and operated by a single entity for a fixed users' community.

Users of different services must have different accounts and therefore reenter the same information about their identity which increases the difficulty of management. Moreover, the users are overloaded by identity and password to memorize which produces a significant barrier to usage.

A real problem is the forgetfulness of passwords due to the infrequent use of some of these data. This can obviously lead to a higher cost of service provisions.

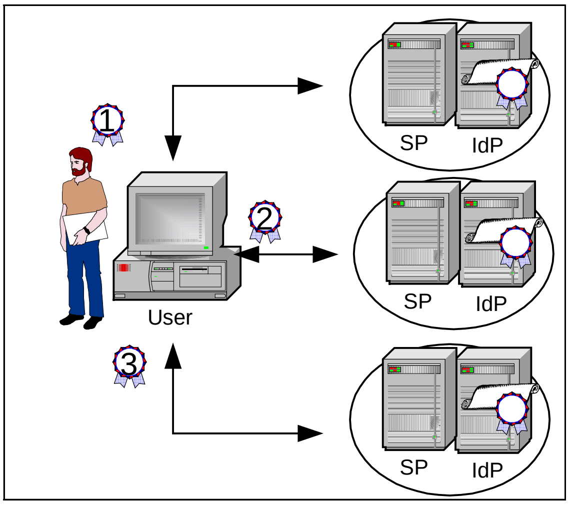


Fig. 6. Identity silo model

This is for single users, what about Enterprises that have automated their procedures and have a proliferation of applications with de-provisioning but are still in a domain-centric model? What about resources shared between domains?

Silo model is not interoperable and is deficient in many aspects. That's why federated identity management model is now emerging and it is very appreciated by enterprises. A federated identity management system consists of software components and protocols that handle in a decentralized manner the identity of individuals throughout their identity life cycle. [11]

3.1.3 Solution by Aggregation

Aggregating identity information and finding the relationship between identity records is important to aggregate identity. There are some alternatives:

- ➢ The first approach consolidates authentication and attributes in only one site and is called a centralized management solution like Microsoft Passport. This solution avoids the redundancies and inconsistencies in the silo model and gives the user a seamless experience [6]. The evolution was as follows [9, 11]:
  - ✓ Building a single central identity data store which is feasible only for small organizations.
  - ✓ Creating a meta-directory that synchronizes data from other identity data stored elsewhere
  - ✓ Creating a virtual directory that provides a single integrated view of the identity data stored
  - ✓ A Single Sign On Identity model which allows users to be authenticated by one service provider
- ➢ The second approach decentralizes the responsibility of IdP to multiple such IdPs which can be selected by the end users. This is a federate system where some attributes of identity are stored in distributed IdPs. A federated directories model, by linking identity data stored together, has emerged. Protocols are

defined in several standards such as in Shibboleth [19], Web services federation language 2003.

3.1.4 Centralized vs. Federation Identity Management.

Microsoft Passport is a centralized system, entirely controlled by Microsoft and closely tied to other Microsoft products. Individuals and companies have proven to be reluctant adopters of a system so tightly controlled by one dominant company.

Centrally managed repositories in centralized identity infrastructures can't solve the problem of cross-organizational authentication and authorization. This approach has several drawbacks as the IdP does not only become a single point of failure but may also not be trusted. That's why Microsoft Passport was not successful. In contrast, the federation identity will leave the identity resources in their various distributed locations but produce a federation that links them to solve identity duplication, provision and management.

3.1.5 A Simple Centralized Model

A relatively simple centralized identity management model is to build a platform that centralizes identities. A separate entity acts as an exclusive user credentials provider for all service providers. This approach merges both authentication and attributes in only one site. This architecture, which could be called Common user identity management model, is illustrated in Fig.7. All identities for each SP are gathered to a unique identity management site (IdP). SPs have to provide each identity to IdP.

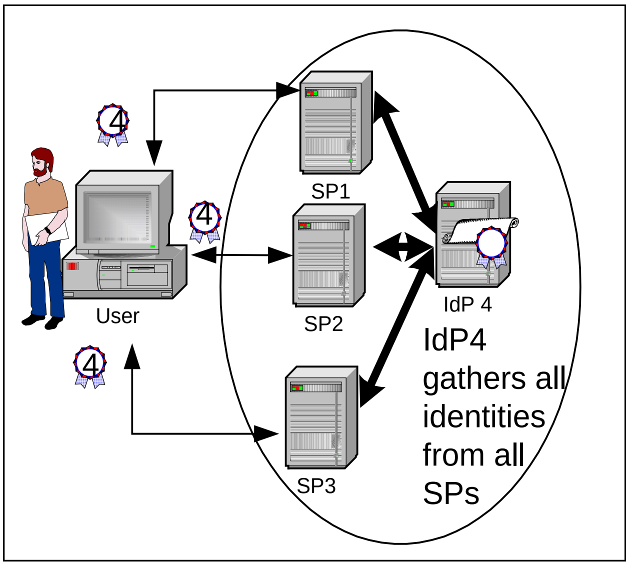


Fig. 7. Simple centralized identity management

In this environment, users can have access to all service providers using the same set of identifiers and credentials. A centralized certificated CAs could be implemented with a PKI or SPKI [23]. This architecture is very efficient in a close domain where users could be identified by a controlled email address. Although such architecture seems to be scalable, the concentration of privacy related information has a lot of difficulties in social acceptance in terms of privacy [32].

3.1.6 Meta-directories

SPs can share certain identity-related data on a meta-level. This can be implemented by consolidating all service providers' specific identities to a meta-identifier linked to credentials.

There are collections of directories information from various directory sources. We aggregated them to provide a single view of data. Therefore, we can show these advantages:

- A single point of reference provides an abstraction boundary between application and the actual implementation. A single point of administration avoids the multiple directories, too.
- Redundant directory information can be eliminated, reducing the administration tasks.

This approach can be seeing from the user's point of view to his/her password as synchronization across multiple service providers. Thus, the password is automatically changed with all the others.

This architecture can be used in large enterprises where all services are linked to a meta-directory. In this case, the ease-of-use is clear as the administration is done by a single authority.

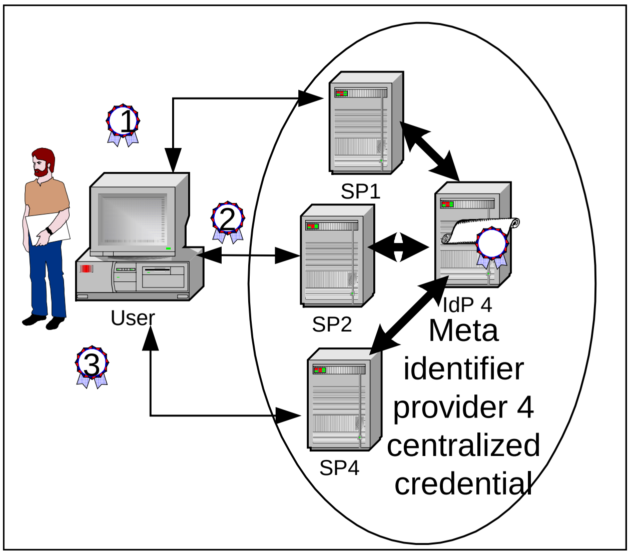


Fig. 8. Meta-directory model

3.1.7 Virtual Directories

Virtual directories are directories that are not located in the same physical structure as the Web home directory, but look as if they were to Web clients. The actual directories may be at a completely different location in the physical directory structure; for example, on another hard disk or on a remote computer. They are similar in concept to meta-directories in that they provide a single directory view from multiple independent directories. They differ in the means used to accomplish this goal. MD software agents replicate and synchronize data from various directories in what might be batch processes. In contrast, VD provide a single view of multiple directories using real-time queries based on mapping from fields in the virtual scheme to fields in the physical schemes of the real directories.

3.1.8 Single-Sign-On (SSO)

We use multiple user's Ids and passwords just to use the emails systems and file servers at work and we feel pain from managing multiple identities. The second problem is the scattering of identity data which causes problem for the integration of IT systems.

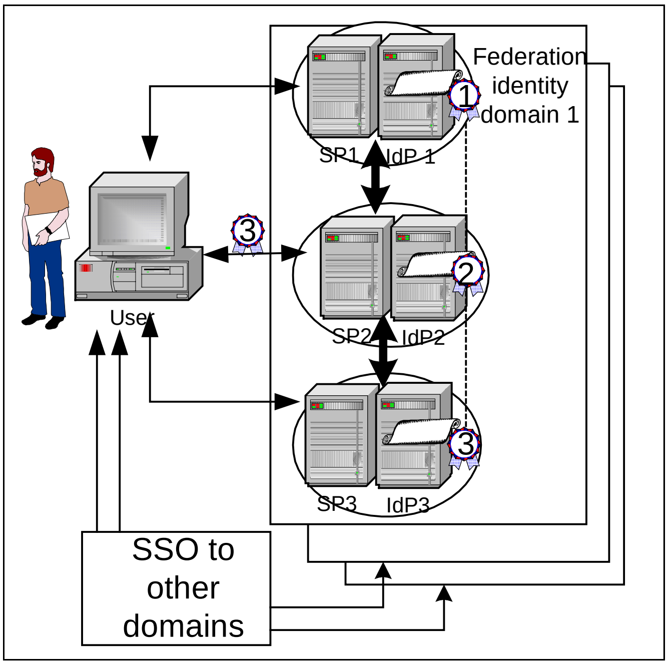


Fig. 9. Single-Sign-On model

Single Sign-On is a solution proposed to eliminate multiple password issues and dangerous password. Moreover, it simplifies the end-user experience and enhances security via identity-based access technology.

Therefore, it offers these features:

- flexible authentication,
- seamless
- directory independence and
- session and password management.

3.1.9 Federated Identity Management

We have seen different approaches to manage user's identity; they are not clearly interoperable and are deficient in unifying standard-based frameworks. On one hand, maintenance of privacy and identity control are fundamental when offering identity to users, on the other hand the same users ask for more easy to use and rapid access. The balance of the two sides leads to federated network identity. That's why these environments are now emerging. A federated identity management system consists of software components and protocols that handle the identity of individuals throughout their identity life cycle.

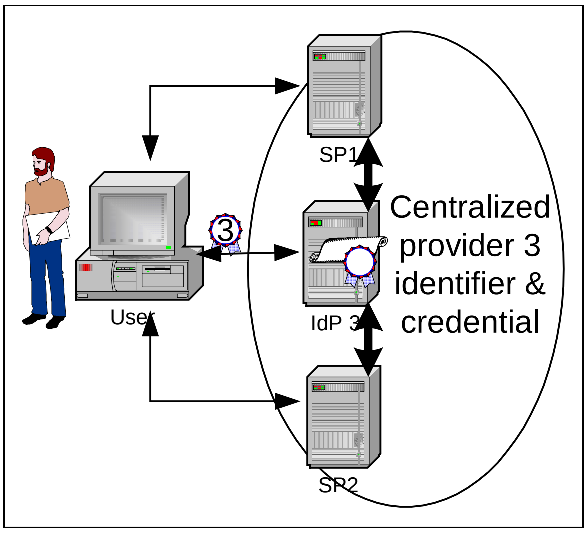


Fig. 10. Federated identity management model

This architecture gives the user the illusion that there is a single identifier authority. Even though the user has many identifiers, he doesn't need to know exactly all of them. Only one identifier is enough to have access to all services in the federated domain.

Each SP is responsible for the name space of his users and all SPs are federated by linking the identity domains. Thus, the Federated identity model is based on a set of SPs called a circle of trust by the Liberty Alliance. This set of SPs follows an agreement on mutual security and authentication in order to allow SSO. Indeed, the federated identity management combines SSO and authorization tools using a number of mutual SPs' technologies and standards. This practice makes the recognition and entitlement of user identities by other SPs easy. The Fig.10 shows the set of federated domains and the possibility for other SPs to have access to the same user with different identifiers.

The essential difference between federated identity systems and centralized identity management is that there is no single entity that operates the identity management system. Federated systems support multiple identity providers and a distributed and partitioned store for identity information. Therefore, a federated identity network allows a simplified sign-on to users by giving rapid access to resources, but it doesn't require the user's personal information to be stored centrally. With this identity network approach, users authenticate themselves once and can control how their personal information and preferences are used by the service providers.

Federated identity standards, like those produced by the Liberty Alliance, provide Single-Sign-On over all offered services and enable users to manage the sharing of their personal information through identity and service providers as well as the use of personalized services in order to give access to convergent services. The interoperability between disparate security systems is assumed by an encapsulation layer through a trust domain which links a set of trusted service providers.

However there are some disadvantages with federated identity management. The first one is the lack of privacy of the user as his/her personnel attributes and information can be mapped using correlation between identifiers. Anonymity could be violated. The second one is the scalability of users as they have access to the network from different domains by authentication to their relative IdPs. Therefore, the problem of passwords will continue across multiple federated domains. A major challenge is to integrate all these components into a distributed network and to deal with these drawbacks. This challenge cannot be taken up without new paradigms and supported standards.

The evolution of identity management system is toward also simplification of user experience and reinforcing authentication. It is very known that a poor usability implies the weakness of authentication. A new paradigm should be introduced to solve those problems while still being compatible at least with federated identity management.

That is why user-centric identity management, has emerged [6,9]. This paradigm is embraced by multiple industry products and initiative such as Microsoft Cardspace [12], Sxip [13] and Higgins Trust Framework [14]. This is Identity 2.0.

3.2 Identity 2.0

The user of Internet services is overcome with identities. he/she is seldom able to transfer his/her identity from one site to another. The reputation that he/she gains in one network is useful to transfer to other networks. Nevertheless, he/she can not profit from his/her constructed reputation and he/she should rebuild his/her identity and reputation another time, and so on. The actual systems don't allow users to decide about the sharing of their attributes related to their identity with other users. This causes a lack of privacy control. Some solutions propose an advanced social system that would model the social interaction like the real world.

The solutions must be easy to use and enable users to share the credentials among many services and must be transparent from the end-user perspective. The principle of modern identity is to separate the acquisition process from the presentation process. It is the same for the identification process and authorization process. Moreover, it provides scalability and privacy. Doing so, we can have more control on my identity.

The scale, security and usability advantages of user-centric identity are what make it the underpinning for Identity 2.0. The main objective of Identity 2.0 protocol is to provide users with full control over their virtual identities. An important aspect of Identity 2.0 is protection against increasingly Web attacks like Phishing attacks as well as the inadvertent disclosure of confidential information while enabling convenient management.

Identity 2.0 would allow users to use one identity respecting transparency and flexibility. It is focused around the user and not around directory or identity provider. It requires identified transactions between users and relaying party using credentials, thus providing more traceable transactions. To maximize the privacy of users, some credentials could be given to the users in advance. Doing so, the IdP could not easily know when the user is utilizing the credentials.

The Identity 2.0 endorses completely the paradigms of user-centric identity management enabling the full control of user on his/her identity. Service Provider will therefore be required to change their approaches by including request and authentication of users' identity. Identity 2.0 systems are interested in using the concept of a user's identity as credentials about the user, from their attributes like their name, address, to less traditional things like their desires, customer service history, and other attributes that are usually not so much associated with a user identity.

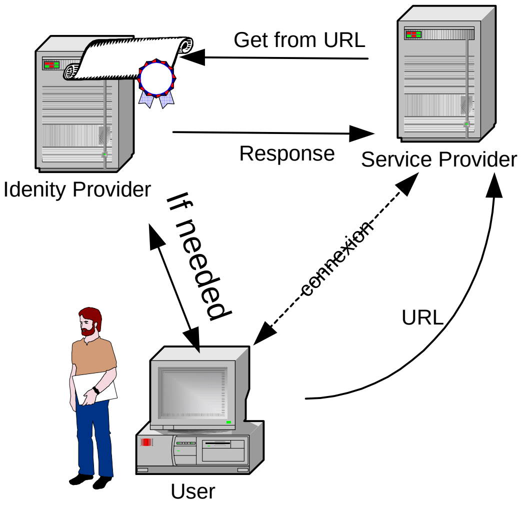


Fig. 11. URL-based Id 2.0

3.2.1 Identity 2.0 initiatives

When a Website collects data from users he cannot confirm whether or not the collected data is pertinent and reliable as the users often enter nonsense information into online forms. This is due to the lack of Website to control and verify the users' data. Furthermore, due to the law limitation on the requested data, the Website cannot provide true customized services even though users require them. In the other side, users have not direct control on what the Website will do with their data. In addition, the users enter many times the same data when accessing the first time different Websites. Doing so, they have a huge difficulty to manage their large number of identities.

To mitigate these problems, different models of identity management have been considered. One such model, Identity 2.0, proposes an Internet-scalable and user-centric identity architecture that mimics real world interactions.

Many research labs have collaborated to develop the Identity 2.0 Internet based Identity Management services. It is based on the concept of user-centric identity management, supporting enhanced identity verification and privacy, and user consent and control over any access to personal information for Internet-based transactions.

There are various Identity 2.0 initiatives:

a. LID
b. XRI
c. SAML
d. Shibboleth
e. ID-WSF
f. OpenID
g. Microsoft's CardSpace (formerly InfoCard)
h. SXIP,
i. Higgins

a. LID

Like LDAP, LID is under the principle of simplicity because many existing identity schemes are too complicated to be largely adoptable. It simplifies more complex protocol; but instead of being less capable due to fewer features, it has run success that their more complex predecessors lacked. This was because their simplification reduced the required complexity to the point where many people could easily support them, and that was one of the goals of LID.

Light-Weight Identity (LID) is a set of protocols capable of representing and using digital identities on the Internet in a simple manner, without relying on any central authority. LID is the original URL-based identity protocol, and part of the OpenID movement.

LID supports digital identities for humans, human organizations and non-humans (e.g. software agents, things, Websites, etc.) It implements Yadis, a meta-data discovery service and is pluggable on all levels.

b. XRI/XDI

We have XRI E**X**tensible **R**esource **I**dentifier and XDI which fractional solution without Web services integrated**. They are open standards** as they are royalty-free open standards. XRI is about Addressing. XDI is about Data Sharing protocol and uses basically XRI. Both XRI and XDI are being developed under the support of OASIS. I-name and I-number registry services for privacy-protected digital addressing use XRI. It can be used as an identifier for persons, machines and agents.

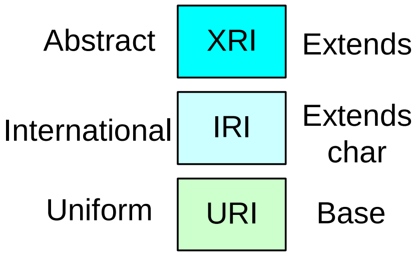


Fig. 12. XRI layers

XRIs offer a human-friendly form of persistent identifier. That's why it is convenient identifier for SSO system. They Supports both persistent and reassignable identifiers in the same syntax and establish a global context symbols. Moreover, they enable identification of the same logical resource across multiple contexts and multiple versions of the same logical resource.

XDI (XRI Data Exchange) is a Secure Distributed Data Sharing Protocol. It is also an architecture and specification for privacy-controlled data exchange where all data is identified using XRIs. The XDI platform includes explicit specification for caching with both push and pull synchronization. XDI universal schema can represent any complex data and have the ability of cross context addressing and linking.

c. SAML

The Security Assertion Markup Language (SAML) is an OASIS specification [40] that provides a set of rules for the structure of identity assertions, protocols to move assertions, bindings of protocols for typical message transport mechanisms, and profiles. Indeed, SAML is a set of XML and SOAP-based services and formats for the exchange of authentication and authorization information between security systems. The initial versions of SAML v1.0 and v1.1 define protocols for SSO, delegated administration, and policy management. The most recent version is SAML 2.0. It is now a common language to the majority platform to change secure unified assertion. He is very useful and simple as it is based on XML.

An assertion is a datum produced by a SAML authority referring to authentication, attribute information, or authorizations applying to the user with respect to a specified resource.

This protocol enables interoperability between security systems (e.g. Browser Single Sign On, Web Services Security, etc.). Other aspects of federated identity management as permission-based attribute sharing are also supported.

This protocol enables interoperability between security systems (e.g. Browser Single Sign On, Web Services Security, etc.). Other aspects of federated identity management as permission based attribute sharing are also supported.

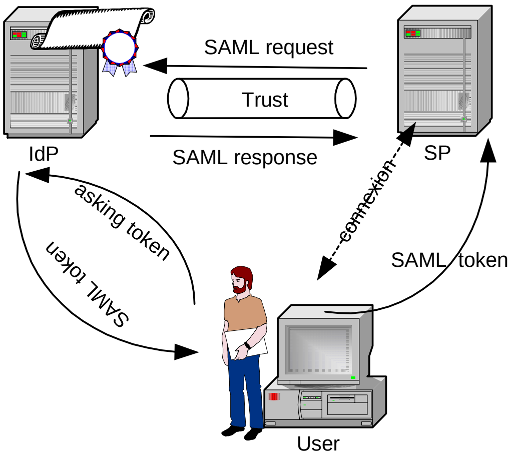


Fig. 13. SAML token exchange

SAML is sometimes criticized for its complexity of the specifications and the relative constraint of its security rules. Recently, the SAML community has shown significant interest in extending SAML to reach less stringent requirements for low-sensitivity use cases. The advantages of SAML are robustness of its security and privacy model and the guarantee of its interoperability between multiple vendor implementations through the Liberty Alliance's Conformance Program.

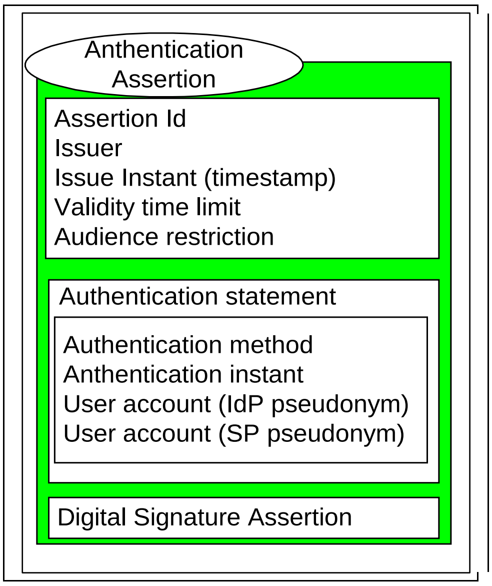


Fig. 14. SAML assertion

d. Shibboleth

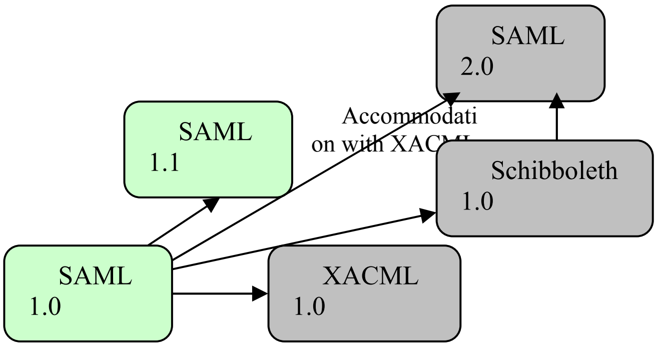


Fig. 15. Convergence between SAML & Shibboleth

Shibboleth [19] is a project which goal is to allow universities to share the Web resources subject to control access. Thereafter, it allows inter-operation between institutions using it. It develops architectures, policy structure, practical technologies, and an open source implementation. It is building components for both the identity providers and the reliant parties. The key concept includes “federated” management identity whose meaning is almost the same as the Liberty term’s [33]. Access control is fundamentally based on user attributes, validated by SAML Assertions. In Fig.15, we can see the evolution of SAML, Shibboleth and XACML[34].

e. ID-WSF

In 2001, a business alliance was formed to serve as open standards organization for federated identity management and it was named Liberty alliance [16]. Its goals are to guaranty interoperability, support privacy, and promote adoption of its specifications, guidelines and best practices. The key objectives of the Liberty Alliance are to :

- Enable users to protect their privacy and identity
- Enable SPs’ to manage their clients
- Provide an open federated SSO
- Provide a network identity infrastructure that supports all current emerging network access devices

The Liberty Alliance’s work in the first phase is to enable federated network identity management. It offers among others SSO and linking accounts in the set of SPs’ in the boundary of the circle of trust. This work of this phase is referred to as Identity Federation Framework (ID-FF).

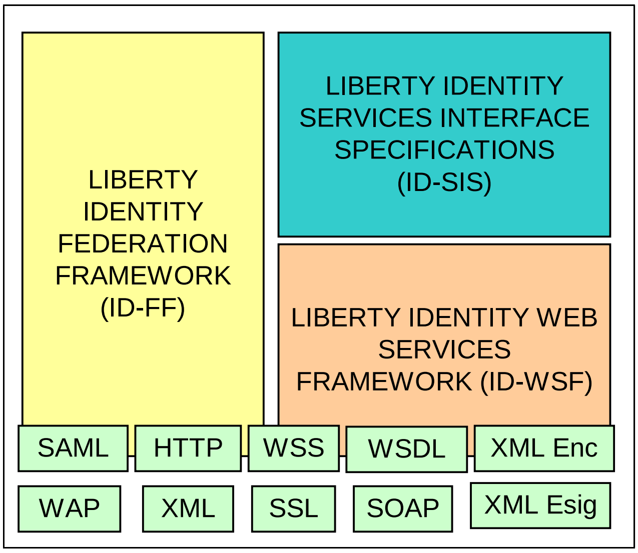

LIBERTY IDENTITY FEDERATION FRAMEWORK (ID-FF)
LIBERTY IDENTITY SERVICES INTERFACE SPECIFICATIONS (ID-SIS)
LIBERTY IDENTITY WEB SERVICES FRAMEWORK (ID-WSF)
SAML
HTTP
WSS
WSDL
XML Enc
WAP
XML
SSL
SOAP
XML Esig

Fig. 16. High-Level Overview of the Liberty Alliance Architecture

In the second phase, the specifications offer enhancing identity federation and interoperable identity-based Web services. This body is referred to as Identity Web Services Framework (ID-WSF). This framework involves support of the new open standard such as WS-Security developed in OASIS. ID-WSF is a platform for the discovery and invocation of identity services – Web services associated with a given identity-. In the typical ID-WSF use case, after a user authenticates to an IdP this fact is asserted to a SP through SAML-based SSO. Embedded within the assertion is information that the SP can optionally use to discover and invoke potentially numerous and distributed identity services for that user. For some scenarios which present an unacceptable privacy risk as it suggests the possibility of a user's identity being exchanged without their consent or even knowledge. ID-WSF has a number of policy mechanisms to guard against this risk but ultimately, it is worth noting that many identity transactions (e.g. automated bill payments) already occur without the user's active real-time consent – and users appreciate this efficiency and convenience.

To build additional interoperable identity services such as registration services, contacts, calendar, geo-location services, alert services, it's envisaged to use ID-WSF. This specification is referred to as the Identity Services Interfaces Specifications (ID-SIS). The Liberty Alliance specifications define the protocol messages, profiles, and processing rules for identity federation and management. They rely heavily on other standards such as SAML and WS-Security which is another OASIS specification that defines mechanisms implemented in SOAP headers.

These mechanisms are designed to enhance SOAP messaging by providing a quality of protection through message integrity, message confidentiality, and single message authentication. Additionally, Liberty has contributed portions of its specification back into the technical committee working on SAML. Other identity management enabling standards include:

- Service Provisioning Markup Language (SPML)
- XML Access Control Markup Language (XACML)
- XML Key Management Specification (XKMS)
- XML Signature
- XML Encryption

The WS-* (the Web Services protocol specifications) are a set of specifications that is currently under development by Microsoft and IBM. It is a part of larger effort to define a security framework for Web services, the resultant of proposals are often referred to as WS-*. It includes specifications as WS-Policy, WS-Security Conversation, WS-Trust, and WS-Federation. This last one has functionality for enabling pseudonyms and attribute-based interactions. Therefore, WS-Trust's has the ability to ensure security tokens as a means of brokering identity and trust across domain boundaries [32].

The Liberty Alliance is developing and delivering specification that enables federate network identity management. Fig.16 shows an Overview of the Liberty Alliance architecture as describe in the introduction to the Liberty Alliance identity architecture.

f. OpenID 2.0

Brad Fitzpatrick is at the origin of the development of the OpenID 1.0. The intent of the OpenID framework is to specify layers that are independent and small enough to be acceptable and adopted by market [21]. OpenID is basically providing simple attribute sharing for low-value transactions. It does not depend on any preconfigured trust model. The version 1.0 has deal with http based URL authentication protocol. OpenID authentication 2.0 is becoming an open platform that supports both URL and XRI user identifiers. In addition, it would like to be modular, lightweight and user oriented. Indeed, OpenID auth. 2.0 allows user to choose/control/manage his/her identity address. Moreover, the user choose his/her Identity Provider and have a large interoperability of his/her identity and can dynamically uses new services that stand out attribute verification and reputation without any loose of features. No software is required on user's side as the user interacts directly with the identity provider's site. This approach jeopardize the user identity because it could be hacked or theft. Moreover the user has no ability to examine tokens before they are sent. At the beginning of identity management each technology came with its own futures without any interest for others. Later, the OpenID 1.0 community has realized the importance of integrating other technologies as OASIS XRDS which is useful for his simplicity and extensibility.

OpenID Stack: The first layer is for supporting users' identification. Using URL or XRI form, we can identify an user. URL use IP or DNS resolution and is unique and ubiquitously supported. It can be as a personal digital address as used by blogers even though it is not yet largely used.

XRI (EXtensible Resource Identifier) is being developed under the support of OASIS and is about Addressing.

I-names are a generic term for XRI authority names that provide abstract identifiers for the entity to which they are assigned. They can be used as the entry point to access data under the control of that authority. Like a domain name, the physical location of the information is transparent to the requester.

OpenID 2.O provides a private digital address to allow a user to be only identified in specific conditions. This is guaranty the user privacy in a public domain.

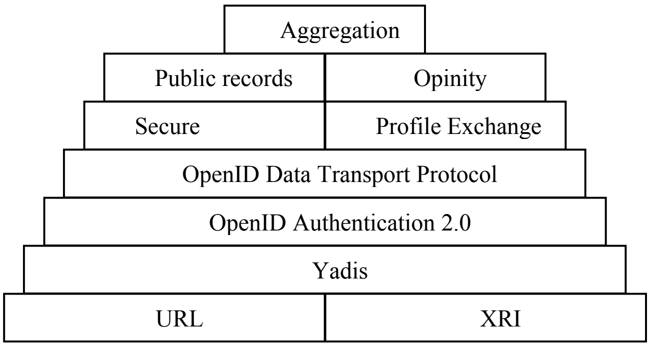


Fig. 17. OpenID protocol stack

Discovery. Yadis is used for identity service discovery for URLs and XRI resolution protocol for XRIs. The both use OASIS format called XRDS (Extensible Resource Description Sequence). The protocol is simple and describes any type of service.

Authentication. This service lets a user to prove his/her URL or I-name using credentials (cryptographic proof). This protocol is explained in Fig.17. The OpenID doesn't need a centralized authority for enrollment and it is therefore a federated identity management. With the OpenID 2.0 the IdP offers the user the option of selecting a digital address to send to the SP. To ensure anonymity, IdP can randomly generate a digital address used specially for this SP.

Data Transport. This layer ensures the data exchange between the IdP and SP. It supports push and pulls methods and it is independent from authentication procedures. Therefore, the synchronization of data and secure messaging and other service will be enabled. The data formats are those defined by SAML, SDI (XRI Data interchange) or any other data formats. This approach will enable evolution of the OpenID platform.

The four layers construct the foundation of the OpenID ensuring user centricity. There are three points to guaranty this paradigm:

a. User choose his/her digital identity
b. User choose IdP
c. User choose SP

OpenID is decentralized and well founded and at the same time simple, easy to use and to deploy. It provides open development process and single sign-on for the

Web and ease of integration into scripted Web platforms (e.g. Drupal, WordPress, etc). So, it is a greater future for him. You can learn about OpenID at openidenabled.com also the community of OpenId can be joined at opened.net.

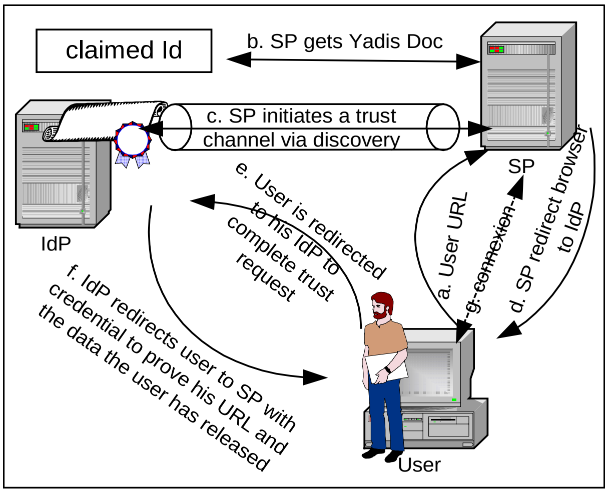

claimed Id
b. SP gets Yadis Doc
c. SP initiates a trust channel via discovery
IdP
SP
e. User is redirected to his IdP to complete trust request
f. IdP redirects user to SP with credential to prove his URL and the data the user has released
a. User URL
g. connexion
d. SP redirect browser to IdP
User

Fig. 18. OpenID 1.1 protocol flow

g. InfoCard

Rather than invent another technology for creating and representing digital identities, Microsoft has adopted the federated user-centric identity meta-system. This is a serious solution that provides a consistent way to work with multiple digital identities. Using standard protocols that anyone can implement on any platform, the identity meta-system allows the acquisition and use of any kind of security tokens to convey identity.

The "Infocard" is the Microsoft's codename for this new technology that tackles the problem of managing and disclosing identity information. The "InfoCard" implements the core of the Identity Metasystem, using open standard protocols to negotiate, request and broker identity information between trusted IdPs and SPs. "InfoCard" is a technology that helps developers to integrate a consistent identity infrastructure into applications, Web sites and Web services.

By providing a way for users to select identities and more, Windows CardSpace [12] plays an important part in the identity meta-system.

It provides the consistent user experience required by the identity meta-system. It is specifically hardened against tampering and spoofing to protect the end user's digital identities and maintain end-user control. Windows CardSpace enables users to provide their digital identities in a familiar, secure and easy way.

In the terminology of Microsoft, relying party is in our model service provider (SP). To prove an identity over a network, the user emitted credentials which are some proofs about his/her identity. For example in the simplest digital identity the user name is the identity while the password is said to be the authentication credential. In the terminology of Microsoft and others, there are called security token and contain one or more claims. Each claim contains information about the users, like the user name or home address, etc. In addition, security token encloses prove that the claims are correctly emitted by the real user and are belonging to him. This is could be done cryptographically using different forms such as X.509 certificates and Kerberos tickets but unfortunately there are not practical to convoy different kind of claim. The standard SAML as seen before is the indicated one for this purpose as it can be used to define security tokens. Indeed, SAML token could enclose any desired information and thus become as largely useful in the network to show and control digital identity.

CardSpace runs on Windows Vista, XP, Server 2003 and Server 2008, based on .NET3, and also uses Web service protocols :

. WS-Trust
. WS-Policy
. WS-SecurityPolicy
. WS-MetaDataExchange

CardSpace runs in a self virtual desktop on the PC. Thereby, it locks out other processes and reduces the possibility of intercepting information by a spyware.

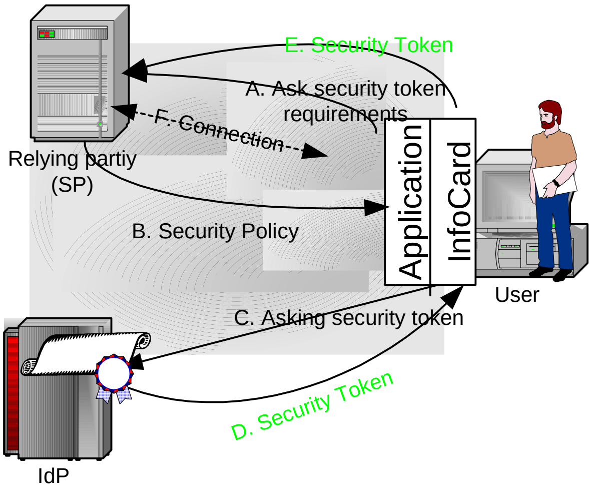


Fig. 19. Interactions among the users, identity providers and relying party

Fig.19 shows that the architecture is fitting exactly to the principle of Identity 2.0. The user access one of any of his/her relying parties (SPs) using an application that supports CardSpace. When the choice is made, the application asks for the requirement of security token of this specific SP that will answer with SP policy. It really contains information about the claims and the accepted token formats. Once is done the application passes these requirements to CardSpace which asks the security token from an appropriate identity provider. Once this security token has been received, CardSpace transmits via application to the relying party. The relying party can then use this token to authenticate the user.

Please note that each identity is emitted by an identity provider and is stored at the user side. It contains the emitter, the kind of security token he/she can issue and the details about the claims' enclose. All difficulties are hidden to the user as he/she has only to choose one of InfoCard when the process of authentication is launched. Indeed, once the requirements information are returned and passed to CardSpace, the system displays the card selection matching the requirements on screen. In this regard, the user has a consistent experience as all applications based on CardSpace will have the same interface, and the user do not have to worry about the protocol used to express identity's security token. The PIN number is entered by user and the choice of his/her card is done in a private Windows desktop to prevent locally-running process.

h. SXIP 2.0

In 2004, The SXIP 1.0 grows from efforts to build a balanced online identity solution that met the requirements of the entire online community. Indeed, SXIP 2.0 is the new generation of the SXIP 1.0 protocol that was a platform that gives users control over their online identities and enables online communities to have a richer relationship with their membership. SXIP 2.0 defines entities' terminology as:

- Homesite: URL-based identity given by IdP.
- Membersite : SP that uses SXIP 2.0.
- User: equivalent to the user in our model.

The Simple eXtensible Identity Protocol (SXIP) [13] was designed to address the principles defined by the Identity 2.0 model, which proposes an Internet-scalable and user-centric identity architecture that mimics real-world interactions.

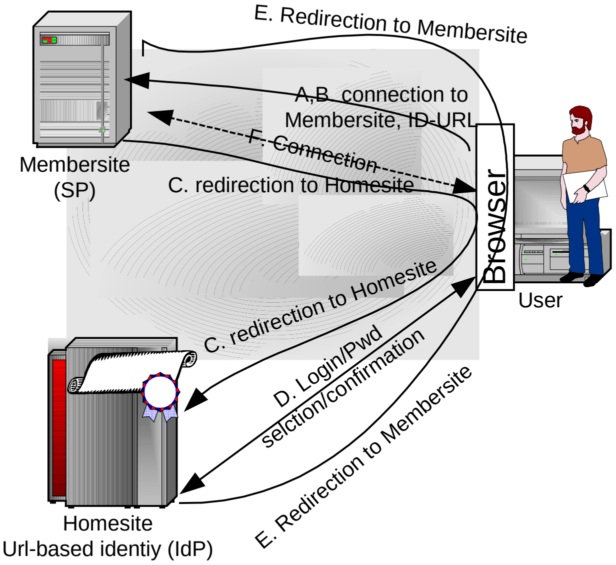

E. Redirection to Membersite
A,B. connection to Membersite, ID-URL
F. Connection
Membersite (SP)
C. redirection to Homesite
Browser
User
C. redirection to Homesite
D. Login/Pwd selction/confirmation
E. Redirection to Membersite
Homesite
Url-based identiy (IdP)

Fig. 20. SXIP entities interactions

If a SP has integrated a SXIP to his Website, which is easy done by using SDKs, he is a Membersite. When a subscriber of SXIP would like to access this Membersite:

a) types his/her URL address and clicks on [Sxip in],
b) types his/her URL identity issued by IdP (called Homesite)
c) browser is redirected to the Homesite
d) enters his/her username and password, being informed that the Membersite has requested data, selects the related data and verify it and can select to automatically release data for other visit to this Membersite and confirms
e) browser is redirected to the Membersite
f) have access to the content of the site.

SXIP 2.0 is a platform based on a fully decentralized architecture providing an open and simple set of process for exchanging identity information. SXIP 2.0 has significantly reduced the problems resulting from moving identity data form one site to another. It is URL-based protocol that allows a seamless user experience and fits exactly to user-centric paradigm. In that sense, the user has full control on his/her identity and has an active role in the exchange of his/her identity data. Therefore, he/she can profit from portable authentication to connect many Websites. Doing so, user has more choice and convenience when exchanging his/her identity data and enables indirectly Websites to offer enhanced services to their subscribers.

SXIP 2.0 provides the following features:

• Decentralized architecture: SXIP 2.0 is completely decentralized and is a federated identity management. The online identity is URL-based and the user identity is separated from the authority that issues the identifiers for this identity. In this regard, we can easily move the location of the identity data without losing the associated identifier.

• Dynamic discovery: A simple and dynamic discovery mechanism ensures that users are always informed online about his/her home site that is exporting identity data.

• Simple implementation: SXIP 2.0 is open source using different high level development languages such as Perl, Python, PHP, and Java. Therefore, the integration of SXIP 2.0 into a Website is effortless. It does not require PKI as it uses a URL-based protocol that do not need it.

• Support for existing technologies: SXIP 2.0 uses simple Web browsers, the primary client and means of data exchange, providing users with choice in the release of their identity data.

• Interoperability: SXIP 2.0 can coexist with other URL-based protocols.

• Richer data at an Internet scale: SXIP 2.0 messages consist of lists of simple name value pairs. It can exchanged simple text, claims using SAML and third-party claims in one exchange and present them in many separate exchange. In addition, the Identity provider is not bothersome every time identity is requested.

Finally by using SXIP 2.0, Websites can also be authoritative about users for data, such as third-party claims. Those are keys to build online reputation, further enriching the online exchange of identity data.

i. Higgins

Higgins [36] is a project supported principally by IBM and it is a part of IBM's Eclipse open source foundation. It will also offer libraries for Java, C and C++, and plug-ins for popular browsers. It is really an open source trust framework which goals are to support existing and new applications that give users more convenience, privacy and control over their identity information. The aim objective is to develop an extensible, platform-independent, identity protocol-independent, software framework that provides a foundation for user-centric identity management. Indeed, it enables applications to integrate identity, profiles and relationship across heterogeneous systems.

The main goal of Higgins as an identity management systems are interoperability, security and privacy that are a decoupled architecture. This system is a real user-centric based on a federated identity management. The user has the ability to use a pseudonym or simply reply anonymously in case you would not give your name. We use the term context to cover a range of underlying implementations. A context can be thought of as a distributed container-like object that contains digital identities of multiple people or processes.

The platform intends to address four challenges:

- ✓ the need to manage multiple contexts,
- ✓ the need for interoperability,
- ✓ the need to respond to regulatory, public or customer pressure to implement solutions based on trusted infrastructure that offers security and privacy, and
- ✓ the lack of common interfaces to identity/networking systems.

Higgins matches exactly the user-centric paradigms because it offers consistent user experience based on card icons for the management and release of identity data. Thereby, there is less vulnerability to Phishing and other attacks. Moreover, user privacy is enabled by sharing only what is needed. Thus, the user has a full control on his/her personal data. Identity Attribute Service enables aggregation and federation of identity systems and even silos. For enterprises, it integrates all data related to identity, profile, reputation, and relationship information across and among complex systems.

Higgins is a trust framework that enables users and enterprises to adopt, share across multiple systems and integrate to new or existing application, digital identity, profiles, and cross-relationship information. In fact, it facilitates as well the integration of different identity management systems as the management of identity, profile, reputation and relationship data across repositories. Using context providers, directories and communications technologies (e.g. Microsoft/IBM WS-*, LDAP, email, etc.) can be plugged into the Higgins framework. Higgins has become an Eclipse plug-in, and is a project of the Eclipse Foundation. Any application developed with Higgins will enable users to share identity with other users under a strict control.

Higgins is benefic for developers, users and enterprise. Higgins relieves the developers from knowing all the details of multiple identity systems, thanks to one API that support many protocols and technologies: CardSpace, OpenID, XRI, LDAP, etc. An Application written to the Higgins API can integrate the identity, profile, and relationship information across these heterogeneous systems. The goal of the framework is to be useful in the development of applications accessed through browsers, rich clients, and Web services. Thus, the Higgins Project is supported by IBM and Novell and thwart InfoCard Microsoft's project.

The Higgins framework intents to define in terms of service descriptions, messages and port types consistent with an SOA model and to develop a Java binding and implementation as an initial reference. Applications can use Higgins to create a unified, virtual view of identity, profile and relationship information. A key focus of Higgins is providing a foundation for new "user-centric identity" and personal information management applications. Finally, Higgins provides virtual integration; user-centric federated management model and trust brokering that are applied to identity, profile and relationship information. Furthermore, Higgins provides common interfaces to identity and thanks to data context it encloses enhanced automation process. Those features are also offered across multiple contexts, disparate systems and implementations. In this regard, Higgins is a full interoperable framework.

The Higgins service acts together with a set of so-called context providers which can represent a department, association, informal network and so on. A context is the environment of Higgins and digital identities, the policies and protocols that govern their interactions. Context providers adjust existing legacy systems to the framework, or implement new ones Context providers may also contain the identities of a machine or human. A context encloses a group of digital identities and their related claims and links. A Context maintains a set of Claims about properties and values (e.g. name, address, etc.). It is like security token for Cardspace. The set of profile properties, the set of roles, and the access rights for each role are defined by and controlled by the Context Provider.

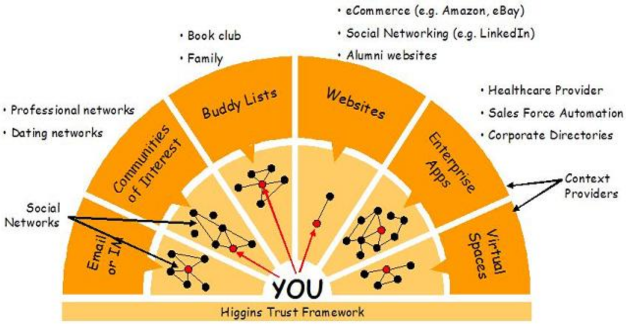


Fig. 21. Higgins Trust Framework and context [14]

Context providers act as adapters to existing systems. Adapter providers can connect for example to LDAP servers, identity management systems like CardSpace, mailing list and social networking systems. A Higgins context provider has the ability to implement the Context interface and thus empower the applications layered on top of Higgins.

3.2.2 Summarizing Table

The 10 requirements at the top of the following table are those discussed in Section 2.4. In this table (table 1), white means that the requirement is not covered, grey partially and black fully fulfilled.

Table 1. Evaluation of identity 2.0 technologies

Requirement	Empowering the total control of users over their privacy rivacy	Usability, as users are using the same identity for each identity transaction	Giving a consistent user's experience thanks to uniformity of identity interface	Limiting identity attacks i.e. Phishing	Limiting reachability/disturbances, such as spam	Reviewing policies on both sides when necessary, identity providers and service providers	Huge scalability advantages as the Identity Provider does not have to get any prior knowledge about the Service	Assuring secure conditions when exchanging data	Decoupling digital identity from applications	Pluralism of Operators and Technologies
XRI/XDI	▨	▨	□	□	□	□	□	▨	▨	□
ID/WSF	▨	▨	▨	□	□	▨	▨	▨	▨	▨
Shibboleth	▨	□	□	□	□	□	□	▨	□	□
CardSpace	▨	▨	▨	□	□	▨	▨	▨	▨	□
OpenID	▨	▨	▨	□	□	□	▨	▨	▨	▨
SXIP	▨	▨	▨	□	▨	□	▨	▨	▨	▨
Higgins	▨	▨	▨	□	▨	▨	▨	▨	▨	▨

At the moment, service providers have to choose between so many authentications and identity management systems and users are left to face the non-convenience of a variety of digital identities. The main initiatives have different priorities and some unique advantages, while overlapping in many areas. The most pressing requirements for users are interoperability, usability and centricity. Thanks to Higgins the majority of identity requirements are guaranty. Therefore, using it, the user is free to visit all Web sites without being worried about the identity management system used by the provider.

4 Identity 2.0 for Mobile Users

In this section, we will talk about mobility, his evolution and his future.

4.1 Introduction

The number of devices such as mobile phones, smart cards and RFIDs [41], is increasing daily and becoming huge. Mobile phones have attracted particular interest because of their large penetration and pervasiveness that exceeds that of personal computers. Furthermore, the emergence of both IP-TV and wireless technology has facilitated the proliferation of intelligent devices, mobile phones, RFIDs, and other forms of information technology that are developing at a rapid speed. These devices include a fixed identifier that could be linked to the user's identity. This identifier provides a mobile identity which takes into account information about the location and the mobile user's personal data.[37]

4.2 Mobile Web 2.0

Mobile Web 2.0 as a content-based service is an up-to-date offering of services within the mobile network. As the number of people having access to mobile devices exceeds those using a desktop computer, *mobile Web* will be a key factor for the next generation network. At the moment, mobile Web suffers from lack of interoperability and usability due to the small screen size and lower computational capability. Fortunately, these limitations are only temporary and within 5 years they will be easily overcome. There will be convergence in the next generation public networks towards the mobile network which will bring mobility to the forefront. Thus, mobile identity management will play a central role in addressing issues such as usability, privacy and security which are key challenges for researcher in the mobile network. Since the initial launch of mobile Web services, customers have increasingly turned to their wireless phones to connect with family and friends and also to obtain the latest news and information or even to produce content with their mobile and then publish them. Mobile Web 2.0 [40] is the enforcement of evolution and will enhance the experience of users by providing connections in an easier and more efficient way. For this reason, it will be welcome by the key actors as a well-established core service identity management for the next generation mobile network. This mobile identity management will be used not only to identify, acquire, access and pay for services but also to offer context-aware services as well as location based services.

4.3 Mobility

The mobile identity may not be stored at the same location but could be distributed among many locations, authorities and devices. Indeed, identity is mobile in many respects [1]:

a. There is a device mobility where a person is using the same identity while using different devices;
b. There is a location mobility where a person is using the same devices while changing the location; and
c. There is context mobility where a person is receiving services based on different societal roles: as a parent, as a professional and so on.

The three kind of mobility are not isolated but they interacted more often and became concurrently modified creating much more complex situations that what implied from single mode. Mobile identity management addresses three main challenges: a. usability via context awareness b. trust based on the perception of secure operation and c. the protection of privacy [1].

4.4 Evolution of Mobile Identity

Mobile identity management is in its infancy. GSM networks, for example, provide management of SIM identities as a kind of mobile identity management, but they do not meet all the requirements for a complete Mobile identity management.

Unlike static identity, already implemented in Web 2.0 identity, dynamic aspects, such as the user's position or the temporal context, gain increasingly importance for new kinds of mobile applications.[35]

Mobile identity (MId) infrastructure solutions have evolved over time and can be classified into three solutions. The first proposed solution is just an extension of wired identity management to mobile Internet. This is the widespread solution, which is limited to the users of mobile devices running the same operating system as wired solution. This limitation is expected to evolve over time mainly with the large deployment of Web services. Some specifications, such as Liberty Alliance specifications, have been developed for identity management including mobility. However, several limitations are observed when the MId system is derived from fixed context. These limitations are principally due to the assumptions during their design and they do not match well with extra requirement of mobility [1].

Many improvements such as interoperability, privacy and security are to be operated and also older centralized PKI must be replaced by modern trust management system or at least a decentralized PKI. The second solution is capable of providing an alternative to the prevalent Internet derived MId infrastructure consisting of either connected (Cellular phones) or unconnected (Smartcards) mobiles devices.

The third one consists of using implantable radio frequency identity (RFID) devices. This approach is expected to increase rapidly even if the market penetration is smaller than cellular phones.

In addition, the sensitivity risk of data related to different applications and services are seldom at the same level and the number of identifiers used by a person is in constant increasing. Thus, there is a real need of different kind of credentials associated with different kind of applications. Indeed, a tool at the user side capable of managing the credentials and identifies is inevitable. With the increasing capacity of CPU power and the spreading number of mobile phone with a SIM card, mobile phones can be considered as a Personal Authentication Device (PDA). They can hold securely the users' credentials, password and even identities. Thereby, we introduced a new efficient Identity management device at the user side able to facilitate the memorization in one hand, and strengthen the security by limiting the number of passwords and their weakness in other hand. All wired identity management can be deployed using PDA. In addition, many different authentication architectures become possible and easy to implement such as dual channel authentication.

4.4.1 PDA as Solution to Strong Authentication

PDA is a tamper-resistant hardware device which could include smart card and sensors or not. As it is used for authentication it is called a personal authentication device (PDA) [42]. This term has been early used in the context of security by Wong and al. [43]. The approach is the same and the only thing change so far is the performance of the mobile device has radically changed. This is the opportunity to emphasis the user centricity as the PDA could strengthen the user experience and to facilitate the automation and system support of the identity management at the user side. The Fig.22 illustrated the combination of PDA and silo model. The user stores his/her identity in the PDA. Whenever he/she would like to connect to a Service provider

a. he/she authenticates her/himself with a PIN code to use the PDA.
b. the user choose the Password to be used for his/her connection to the specific service provider
c. the user launch and log to the specific service provider by entering his/her Username and the Password.

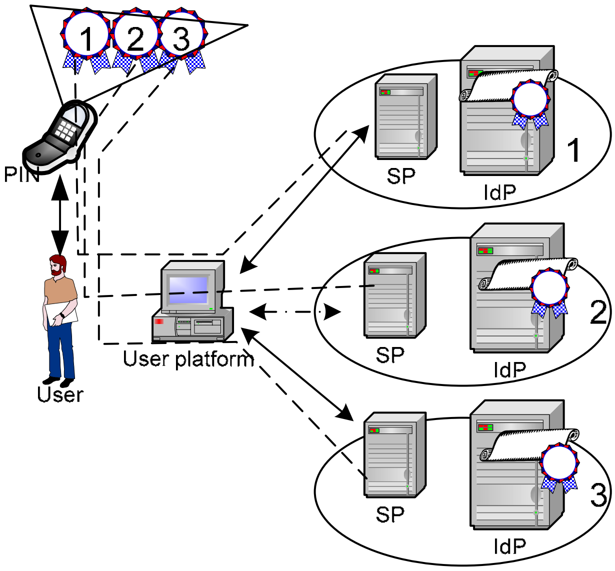


Fig. 22. Integration of PDA in silo model

The PDA is a good device to tackle the weakness and non-convenience of password authentication due to its Thereby, we have a user friendly and user centric application and even introducing stronger authentication. The fundamental advantage of PDA comparing with common PC using common operating systems such as windows or linux is that PDA has a robust isolation of processes. Therefore, compromising one application does not compromise all the applications. This advantage is becoming less important for mobile phone as flexibility is introduced by manufacturers a lot of vulnerabilities is also introduced. We have seen many viruses for mobile phones and even nowadays we have viruses for RFID. This vulnerability can compromise authentication and even biometrics authentication. That's why we should be very vigilant in implementing security in PDA devices. An ideal device is the USB stick running a standalone OS, and integrating a biometric reader and mobile network access. A can find some of them with fingerprint reader for a reasonable price.

Two main categories can group many authentication architectures that could be implemented in a PDA. There are single and dual channel authentications. Thereby, the cost, the risk and the non-convenience could be tackled at the same time.

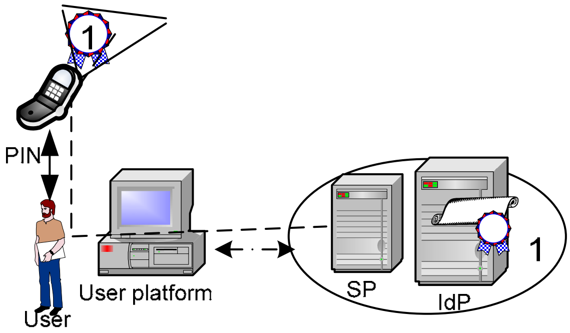


Fig. 23. Single channel authentication

Fig.23 illustrates the principle of single channel authentication which is the first application of the PDA. In Fig.25 the second principle of double channel authentication which is more secure as the

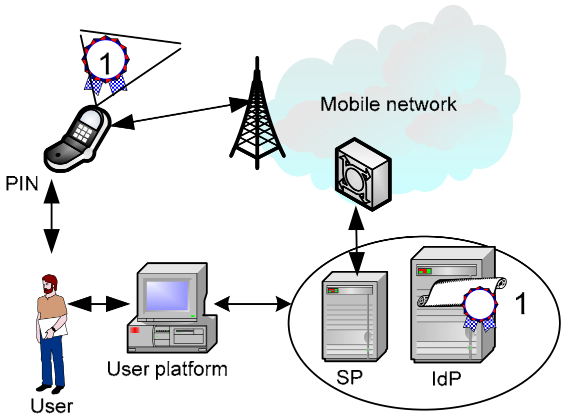


Fig. 24. Dual channel authentication

4.4.2 Different Kinds of Strong Authentication through a Mobile PDA

The mobile network mainly GSM can help to overcome a lot of security vulnerabilities such as phishing or man-in-the-middle. It attracts all business that would like to deploy double channel authentication but are worry about cost and usability. The near-ubiquity of the mobile network has made feasible the utilization of this approach and even being adopted by some banks.

a. SMS based One-Time Password (OTP)

The main advantages in mobile network are the facility and usability to send and receive SMSs. Moreover, they could be used to setup and download easily Java program to the mobile device. In addition, mobile devices are using smart card that can securely calculate and store claims. The cost is minimized by adopting a mobile device using SMS to receive OTP instead of a special hardware that can generate OTP.

The scenario implemented by some banks is illustrated in Fig.26 and it is as follow:

First of all, the user switches his/her mobile phone and enters his PIN code then

a. The user log into his online account by entering his/her Username and Password (U/P).
b. The Web site received the couple U/P.
c. The server verifies the couple
d. Send a SMS message with OTP
e. The user reads the message
f. The user enters the OPT into online account
g. The server verify the OPT and give access

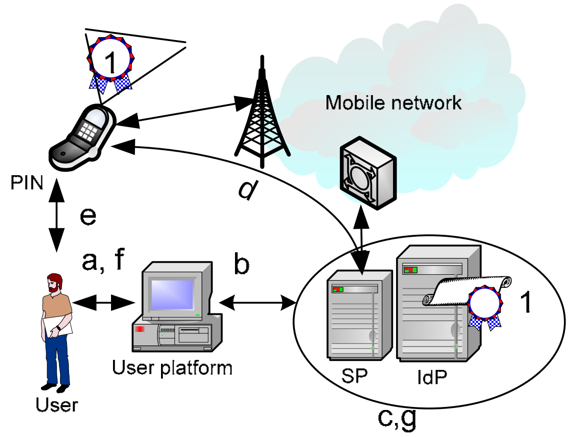


Fig. 25. Scenario of SMS double channel authentication

The problem of this approach is the fact that the cost is assumed by the service provider. In addition, some drawbacks are very common mainly in some developing countries such as lack of coverage and SMS latency. Of course, the attack of the man-in-the-middle is not overcome by this approach.

b. Soft Token Application

In this case, the PDA is used as a token emitter. The application is previously downloaded. SMS could be sent to the user in order to set up the application that will play the role of soft token.

The scenario is exactly identical to the SMS but only the user generates his/her OTP using the soft token instead of waiting for a SMS message. The cost is less than the SMS based OTP. This approach is a single channel authentication that is not dependent on mobile network coverage neither on his latency. Furthermore, the attack of the man-in-the-middle is not tackle.

c. Full Option Mobile Solution

We have seen in the two previously scenarios that the attacks of the man-in-the-middle is not addressed. It exist a counterattack to this security issue consisting of using the second channel to completely control all the transactions over the online connection. Of course, the security of this approach is based on the assumption that it is difficult for an attacker to steal the user's personal mobile phone or to attack the mobile network. Anyway, we have developed an application to crypt the SMS message which minimizes the risk of attacks.

The scenario is illustrated in the Fig.26 and it is as follow:

a. The user login on online account using token
b. The server receives the token
c. The server verifies the token
d. the access is given to the service
e. the user request a transaction
f. SMS message is send with the requested transaction and a confirmation code
g. The user verifies the transaction
h. He enters the confirmation code
i. The server verifies and execute the transaction
j. The server sends a transaction confirmation

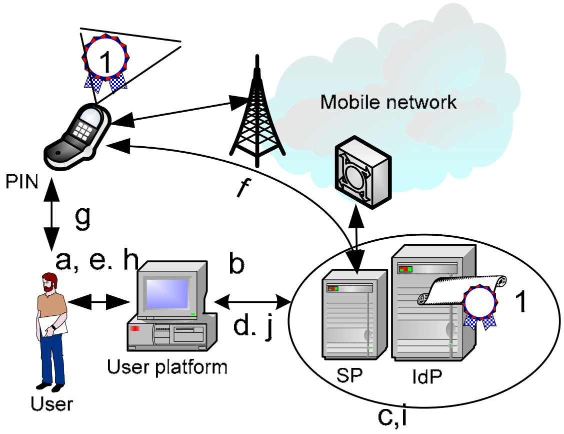


Fig. 26. Secure transaction via SMS

4.5 Future of Mobile User-Centric Identity Management in an Ambient Intelligence (AmI) World

AmI manifests itself through a collection of everyday devices incorporating computing and networking capabilities that enable them to interact with each other, make intelligent decisions and interact with users through user friendly multimodal interfaces. Ambient intelligence is driven by users' needs and the design of its capabilities should be driven by users' requirements.

Ambient Intelligence technologies are expected to combine concepts of ubiquitous computing and intelligent systems putting humans in the center of technological developments. In deed, the Internet extension to home and mobile networks, the multiplication of modes of connection will make the individual the central point. Therefore, the identity is a challenge in this environment and will guarantee the infatuation with Ambient Intelligence. Moreover, AmI will be the future environment where we shall be surrounded by mobile devices which will be more and more used for mobile interactions with things, places and people.

The low-cost and the shrinking size of sensors as well as the ease of deployment will aid ambient intelligence research efforts for rapid prototyping. Evidently, a sensor combined with unique biometric identifiers is becoming more frequently utilized in access a system, and supposedly provide proof of a person's identity and thus accountability for subsequent actions.

To explore these new AmI technologies, it is easier to investigate a scenario related to ubiquitous computing in an ambient intelligence environment.

4.5.1 AmI Scenario

A person having a mobile device, GPS (or equivalent) and an ad-hoc communication network connected to sensors, visits an intelligent environment supermarket and would like to acquire some merchandise. We illustrate below how this person can benefit from mobile identity.

When this person enters the supermarket, he/she is identified by means of his/her mobile device or implemented RFID tag and a special menu is displayed to him/her. His/her profile, related to his/her context identity, announces a discount if there is one. The members of his/her social network could propose him/her a connection if they are present and even guide him to their location.

Merchandise on display could communicate with his/her device to show prices and details. Location-based services could be offered to quickly find his/her specific articles. His device could help him to find diabetic foods or any restrictions associated with specific articles. A secure Web connection could be initiated to give more information about purchases and the user account.

An adaptive screen could be used by the supermarket to show him/her information that is too extensive for his/her device screen. Payment could be carried out using payment identity stored in his/her device and even a biometric identity to prevent identity theft. Identity information and profiling should be portable and seamless for interoperability. The identity must be managed to ensure user control. Power and performance management in this environment is a must.

The concept of authentication between electronic devices is also highlighted. In order to use identity management, the user needs an appropriate tool to facilitate the management for the disclosure of personal data. A usable and secure tool should be proposed to help even inexperienced users manage their general security needs when using the network.

We need mobile identity management, which is a concept that allows the user to keep his or her privacy, depending on the situation. By using identity management, the user's device acts in a similar way to the user. In different contexts, the user presents a different appearance. Devices controlled by identity management change their behavior similar to the way in which a user would.

4.5.2 Requirements for Mobile User-centric Identity Management in an AmI world

As the network evolution is toward mobility with the proliferation of ubiquitous and pervasive computing systems, the importance of identity management to build trust relationships in the context of electronic and mobile (e/m) government and business is evident [30, 31]. Thereby, all these systems require advanced, automated identity management systems in order to be cost effective and easy to use.

Several mobile devices such as mobile phones, smart cards of RFID are used for mobility. As mobile devices have fixed identifiers, they are essentially providing a mobile identity that can be liked to a user. Mobile identity takes into account location data of mobile users in addition to their personal data. A recent court decision in the UK has established as proof of location of the accused the location trace of his mobile phone which implies a de facto recognition of the identity of a citizen as the identity of her mobile telephones. [1]

That is why Mobile identity management (MIdm) is necessary to empower mobile users to manage their mobile identities to enforce their security and privacy interests. Mobile identity management is a special kind of identity management. For this purpose, mobile users must be able to control the disclosure of their mobile identity dependent on the respective service provider and also their location via mobile identity management systems.

Ambient Intelligence emphasizes the principles of secure communication anywhere, anytime, with anything. The evolution of AmI will directly influence identity management with this requirement to ensure mutual interaction between users and things. *Being Anywhere* will imply more and more mobility, interoperability and profiling. *At Anytime* will imply online as well as offline connection as the network does not have a 100% coverage and will imply power as well as performance management in order to optimize use battery. *With anything* will imply sensor use, biometrics and RFID interaction; and *Securely* implies more and more integration of privacy, authentication, anonymity and prevention of identity theft.

From multilateral security [39,38], Jendricke [36] has derived privacy principles for MIdm and we have completed them below with a few other important principles.

Management systems:

1. Context-detection
    a. Sensors
    b. Biometrics
    c. RFID
2. Anonymity
3. Security
    a. Confidentiality
    b. Integrity
    c. Non-repudiation

   d. Availability
4. Privacy
   a. Protection of location information
5. Trustworthiness
   a. Segregation of power, separating knowledge, integrating independent parties
   b. Using Open Source
   c. Trusted seals of approval seal
6. Law Enforcement / Liability
   a. Digital evidence
   b. Digital signatures
   c. Data retention
7. Usability
   a. Comfortable and informative user interfaces
   b. Training and education
   c. Reduction of system' complexity
   d. Raising awareness
8. Affordability
   a. Power of market :Produce MIMS that are competitive and are able to reach a remarkable penetration of market
   b. Using open source building blocks
   c. Subsidies for development, use, operation, etc.
9. Power management : the energy provided by the batteries of mobile devices is limited and that energy must be used with care on energy-friendly applications and services
10. Online and offline identity proof
11. Small screen size and lower computational capability
12. Interoperability
13. Identity needs to be portable to be understood by any device.

4.6 Research Directions

Future identity management solutions will play a more central role in the IT industry due to the pervasiveness and the increased presence of identity information in all components of the IT stack. The Liberty Alliance specifications provide a standardized solution to the problem of lack of capability for mobile identity. The specified architecture will have an impact on the architecture of mobile services. However many open issues remain to be considered.

There will be many issues raised concerning identity and mobility. All strong platform combining identity management and mobility play a central role and will be key elements for the progress of ambient intelligence network; it will also represent the main vehicle for the Information society.

We list below a few of the most important research questions in the field of mobile user-centric identity management:

- **Requirements.** How can we satisfy the requirements of mobile identity management in mobile system and devices?

- **Mobility.** How can we mange identity when the device is off-line? How can we manage biometric information? How can mobility and biometrics help in authentication?

- **Privacy.** What are the identity management needed controls to preserve individual privacy? How can we guarantee anonymity? How can we protect the user from location/activity tracking?

- **Forensic science**. What is the reliability of the identity management system, and how can evidence extracted from the system be used in court? What protections are in place for the identity holder or for the relaying party?

- **Costs of infrastructure.** How can we limit the cost of mobile identity management?

- **Interoperability**. How can we evolve towards higher levels of interoperability?

- **Efficiencies**. How can we efficiently integrate sensors, RFIDs and biometric into mobile identity management systems? How can we manage performances of mobile devices? How can we integrate usability into mobile identity usage?

- **Identity theft.** Does large scale deployment of identity management systems make it easier or harder to perpetrate identity theft or identity fraud? How can we deal with theft (device or identity)?

- **Longevity of information.** Do mobile identity management systems provide adequate care in tracking changes in identity information over time?

- **Authenticity of identity.** What are the trust services that must be in place in order to generate confidence in the identity management service?

6. Conclusion

Internet is more and more used but the fact that the Internet has not been developed with an adequate identity layer is a major security risk. Password fatigue and online fraud are a growing problem and are damaging user confidence.

Currently, there are major initiatives trying to provide a more adequate identity layer for the Internet but their convergence has not been achieved yet. Higgins and Liberty Alliance seem to be the most promising ones. Anyway, future identity management solutions will have to work in mobile computing settings, anywhere and anytime.

We have underlined the necessity of mobility and the importance of identity in future ambient intelligent environments. Mobile identity management will have to support a wide range of information technologies and devices with critical requirements such usability on the move, privacy, scalability and energy-friendliness.

References

[1] George Roussos, Uma Patel, “Mobile Identity Management: An Enacted View”, Birkbeck College, University of London, City University, London, 2003.

[2] Westin, A.: “Privacy and Freedom”. Athenaeum, New York, NY (1967)

[3] John Madelin and al., BT report on: “Comprehensive identity management Balancing cost, risk and convenience in identity management”, 2007

[4] J.-Marc Seigneur, “Trust, Security and Privacy in Global Computing”, PhD Thesis, Trinity College Dublin, 2005.

[5] Introduction to the Liberty Alliance Identity Architecture. Rev. 1.0, March 2003.

[6] Abhilasha Bhargav Spantzel and al. “User Centricity: A Taxonomy and Open Issues”, IBM Zurich Research Laboratory, 2006

[7] Independent Center for Privacy Protection (ICPP) and Studio Notarile Genghini(SNG), “Identity Management Systems (IMS): Identification and Comparison Study”, 2003

[8] CAMERON, K. “Laws of Identity”, 5/12/2005

[9] A. Jøsang and S. Pope, “User Centric Identity Management”, AusCERT Conference 2005.

[10] ITU (International Telecommunication Union), Geneva, http://www.itu.org/

[11] A. Jøsang, al., “Usability and Privacy in Identity Management Architectures”, (AISW2007), Ballarat, Australia, 2007.

[12] [Microsoft, A technical ref. for InfoCard in windows http://msdn.microsoft.com/winfx/reference/infocard/,2005

[13] [J. Merrels, SXIP Identity. DIX: Digital Identity Exchange Protocol. Internet Draft, March 2006.

[14] Higgings Trust Framework project, http://www.eclipse.org/higgins/ 2006,

[15] OASIS Working Draft Version 04, “An Introduction to XRIs”, 14-March-2005

[16] Liberty Alliance, Liberty ID-FF Architecture Overview. Liberty Alliance Project, 2005.

[17] OASIS, Conformance Requirements for the OASIS Security Assertion Markup Language (SAML) V.20, 2005.

[18] Teruko MIYATA and al., “A Survey on Identity Management Protocols and Standards”, IEICE TRANS. INF & SYST, 2006

[19] Internet2, Shibboleth project, http://shibboleth.Internet2.edu

[20] Liberty Alliance, “Liberty Developer Tutorial” http://www.projectliberty.org/

[21] David Recordon VeriSign Inc, Drummond Reed, “OpenID 2.0: A Platform for User-Centric Identity Management”, 2006.

[22] Yadis, Yadis specification 1.0, released http://yadis.org, March 2006.

[23] C. Esslison et al., RFC 2693- SPKI Certification Theory. IETF, Sep. 1999, http://www.ietf.org/rfc/rfc2693.txt.

[24] T. M. Cooley, "A Treatise on the Law of Torts", Callaghan, Chicago, 1888.

[25] Identity Management Systems (IMS): Identification and Comparison Study, Independent Center for Privacy Protection (ICPP) and Studio Notarile Genghini(SNG), 2003

[26] A User Centric Anonymous Authorisation Framework in Ecommerce Environment Richard Au, Harikrishna Vasanta, KimKwang Raymond Choo, Mark Looi, Information

Security Research Centre Queensland University of Technology, Brisbane, AUSTRALIA

[27] Federal Financial Institutions Examination Council. Authentication in an InternetBanking Environment, October 2005. http://www.ffiec.gov/press/pr101205.htm.

[28] A. erzberg and A. Gbara. TrustBar : protecting even Naïve) Web Users from Spoofing and Phishing Attacks. 2004.

[29] Introduction to usability, http://www.usabilityfirst.com/intro/index.tx1, 2005

[30] MyGrocer Consortium MyGrocer Whitepaper, 2002

[31] M. Wieser The Computer for the Twenty-First Century, Scientific American, 1991.

[32] Teruko MIYATA and al., A Survey on Identity Management Protocols and Standards, IEICE TRANS. INF & SYST, 2006

[33] Liberty Developer Tutorial http://www.projectliberty.org/resources/LAP_DIDW_Oct-15_2003_jp.pdf

[34] SACML, http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml

[35] User-centric Identity Management in Open Mobile Environments Mario Hoffmann Fraunhofer-Institute for Secure Telecooperation (SIT)

[36] Uwe Jendricke and al., Mobile Identity Management, UBICOMP 2002

[37] Weis, Stephen A. and al. : Security and Privacy Aspects of low-Cost Radio Frequency Identification Systems; Proc. Of First International Conference On Security in Pervasive Computing, March 2003

[38] K. Reichenbach and al.: individual management of personal reachability in mobile communications, Proc. IFIP TC11 (Sec'97).

[39] K. Rannenberg. Multilateral security? a concept and examples for balanced security Proc. 9th ACM new security Paradigms Workshop, 2000.

[40] Ajit Jaokar, Tony Fish, Mobile Web 2.0, A book, 2007.

[41] Garfinkel, S., Rosenberg, B., RFID, Applications, Security and Privacy. Addison Wesley, Boston 2006.

[42] A. Jøsang, and al. Trust Requirements in Identity Management, AISW 2005.

[43] Wong, and al.: Polonius: an identity authentication system, Proceedings of the 1985 IEEE Symposium on security and Privacy.

[44] A. Jøsang, and G. Sanderud. Security in Mobile Communications :Challenges and Opportunities. Proccedings of the Australasian Information Security Workshop, 2003.

LSPEnv: Location-based Service Provider for Environmental Data

Katarzyna Wac and Lemonia Ragia

Abstract. This paper presents an approach for forecasting environmental data for location based services. The environment becomes a very important issue and especially people with health problems need more information and support in their daily life. In this work we propose a system for making predictions for spatial-temporal variables using the Bayesian Network method as a machine learning. To handle the missing values in our data we use the Structural Expectation Maximization Algorithm. The architecture of our system is based on a three-tier architecture which assists the distribution of the evaluation process. The case study is based on real environmental data from the Swiss national network. The provided data represent different types of location, e.g. rural, urban, etc. and are taken in different time. The results can be presented on a mobile device, in Internet and to any mobile user.

Keywords: location based services, machine learning, environmental data, prediction

1 Introduction

Location based services known as LBS are developed the last two decades because of new technology. They are widely used for advertising via cell phones, for sending information when somebody visits a new place, etc. Especially with the usage of GPS the position of a person is known and new challenges are open. LBS has proven to be a new growth business. Location-based services blend information about a person's location with other useful content, providing *relevant, timely and local information* to consumers *when* and *where* they need it.

There are a lot of efforts from vendors and governmental institutions to provide such services, to improve the existing systems or to use efficiently the technology. Lots of products appeared in the market. One can use the services of a product after installing it or after payment of a fee. For example, navigation software connected to mobile phones makes navigation easier for mobile people. GIS companies try to share their information via cell phones and they spend a lot of money to improve their software, e.g. using new technology for visualization.

High bandwidth in the mobile nets (UMTS, GPRS, HSCSD), personal digital assistants (PDA), wireless connection via WLAN and the accurate position location via GPS afford new opportunities for location based services. Analysts expect that there will be a huge impact on this kind of business and the services will become more easily available for every user. A very critical factor in mobile application is security. It gives assurance to the users to use the mobile application.

In this paper we present an approach for forecasting environmental data for LBS. It is widely accepted that there is a link between the state of the environment and human's health condition. Statistics show that many diseases can be caused by environmental pollutants [3].

Here is the description of two scenarios where people with health problems caused by the particular state of the environment can be helped in their daily life using our system.

Scenario 1: Asthma is a chronic disease in which a person experiences breathing problems. The occurrence of this disease is very much influenced by the air quality. Major factors influencing air pollution are related to transportation: gasoline and fuel fumes from cars, trains and planes. Industry also produces massive air pollutants. Smog is one of the results of this situation. Particularly, smog is produced by the existence of nitrogen dioxide in the air and occurs especially in big cities in the high traffic hours. Another factor influencing asthma is related to the house heating systems based on burning of fossil fuels. Concerning the state of the environment in a city, there are studies arguing the strong association of environmental causes of asthma and the health of inner city children [4]

Asthmatic people are interested to travel and visit new places. However, due to their condition, it would be of high importance for them to know in advance the forecasts for best and worst hours in a city center of the visited city, or generally to know which places in the city would be comfortable for them to visit in which hours and days. With such information at hand, asthmatics could be empowered to enhance their quality of experience and avoid health risks, by adapting their trip schedule in advance to their health condition and state of the environment.

Scenario 2: Allergies are chronic diseases in which a person reacts in a sensitive way to some substances in the air, food or water, having no comparable effect on the average individual. For air-related allergies, causes for such illness can be products of particular plants or other substances in the air. For example, the existence of pollen in an area with vegetation or some kind of dust at home, or air pollutants in the atmosphere can create allergic reactions. What is worse, some substances like tree pollen can be dispersed kilometres from the original area where the trees are located.

In general, allergic persons have to relocate themselves and avoid particular places in which the surrounding environment stimulates their disease. These people require (beforehand) detailed information about the surroundings environment in order to better schedule their trip and better manage their disease. Typically this information would be provided to them before they relocate, and it would contain a pollen forecast for a specific season in a given region or pollution and smog level in the city they want to visit.

In order to support these scenarios we propose in this paper the development of the Location-based Service Provider for Environmental data, denoted further on as LSPEnv. The LSPEnv provides forecasts for environment state in a given location and time over the wireless or wired access networks, i.e. to mobile or fixed service users. The information provided by the LSPEnv can be used further by people suffering from chronic diseases like asthma or allergies, to better manage the health risk taken while visiting unknown places.

2 Previous work

Location based services are “information services accessible with mobile devices through the mobile network and utilizing the ability to make use of the location of the mobile device” [5]. Similar definition is given also according to the international OpenGeospatial Consortium [6]. They are position specific information where the current position is given by a GPS. Is is also a communication means between different people when they share information. The messages are provided to the users by maps, or in textual form.

The domain of location based services is defined by the GSM Alliance Service Working Group [1] as: Asset management, Fleet management, Emergency Services, Person Tracking, Localized Advertising, Mobile Yellow Pages, Network Planning, Dynamic Network Control, Traffic Congestion Reporting, Routing to nearest enterprise, Roadside Assistance, Navigation, City Sightseeing.

Some examples are finding the next hospital or the next exit in a highway, or via SMS the most important tourist attraction in a city. Local authorities use LBS often for tourism purposes. In the beginning without GPS people could have information about a tourist attraction only when they are in a place which is covered by network based tracking. Now the tourists can inform others about interesting places or specific monument with their exact position. In traffic management it plays an important role to avoid accidents or traffic jam. For example, one driver can publish a geo-referenced message that informs about a street with heavy traffic. The same is for medical emergency information or for weather forecasting. In some cases, e.g. in the mountains, the skiers can get information for the current status of weather condition or special information about a ski region like number of mountains railways, ski rental etc. Regarding the location based services for environment there are only some concepts about web services which are related to spatial information exchange [1].

In location specific health information some examples are local disease rates including maps and guidelines, local health news, local weather, pollen and air quality alerts and maps (e.g. for allergic people), local health risks and hazards, addresses or local healthcare facilities, travelers’ health information, local drugs, etc. Similar questions that have to do with the physical situation of a patient can be answered in a system where medical sensors are also connected wireless [19]. Finally the concept of location has different meanings depending on the application domain. Therefore it is impossible to cover all the different application with a unique representation.

Machine learning techniques have been used for environmental data where data mining approach has been developed for the analysis and mapping of spatially distributed data [16]. They are also used for air quality assessment which is estimated as a classification problem of real time air pollutants data [17]. In another paper three algorithms of machine learning are used to predict the daily peak concentration of an air pollutant for air quality control [18].

3 Architecture Framework

3.1 Requirements

To concentrate the operational functions that are needed for effective support in making decisions we have to define some restrictions that have to be taken into account. We consider that not only professional users but all people can be potential users of our system. Based on the given scenarios we consider the following *functional requirements* for the LSPEnv:

a. Environmental state information gathering. This requirement has to do with the available data, which can be provided by different sources and in different time frames. This category includes following characteristics:
   - From environmental sensors
   - In different locations
   - Continuously in time, but at least 1 sample per hour

b. Environmental state information processing. The modelling of the data is a key issue for the further analysis of the data. It involves:
   - Transformation of sensor data into a standard format as used by LSPEnv
   - Location attribute processing along the GPS coordination
   - Date attribute processing, derivation of day of week
   - Variables processing and assignment of ranges corresponding to levels of influence of a given environmental state variable on a person's health.

c. Environmental state information forecast. The main aims of this requirement are:
   - Prediction engine running continuously over the collected data
   - Possibility to forecast for now or any given time in a future
   - Possibility of forecast estimation for locations, times for which the data does not exist
   - Possibility of forecast variables, for which the data does not exist, based on the existing data; however that would require incorporation of specialized knowledge by the LSPEnv

d. Environmental state forecasts dissemination. The user graphical interface has to be considered in order to provide clear and understandable information. The results can be:
   - web-based user interface where user inserts a query criteria: location(s), time (s), variables he/she is interested in
   - Available for use on the mobile phone.

The users in our scenario are mainly interested in using the LSPEnv service disseminating the forecasts (function 4 above). Nevertheless, the other three functions (1-3) serve as a basis for the dissemination function. Particularly, the forecast function is a core function of the LSPEnv service provider, and this function has the following functional requirements:

a. Incremental learning upon historical data
b. Forecasting upon incomplete or uncertain historical data
c. Forecasting upon repeating/conflicting historical data

Moreover, the forecast function has the following non-functional requirements:

a. Scalability with number of data items
b. Performance in terms of prediction speed and accuracy
c. Minimal service usage cost (in case if forecasting service is used by mobile users)
d. Ensuring user's anonymity

3.2 Design

The architecture of the LSPEnv is build upon the functional requirements for LSPEnv and hence is as proposed in figure 1. The environmental state information gathering function is responsible for acquiring the sensors readings that act as an input for a environmental state information processing function, which transforms the data for prediction engine in the environmental state information forecast function. Finally, the forecast function uses the dissemination function for user's query processing and predictions provisioning to them accessing this service from a fixed or mobile Internet nodes.

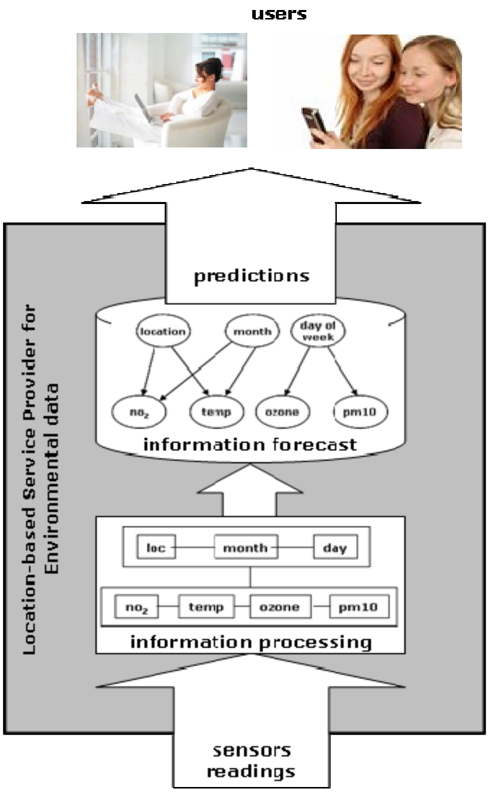


Figure 27 The general LSPEnv architecture

We propose a high-level architecture for the LSPEnv distributed system, based on a prototype application, e.g. for mobile users. The proposed architecture is very generic, since we do not want to limit ourselves to a given set of environmental data, or particular user-dependent technologies. The architecture is based on a widely-accepted distributed three-tier architecture, comprised of a client-, middle-, and resource-tier, that supports a distribution of computational tasks and separation of their concerns. Depending on the environment in which it the LSPEnv may operate, a number of specific requirements may exist, some of which may have a direct impact on the design of the LSPEnv architecture.

A *client-tier* may be composed of any kind of application; it can be application for end-users or application acting on behalf of end-users to retrieve the necessary information from the LSPEnv. The example of the latter one can be a user-agent application or any application that requires environmental information for processing purposes to be able to fine-tune the parameters of related processes e.g. car traffic system regulation in the city. Applications can run on standard PCs or on mobile phones and Personal Digital Assistants (PDAs). The LSPEnv system may have a number of heterogeneous clients, thus, it is crucial that environmental information is accessible over one or several standardized interfaces.

The *middle-tier*, sometimes also called 'business-tier' or 'enterprise-tier', encapsulates all the logic of the LSPEnv system, and particularly the information forecast engine. This tier holds all the logic of the LSPEnv system, it acts as a layer between the client-tier that requests for environmental forecasts and the resource tier that holds the raw data. The forecast engine is accessed over a standardized interface, such as a web service. For example, when a client application requests for information on a particular environmental variable, the engine fetches the required metadata from the metadata repository, sends queries to the data sources specified for the variable, calculates the forecasted value and send the results back to the client.

The forecast engine handles all the tasks related to environmental variables forecast. The metadata repository holds all the data and metadata necessary for the forecast engine to operate. Particularly, it stores metadata such as environmental variables specifications. A rule engine allows handling rules associated to the forecast process. Moreover, the rule engine is a component that can monitor critical events and variables values, deliver alerts to users, and initiate other system actions if the LSPEnv is a part of a bigger system. For example, if the value of a particular environmental variable deviates from a predefined threshold, the rule engine may notify interested users automatically via e-mail or SMS, or may trigger a particular system action. The metadata repository and rule engine, when implemented, can be very specific for a given user/application.

Finally, the *resource-tier*, also called 'database-tier' or 'back-end', is composed of any system capable of providing environmental data to forecast engine. This data can be acquired directly from sensory readings, out of a system processing environmental data, out of (internal or external) historical databases or via invocation of specific remote procedures (Remote Method Invocation or Wireless Services) on a given environmental data source to obtain an environmental specific data.

4 The algorithm from ML

Due to the given functional requirements posed on the forecast function of the LSPEnv, we have chosen the Bayesian Networks (BN) as a machine learning method [7]. Bayesian Networks are useful models in representing and learning complex stochastic relationships between interacting variables and their probabilistic nature is capable of modeling the noise and handle missing values, as inherent in the environmental data. Moreover, that method allows for combining domain knowledge and historical data.

A Bayesian Network is a Directed Acyclic Graph (DAG) that consists of a) the structure or the directed edges that encode the causal relations and conditional independencies between the (mutually exclusive, collectively exhaustive) variables, b) the local parameters or the distribution function and parameters that encode the distribution of a child variable given its parents (CPDs). Bayesian Networks can include continuous and discrete variables [8]. Due to the nature of our problem, we have focused only on the discrete variables, with values in a finite value set. Therefore, for a given example parent B and child A, we denote their relation as:

$$P(A/B) = [P(B/A) * P(A)] / (P(B))$$

Where:

P(A) – the prior probability of A

P(B/A) – the conditional probability of B given A; also called the likelihood function

P(B) – the marginal probability of B

P(A/B) – the posterior probability of A given B

In our learning task, firstly we have focused on the inference of the graph structure from the data. A significant challenge to this task poses the fact that our data has missing values (at random). To tackle BN learning with missing values we use the Structural Expectation Maximization (SEM) Algorithm [9]. SEM searches in the joint space of (graph structure x parameters). It starts with a random structure and its parameters and estimates the probability distribution of missing variables with the SEM algorithm. Then it computes the expected score for each graph of the neighborhood and chooses the one which maximizes the score. It is presented in the following pseudo-code:

```
Loop for n=0,… until convergence
    compute the posterior P(A/B)
        E-step: for each $graph_n$, derive missing values for variables,
                based on knowledge on their distribution, then
                compute an expected score for this graph (via sum of
                log of its prior P(A) and log of its marginal probability
                P(B/A))
        M-step:choose neighborhood $graph_{n+1}$ that, for given values of
                variables maximizes the score of the $graph_n$
        if expected score of $graph_n$ == expected score of $graph_{n+1}$
      return $graph_n$
```

After several updates to the graph structure, we run SEM again to recalculate the missing values for the final graph. This final graph, with the assigned missing values is then used for forecasting service.

The used Bayesian Network has the following structure derived from data based on the SEM algorithm (1) implemented in Matlab Bayes Net Toolkit [10]

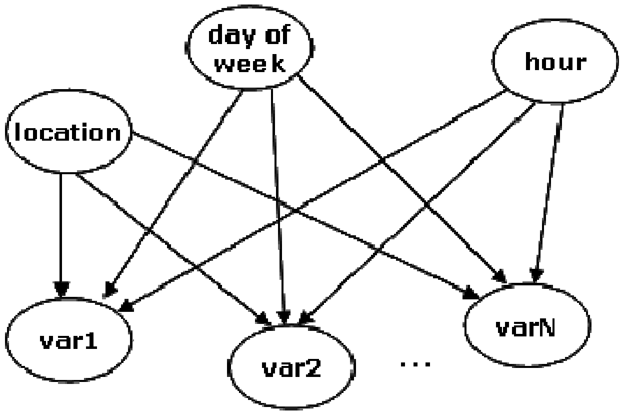


Fig. 2. Bayesian Network Graph Structure

5 Evaluation

The data are provided by the Swiss national network [12] which is representative for the whole country. The measurement stations represents the following locations:

a) urban with very heavy traffic,
b) urban areas with a population more than 100000 people
c) suburban areas
d) rural next to the highway
e) rural with an altitude higher than 1000 m a.s.l.
f) rural with an altitude lower than 1000 m a.s.l. and
g) areas in the high mountains

We have obtained data for 16 different locations in different time frames over Switzerland for temperature, ozone (O_3), nitrogen dioxide (NO_2) and particulate matter (PM10) as environmental variables. Among the air pollutants we choose some of the most important: the ozone, the nitrogen dioxide and the particulate matter because these are high associated with the human health [16]. The ozone is undoubtedly one of the basic causes for many health problems, it has short and long term effects on human health and plays an important role in the mortality rate [13]. Scientific results show that the nitrogen dioxide is a significant factor for the increase of a lot of significant allergic illnesses [15]. The particulate matter is a mixture of organic and inorganic substances and it causes the air pollution. There is scientific work that show the high responsibility of the particulate matter for health effects [14].

For ozone, the limit value is 120 (μg/m^3) as a daily maximum 8-hour mean, above which a person health may get affected. For nitrogen dioxide this limit is 80 μg/m^3 1-hour mean, while for particulate matter its is 50 μg/m^3. Temperature does not have critical levels of values in the range which have been measured in Switzerland.

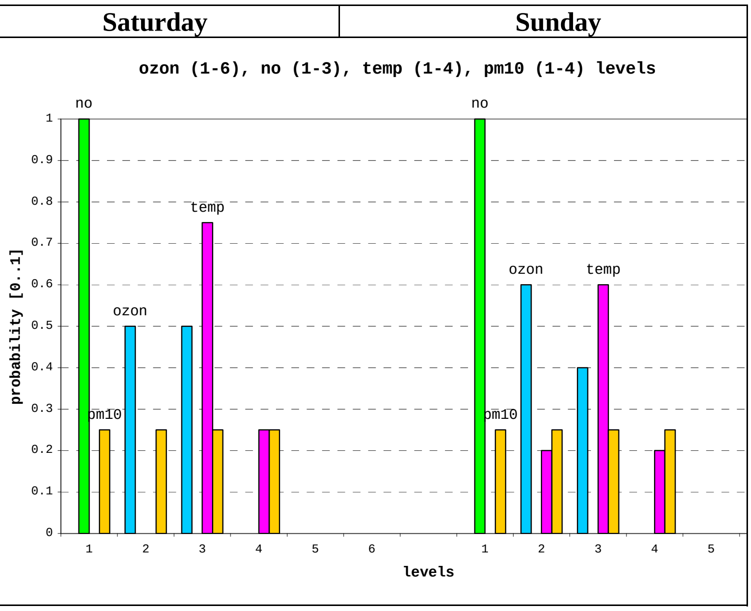
Tab. 1. Values and levels of different environmental state variables

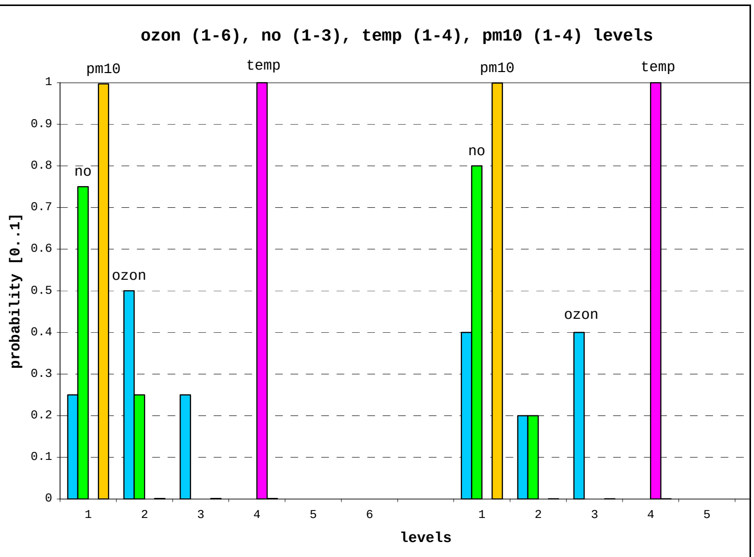
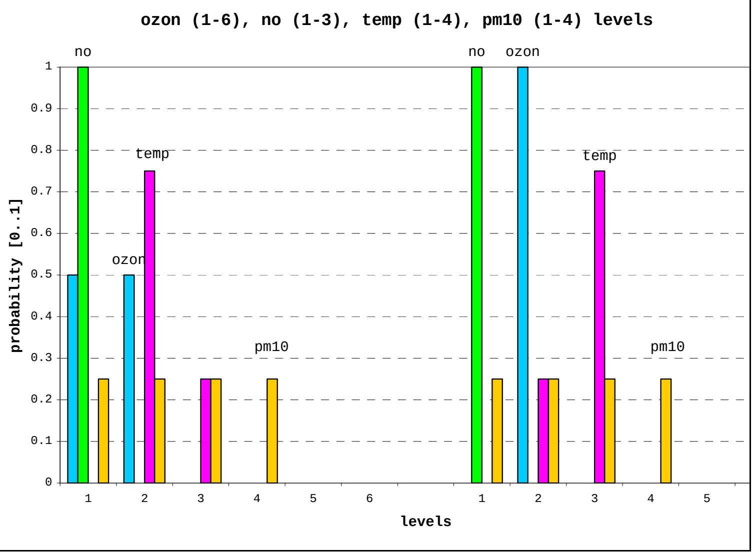
	min	levels of values					max
level	**1**	**2**	**3**	**4**	**5**	**6**	
ozon	2	70	100	120	160	240	270
No$_2$	1	40	80	200			
temp	1	7	12	17	27		
pm10	1	50	75	100	116		

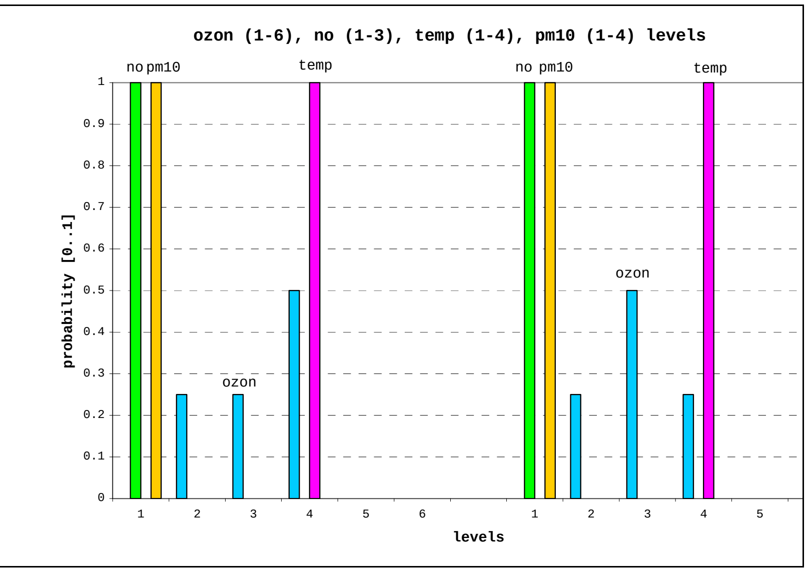
5.1 Evaluation scenario

For the collected data we have evaluated scenario in which an asthmatic user sends the query to the system regarding the environmental state variables in two different cities in Switzerland (5 and 6) for a weekend (Saturday-Sunday) in summer months of July and August. The user plans a weekend trip to one of these cities, and may make a decision based on prediction for environmental state. Prediction results are presented in table 2.

Tab. 2. Probability distribution for different levels of environmental state variables for two different cities for a summer weekend.

		Saturday	**Sunday**
July	**Location 5**		

<table>
<tr>
<td></td>
<td>Location 6</td>
<td>
</td>
</tr>
<tr>
<td>August</td>
<td>Location 5</td>
<td>
</td>
</tr>
</table>

	Location 6	

As we could see from the prediction, the location 6 has higher temperature and level of ozone and nitrogen dioxide (only in July) higher than location 5, hence the person may make a decision upon visiting location 5 rather than 6.

5.2 A prototype LSPEnv application – technical aspects

In order to be able to evaluate the LSPEnv system, at least on a small scale, we propose a prototype application which builds upon the proposed system architecture (Figure 2). The prototype application builds upon a simplified version of our architecture which features only its core elements for the storage of data for and prediction of limited number of environmental variables. Other elements, as those related to metadata repository or enhanced rule engine, have been discarded for the sake of simplicity.

From a technical point of view, our prototype application is based on the architecture proposal presented in previous section (Sec. 3.2). It is composed of a client application that runs on a mobile phone, a environmental variables prediction engine, and a simple database system which holds the raw data for measurements. The use cases for the prototype system are as follows:

1. store set of historical data for set of environmental variables,
2. list available set of environmental variables,
3. provide predictions for given location, month and day of week.

Our prototype application can be based on a distributed, three-tiered architecture, as presented in Figure 3. The client-tier component can run on a mobile phone. The forecast engine of the middle-tier can run on a Java EE platform and can use a database system as metadata repository. Finally, the resource-tier is represented by a database system. A rule engine can be part of our prototype application, but it not yet implemented. Figure 3 shows the overall architecture of the prototype application.

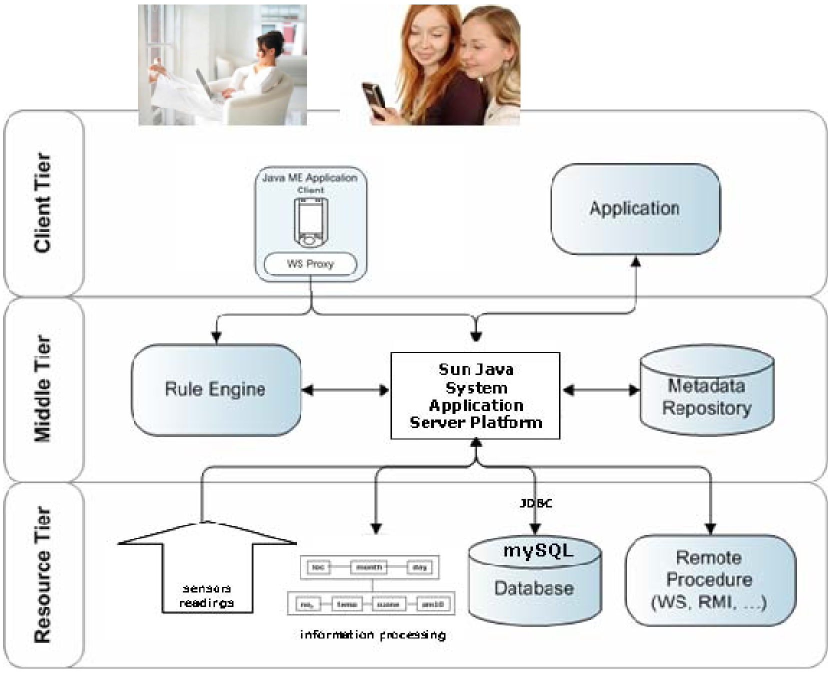


Fig. 3. The implemented LSPEnv system architecture

The application client runs on a Java Platform, Micro Edition (Java ME) which provides an environment for applications running on mobile devices. It allows to display a prediction results to a mobile user, where the screen can be updated automatically every few seconds by a timer task or upon the notification of the prediction change triggered from the LSPEnv system. Furthermore, the application client should allow to list all available environmental variables, and to display detailed information on each one of them on a mobile device. The example prediction for environmental variables can be displayed to the mobile user in a simple graph form, as shown for an example query in Table 2.

Since nearly all information displayed by the application client is provided by the middle-tier, an application may requests data whenever needed. For this purpose, the client can call the web service interface of the middle-tier component. Web services are use open, XML-based standards and transport protocols to exchange data with clients. The Simple Object Access Protocol (SOAP) can be used to exchange data. SOAP defines the envelope structure, encoding rules, and conventions for representing web service invocations and responses. The requests and responses are transmitted over the Hypertext Transfer Protocol (HTTP). The WS Proxy, which represents the remote web service on the application client, is generated by the Java API for XML Web Services (JAX-WS), based on the Web Services Description Language (WSDL) file of the web service. Whenever the application client requires data, it simply invokes the methods on the WS Proxy.

The LSPEnv prediction engine can run on a Sun Java System Application Server Platform, a compliant implementation of the Java EE 5 platform. The prediction algorithms is implemented in Matlab, as indicated in previous sections, and can be called upon the client's prediction request. The metadata repository could be implemented in a simple file-system holding specifications of environmental variables.

Finally, the resource-tier can be represented by a standard mySQL database which stores all the 'raw' data. It is on this database that the prediction engine executes the queries that are specified by the variables stored in the metadata repository. The connection between the prediction engine and the database is established through the Java Database Connectivity (JDBC), which provides methods for querying and updating data in a database.

6 Conclusions and discussion

We propose an open system for forecasting environmenatal issues by integrating different spatial-temporal data and representing them in a graphical user interface. We presented an approach for location based services that takes into account the air quality monitoring and supports people with health problems. The approach is based on the Bayesian Networks method as the prediction function and the Structural Expectation Maximization Algorithm is included to avoid problems with missing values in raw data. We use the three-tier architecture, composed of a client-, middle-, and resource-tier.

We have used real environmental data with four variables from Switzerland which represent lot of different types of the locations. The results are the predictions for these variables for a specific date, or month in the future. The results can be shown in Internet via wireless connection or in a PDA in simple graph form.

We are investigating more functionalities for data retrieval and the visualization of the results. We try to use methods which combine textual and map information for better understanding of the results.

References

[1] GSM The GSM Alliance Services Working Group http://www.gsmworld.com

[2] Ragia L., El Isbihani A., Kiehle C., 2006: Web Service for Groundwater Vulnerability. International Conference, Protection and Restoration of the Environment, July 3-7, Chania, Greece.

[3] European Commission, 2002: "Health statistics - Key data on health 2002- Data 1970 - 2001", ISBN 92-894-3730-8

[4] Mortimer K. M., Tager I. B., Dockery D. W., Neas L. M., and Redline S. 2000: The Effect of Ozone on Inner-City Children with Asthma. In American Journal of Respiratory and Critical Care Medicine, Vol. 162, No. 5, pp. 1838-1845.

[5] Virrantaus K., Markkula J., Garmash A., Terziyan Y. V., 2001 : Developing GIS-Supported Location-Based Services. In: Proc. of WSIS'2001-First International Workshop on Web Geographical Information Systems, Kyoto, Japan, pp. 423-432.

[6] Open Geospatial Consortium (OGS), 2005. Open Location Services.

[7] Charniak, E. (1991). Bayesian networks without tears. *AI Magazine,* 1991. **12**(4): p. 50-63.

[8] Heckerman, D. (1995). A Tutorial on Learning with Bayesian Networks

[9] [Friedman, N. (1998) The Bayesian structural EM algorithm, in G. F. Cooper & S. Moral, eds., Proc. *14th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA.

[10] Murphy, K. (2001). The Bayes Net Toolbox for Matlab. *Computing Science and Statistics*, **33**(1)

[11] Murphy, K. (2001). The Bayes Net Toolbox for Matlab. *Computing Science and Statistics*, **33**(1))

[12] Federal Office for the Environment FOEN, Department of the Environment, Transport, Energy and Communication http://www.bafu.admin.ch/index.html?lang=en

[13] [European Commission (2002): Health statistics Key data on health 2002.

[14] [Dingenen R. V. et al. (2004): European Aerosol Phenomenology - 1 : physical characteristics of particulate matter at kerbside, urban, rural and background sites in Europe in *Atmospheric Environment* Vol. 38, Issue 16, pp. 2561-2577

[15] Jenkins H. S. et al. (1999) The Effect of Exposure to Ozone and Nitrogen Dioxide on the Airway Response of Atomic Asthmatics to Inhaled Allergen. In *American Journal of Respiratory and Critical Care Medicine*, Vol. 160, No. 1, pp. 33-39.

[16] Kanevski M. et al. (2004): Environmental data mining and modeling based on machine learning algorithms and geostatistics. In Environmental Modelling and Software, Vol. 19, Issue 9, pp. 845-855.

[17] Athanasiadis I. N. et al. (2003) Applying Machine Learning Techniques on Air Quality Data for Real-Time Decision Support. In First International Symposium on Information Technologies in Environmental Engineering, Gdansk, Poland, ICSC-NAISO Publishers.

[18] Kalapanidas E. and Avouris N. (1999) Applying Machine Learning Techniques in Air Quality Prediction. In Proc. Advanced Course on Artificial Intelligence, ACAI ´99, Chania, pp.58-64.

[19] Van Halteren A. et al. (2004) Mobile Patient Monitoring: The MobiHealth System. The Journal on Information Technology in Healthcare Vol. 2, Issue 5, pp. 365-373.

Context-Aware Middleware Architecture for Vertical Handover Support to Multi-homed Nomadic Mobile Services[1]

Pravin Pawar, Pierre Maret, Katarzyna Wac, Aart van Halteren, Bert-Jan van Beijnum, Hermie Hermens

Abstract. To accommodate the requirements such as high usability and personalization of 4G (mobile) networks, conventional handheld single network-interface mobile devices are evolving into multi-homed devices. Moreover, owing to the recent advances in the mobile middleware technologies, hardware technologies and association with the human user, handheld mobile devices are evolving into data producers and in turn acting as Nomadic Mobile Service (NMS) providers. For these devices, a vertical handover support is essential for the improved and reliable NMS delivery. Also, the fulfillment of the required QoS by the NMS is bounded by the *end-to-end QoS* (e2eQoS) provided by the underlying heterogeneous networks. To deal with these aspects, we propose a context-aware middleware architecture supporting vertical handover for the NMSes hosted on the handheld mobile devices. We emphasize the following features of the proposed middleware: 1) Context-aware computing based approach which uses an extensive set of context information collected from the mobile device and a fixed network; 2) Provisioning of and interaction with the end-to-end QoS (e2eQoS) predictions context source in the fixed network to obtain near-accurate estimation of the e2eQoS at a certain geographic location and to reduce unnecessary power usage in searching for available networks.

Keywords: Nomadic mobile services, multi-homing, vertical handover, M-health services, context-aware computing, end-to-end QoS.

1 Introduction

In the vision of 4G (mobile) networks, integrated and personalized services are envisaged at any desired time and any location. This vision is, partly, to be realized by making use of mobile, often handheld or otherwise wearable devices, connected to the Internet using one or more of the network interfaces embedded in these devices [1]. Nowadays these mobile devices are evolving into *multi-homed* devices, which are able to connect to the Internet using multiple network interfaces e.g. WLAN, GPRS and UMTS. The forthcoming multi-homing enhancements include the device's ability to simultaneously use multiple networks of the same technology type (e.g. multiple WLANs) using a single network interface [2]. A *vertical handover* (handover between different network technologies) is an adaptation method for multi-homed devices to dynamically redirect the data

[1] This work is supported by Dutch Freeband Awareness project (under grant BSIK5902390), Amigo project (IST-004182, partially funded by the European Commission) and Mobility International region of Rhone Alps (MIRA), France.

communication path of the mobile application to the networking interface different than the currently used one [3].

Until recently, research in the mobile computing community has been focused on use of mobile devices as *data consumers*. However, due to the advances in hardware (e.g. integrated camera, GPS) and software middleware technologies [4], and arising business needs for new services types, the mobile devices are turning also into *data producers*. In such a case, because of the inverted *producer-consumer* roles, the throughput achieved over the uplink by the application producing data is critical and the quality of this data is influenced by the characteristics of the available wireless networks as well as by resources of a mobile device (e.g. CPU clock speed). Moreover, the data producer application has nomadic characteristics because it roams with the device on which it is hosted. In [5], van Halteren et al. extend the concept of Service Oriented Architecture (SOA) to the handheld mobile devices, to realize these applications as *Nomadic Mobile Services* (NMSs) and make these services available to clients located anywhere on the Internet using the *Mobile Service Platform* (MSP) middleware. One of the applications of the nomadic mobile services is a *remote tele-monitoring service* [5] in an (mobile) m-Health domain. Using a *remote tele-monitoring service*, patient's mobile device acquires the vital signs data from the sensors attached to the patient's body, (pre-)processes the data locally at the mobile device, and sends the data to an m-Health portal. At the portal, the data is made available as services that can be used for any desired purpose, e.g. real-time retrieval by a qualified health professional.

There exist a variety of architectures, algorithms and schemes for vertical handover on multi-homed mobile devices. For example [3], [6] and [7] propose context-aware computing-based architectures, [8] and [9] propose to use policy-based approaches, while [10] introduces a generic vertical handover decision function. These approaches are mainly targeted towards providing vertical handover support for the *data consumer* mobile devices at the network layer level (e.g. using Mobile IP).

We consider herewith the problem of providing vertical handover support based on the principles of context-aware computing [11] for the NMS. The work reported in [12] illustrates design, validation and performance evaluation of the context-aware MSP which simply uses the information about the networks the mobile device is connected to, and always selects the network with the highest theoretical throughput to handover to (e.g. always prefers WLAN over GPRS). In [12], only one context source, particularly, the *Communication Context Source (CCS)* provides the knowledge on current state of network resources, which is further processed by the *context processor* and used by the *context reasoner* for decision making. It is important to notice that the CCS does not distinguish between network's uplink and downlink direction. It provides information about network's maximum downlink throughput, and not the uplink, which would be more appropriate for NMS.

This paper extends the work reported in [12] in the following three ways. Firstly, for a handover-decision, besides the theoretical throughput of available networks, the MSP considers other relevant context information acquired directly from the context sources at the mobile device and provided by the context sources placed in the fixed network.

Secondly, we consider the NMS Quality of Service (QoS) requirements in terms of its required uplink throughput and delay. The fulfillment of the NMS's requirements is bounded by the *end-to-end QoS* (e2eQoS) provided by the underlying heterogeneous networks [13]. In most of the cases, the first hop in that path is a wireless (mobile) network, which is a bottleneck in the end-to-end path. Hence the choice of a mobile network provider and network type (e.g. GPRS or UMTS), are critical for a NMS delivery. To deal with that, we propose the *QoS predictions context source*, which is based in the fixed network and provides the following information set to the NMS: a) information about the mobile networks available to a mobile device at a given geographical location and time and b) predicted e2eQoS provided by a particular mobile network. The existence of the QoS predictions context source empowers the mobile device in terms of networks choice at a given location/time, but also along the device's mobility path, as we explain further in this paper. The provisioning of QoS predictions context source also help to overcome the limits of the battery power of a mobile device. Usually, it is required to keep all the network interfaces on the mobile device *always switched on* to search for the mobile networks available in the vicinity. If, as we explained above, network availability is a priori known, then it suffices to switch on the required interface at a given location and time, and keep it off when not used.

Thirdly, once the *context processor* component obtains all the required context information alongwith the e2eQoS predictions, the *context reasoner* component applies a utility function to assign a score to each available network, and further decides whether the handover is necessary.

In summary, the distinguishing aspects of our proposed middleware architecture supporting the vertical handovers decision for multi-homed NMS providers are as follows: 1) The use of an extensive set of context information collected from the mobile device and from a fixed network; 2) Provisioning of and interaction with the QoS predictions context source in the fixed network to obtain a near-accurate estimation of the e2eQoS at a certain geographic location at a given time, reducing unnecessary power usage in searching for the available networks; 3) A utility function-based vertical handover decision mechanism.

The reminder of this paper is organized as follows: Section 2 of the paper discusses the related work. Section 3 presents in details the proposed architecture, its elements and the interactions between them. Section 4 provides the information about the ongoing implementation. Section 5 concludes our finding and provides future work areas.

2 Related Work

In literature a variety of architectures, algorithms and schemes for vertical handover support for the multi-homed mobile devices have been reported. For example, [3], [6] and [7] use context-aware computing-based architectures with the analysis of the contributing factors, vertical handover decision models and experimentation details. While approaches proposed in [3] and [6] use vertical handover decision function based on *Analytical Hierarchy Process* (AHP), [7] proposes use of multi-network optimization protocol after eliminating networks which do not satisfy certain constraints. An anticipated vertical handover scheme based on Fast MIPv6 (FMIPv6) procedure for handover between WLAN and UMTS networks has been proposed in ([14]). In this scheme, the mobile device initiates the registration request while still receiving packets on the old link and after receiving the *Care of Address* (CoA) it starts receiving the packets on the new link. Moreover, a generic vertical handover decision function which uses the weights assigned by the mobile device to different factors affecting handover decision has been proposed in [10]. [2] proposes a handover strategy for streamed video based on the jitter experienced by the mobile device when it is close to the edge of an access network. [15] compares the performance of four vertical handover decision algorithms for the applications which fall into the following four classes: conversational, streaming, interactive and background. The comparison criteria include average bandwidth, delay, jitter and BER of the selected network by these algorithms.

There exist a number of vertical handover approaches based on policies. [8] presents PROTON, a policy based architecture along with context management components. The policy based solution proposed in [9] specifically emphasizes on choosing the correct time and selecting the correct network for the vertical handover. [16] studies in detail the effect of a vertical handover policy on the performance of Internet applications. The simulation results reported in [16] show that the throughput and RTT of the network have a low influence on the overall performance of short-lived TCP connections and user-interactive sessions (e.g. telnet), while delay of the selected network has impact on the CBR traffic over UDP.

Some handover schemes give a special emphasis on handling the resource limitations of the mobile device, user mobility and connectivity. For example, [17] proposes a prediction algorithm for vertical handover by using the speed of the mobile device and mobility patterns. The mobile device compares the predicted time to stay in the network with the handover delay and accordingly takes a handover decision. [18] presents a vertical handover policy based on the remaining battery status. If this value is higher than some specified threshold value, a greedy approach is chosen to obtain a higher throughput; otherwise a conservative approach is chosen to restrict the number of active network interfaces. [19] proposes to use a *Location Service Server* (LSS) in the fixed network which provides information about the coverage area, bandwidth and latency of wireless networks available to a mobile device. The information obtained from LSS is used to switch on the network interface only when the mobile device is in the network coverage area, thus resulting in the power savings. However, this research does not consider e2eQoS provisions.

[20] studies the required buffer size to achieve lossless upward (in terms of network coverage area, e.g. from WLAN to GPRS) vertical handover for the data traffic including multimedia as multimedia applications also generate data at the constant rate. The simulation concludes that with reasonable buffer size (1-20 Kbps) the average packet delays are within acceptable QoS limits even for multimedia traffic. To overcome the problem of dealing with multiple IP addresses

during vertical handover, [21] proposes to use IP tunneling which uses two pairs of the virtual/fixed IP addresses, one pair for the mobile device and one pair for the handover server.

In order to make a vertical handover decision, researchers suggest using a variety of (context) information on the mobile device and in the network. The context information available on the mobile device could be generally classified as user preferences for the network and application, application QoS requirements, experienced network QoS, available network interfaces, reachable networks, *Received Signal Strength Indication* (RSSI), mobility information and device resource usage. On the other hand, the context information available in the (fixed) network includes a service provider's profile, network QoS, network coverage and location of access points.

Since our approach specifically targetS data producing mobile devices hosting NMSs, we do not choose to use one particular approach as reviewed in the related work. However, the proposed architecture customizes the following concepts from the related work: 1) context requirements, 2) AHP based decision making function, 3) a part of e2eQoS prediction context source concept, as similar to LSS proposed in [19].

3 Context Aware Vertical handover Architecture

In this section firstly we describe MSP in brief, secondly we elicit necessary context information contributing to the handover decision algorithm, and thirdly we present the proposed architecture and its components.

3.1 Introduction to Mobile Service Platform

A NMS realized using MSP consists of two components: 1) An application realizing a service running on the mobile device (referred to as a *device service*); and 2) a representation of the device service in the fixed network which is referred to as a *surrogate.* The surrogate functions as a proxy for the device service and participates in the service discovery network. A *Surrogate Host* is responsible for the management of surrogates. The main components of MSP include the following: 1) *MSP-Input/Output (MSP-IO)* resides on a mobile device and interacts with the device service. 2) *MSP-Interconnect* located at the surrogate host and interacts with the surrogate. 3) *MSP-Messages* specifies the structure of messages exchanged between the device service and the surrogate. MSP uses HTTP as a data transfer protocol. The type of interactions between the device service and surrogate are: 1) *One-Way messaging* for unacknowledged message delivery; 2) *Request-Response messaging* for reliable message delivery; and 3) *Streaming* for exchange of continuous data (streams). MSP uses dedicated control plane interactions for control, monitor and lifecycle management of the NMS.

On the instantiation of the first device service on a given mobile device, MSP-IO creates a *Device HTTP Connection* to handle all the interactions from the mobile device. After initialization of the *surrogate* in the surrogate host, MSP-IO creates a *surrogate connection* on top of the *Device HTTP Connection* and provides its handle to the corresponding device service. The *surrogate connection* is later used by the device service to exchange messages with its surrogate. The *message worker* and *stream worker* components in MSP-IO are responsible for the transmission of messages and streams respectively received from the device service. MSP design is based on *Jini* technology. The communication between the device service and surrogate is specified by the Interconnect protocol in *Jini Surrogate Architecture Specification* [22]. We have developed an *HTTP* implementation (referred to as *HTTPInterconnect*) of the Interconnect protocol. The device service is usually implemented using J2ME technology. For information on the architecture, design choices and implementation of MSP, we refer to [5].

4 Architecture and Components Description

The proposed context-aware computing-based architecture consists of a number of *Context Sources* (CS) on the mobile device and in the fixed network, and *Context Processor* (CP) and *Context Reasoner* (CR) components on the mobile device. Traditionally, the handover decision could be taken entirely by the network, for the so-called Network Controlled Handoff (NCHO) or entirely by the mobile device, for the so-called Mobile Controlled Handoff (MCHO) [7]. There also exist some approaches which combine both of these strategies e.g. [23] and [24]. The later is more useful in case of Mobile IP where appropriate communication is required between the Home Agent and new Foreign Agent. We choose to take handover decision on the mobile device because most of the context sources are located on the mobile device and we do not use Mobile IP for handling mobility. Moreover, it supports autonomy of the mobile device such that it does not need to depend on the external entities for the handover decision.

4.1 Context Sources

Traditionally, the decision of a vertical handover to a new mobile network is based on the *Received Signal Strength Indication* (RSSI) of the (to be) connected mobile network. However, RSSI does not exhibit the network conditions adequately [19], and moreover it does not reflect the important NMS-level objective: the *end-to-end QoS* (e2eQoS) experienced by a NMS when a given mobile network is chosen. Moreover, RSSI is local information (on the wireless access link only), whereas we aim to select the mobile network that meets the e2eQoS requirements of the NMS. In our case, the e2eQoS particularly encompasses the NMS-level throughput (in Kbps) and delay (in milliseconds) of the underlying data communication path between the device service and its surrogate placed somewhere in the Internet. In order to consider this objective, we suggest using a variety of context information available on the mobile device and in the fixed network.

Table 1 shows the CSs on the mobile device, context information provided by them, context description and units, motivation behind its selection and the CS interfaces, over which context can be obtained. Similarly, Table 2 presents the context information to be collected from the fixed network. In both tables, the interface methods starting with `get` provide context information only once upon the request, while those starting with `subscribe` continuously provide the context changes to the subscriber. For the sake of brevity, in this paper we do not provide the detailed design of the CS. However, since the architecture of the QoS prediction context source is not obvious from the information provided in Table 2, we choose to provide its brief overview in the further section. Our proposal of the chosen context information as presented in Tables, is based on the earlier experience with the remote tele-monitoring service using Context-Aware MSP [12], current experience with the design and development of the QoS context source and also the literature reviewed in the related work.

Table 1: Context sources on the mobile device

CONTEXT SRC.	*CONTEXT INFORMATION*	*MOTIVATION*	*INTERFACES*

Location And Time Context Source	Coordinates of the device's current geographic location (`longitude, latitude`) and time (`Date, HH:MM:SS`) as obtained from the GPS receiver.	It has been observed that the availability of and the e2eQoS provided by the mobile networks to NMS depends on the location and time.	`getCurrentLocationAndTime();`
			`subscribeLocationAndTimeChanges();`

Device Context Source	For a given mobile device, its model (`String`), CPU type (`String`), CPU clock speed (`MHz`), remaining battery level (`%`) can be obtained using the OS API calls.
	Standard values of power-consumption per network interface (`milliAmps`) obtained from the device specifications.
		For a given device, because of the processing power required for creating and transmitting NMS data, CPU type and clock speed influence the e2eQoS of the (currently used) network.
		Battery level could be combined with the user's power saving preferences. A decision about whether to actively search available networks or not, could also be taken based on the remaining battery power [19]. In [25] we observe that memory and CPU usage are well below limits for MSP.
			`getDeviceType();`
			`getCPUInfo();`
			`subscribeBatteryLevel();`
			`getInterfacePowerRequirements()`

User Preferences Context Source	A user's ranked list of all the mobile network providers, network names, and network technologies a user is subscribed to (`List [String, String, String]`).
	A list of all the device services (`List [String]`) ranked according to their importance to the user.
	User's power preference (`Yes/No`) indicating whether the
		It may not be always the case that user's network preferences are based on only one factor such as usage cost. E.g. a businessman may rank the networks based on the security. Hence it is best to leave the ranking decision to the user.
		A ranked list of services could be used to provide preferential treatment to the services with high rank.
		It is very likely that the interface using more power provides higher e2eQoS. Also,
			`subscribeNetworkPreferences();`
			`subscribeServicePreferences();`
			`subscribePowerPreferences()`

	middleware should consider/or not interface's power usage during selection. All the above information obtained from user using the user interface.	some users always keep their mobile device charged while others do it when needed. Hence power preference context information is selected.	
Communication Context Source	A list of mobile networks along with provider names, technologies, theoretical uplink throughput and delay (`Network Cross Layer Info. in XML)` in the surroundings of a mobile device at a given time and location. We refer to [26] for details.	In case of i) unavailability of the predictions from the QoS Predictions CS or ii) an unpredicted loss of connectivity and lack of further QoS predictions, the information provided by a Communication CS could be used to make a (rough) handover decision to a new mobile network.	`subscribeNetwork CrossLayerInfo()`

Device Service Context Source	Required e2eQoS of every running device service.
	Device service's perception (%) representing its satisfaction for the provided e2eQoS by the currently selected network.
	Criticality alarm of the device service (`Yes/No`) representing its current importance. It depends on the situation in which service is running.
		As observed in [25], each device service has different e2eQoS requirements.
		Device service's perception could be used to validate if the selected network satisfies device service requirements.
		Device Service criticality level can be assigned a higher value e.g. remote health tele-monitoring service in case of emergency health situations (e.g. high probability of seizure) and it can be changed to a lower value in case of non-emergency situation.
			`subscribeDeviceServiceRequirements()`
			`subscribeDeviceServicePerception();`
			`subscribeDeviceServiceCriticality()`

User Trip Information Context Source	User's trip information: in terms of source location (`Location`) and time (`Date, HH:MM:SS`), destination location (`Location`) and estimated time of arrival (`Date, HH:MM:SS`), and transportation mode (`String`) obtained using user i/f.	This information is useful for the QoS prediction CS to calculate the co-ordinates along the user travel path, estimated arrival time and use it further for the prediction of available mobile networks and associated predicted e2eQoS along the user mobility path with certain deviations.	`subscribeTripLocations();`
			`subscribeTransportMode()`

Table 2: Context sources in the fixed network

CONTEXT SOURCE	*CONTEXT INFORMATION DESCRIPTION*	*MOTIVATION*	

Surrogate Host Context Source	Location (`longitude, latitude`), time (`Date, HH:MM:SS`) at which the current mobile network is selected and mobile network	The predictions provided by the QoS context source are based on the historical data combined with learning mechanisms. The context information obtained from	`pushNetworkSelectionInformation();`
			`pushDeviceServiceRequi`

	provider, network name, and technology (`String, String, String`).
	A model (`String`), CPU type (`String`), CPU clock speed (`MHz`) of the mobile device on which the device service is running.
	Device service's e2eQoS requirements and observed e2eQoS as observed at its Surrogate.
		the surrogate host CS is input to the learning of QoS prediction CS.
		However, for learning, the QoS predictions CS does not only depend the context information provided by the surrogate object context source, but also use a variety of other context sources which are out of scope of the reported research. We refer to Section 3.2.1.1 for more details.
			`rements();`
			`pushObservedE2EQoS()`
			(In contrast to `get` and `subscribe` methods, the methods starting with `push` proactively sends context information to QoS predictions CS)

Qos Predictions Context Source	All available mobile networks as specified by provider names, network names and technologies along with their coverage ranges and availability at a given location/time and predicted e2eQoS provisions (in an XML structure similar to Network Cross Layer Info.)	This information is useful to know the possible networks to handover to along the user's mobility path.	`getNetworkPredictionInfo()`

QoS Predictions Context Source

The QoS predictions CS seeks to provide an efficient and accurate method that generates precise e2eQoS forecasts for a broad range of mobile networks at the given geographic location and time. The core of the prediction method is based on a *Dynamic Bayesian Networks* (DBN) [27] model. The choice of DBN has been motivated by the set of requirements for a machine learning task used in our prediction, such as: a) ability to predict, b) incremental learning, c) learning with incomplete, uncertain, redundant and conflicting data, d) learning about causalities combining the domain knowledge and (historical) cases data set, e) prediction with missing values in queries, f) scalability and performance. The DBN based learning model employed in the QoS predictions CS uses valid historical data for learning. One of the real-time ways to obtain this data is to get it from the MSP via Surrogate Host CS. Towards this end, currently we work on calibration and (off-line) evaluation of the prediction method in an extensive set of trials with an m-health tele-monitoring service.

For providing the e2eQoS predictions, the QoS predictions CS needs the following set of data: a) current device location and time, b) device's CPU type, clock speed and battery level, c) user's mobility path information. The source and destination locations of a users's trip are mapped to the respective location coordinates and a route calculation technique (similar to the GPS navigation system) is used to calculate the coordinates along mobility path of the user. The transportation mode information is used to predict the expected time to reach the destination. Our preliminary results show that in case of low mobility, the e2eQoS is predicted with 80% accuracy. However, this work is still ongoing and we are awaiting the conclusive results.

4.1.1 Context Processor

The role of the *Context Processor* (CP) component is to get/subscribe context information from the context sources and provide a necessary aggregated context information to the context reasoner to be able to make a network selection decision at a given time and location. Upon the activation of the first device service on a given mobile device, CP obtains the user's trip information from the user trip information CS, current location and time from the location and time CS, device context information from the device CS and provides this information to the QoS prediction CS to get a complete QoS prediction information along the user mobility path. We refer to a part of QoS prediction information useful at the given location and time as *current QoS predictions*. On the activation of first device service, CP also subscribes to the user preferences CS, device service CS and communication CS. The information obtained from these context sources and the current network predictions together form the *current context snapshot.* The current context snapshot is updated in real-time to accommodate the context changes received from other context sources. CP sends the current snapshot tagged with the relevant *context change event* (refer to Section 3.3.3 for the event types) to CR for the further processing.

Handling Missing/Incomplete/Probabilistic Location and QoS Predictions Context Information

It is very likely that though the user provides the trip information, the actual path taken by the user (e.g. salesman) deviates from the mobility path predicted by the QoS Predictions CS. Moreover, it is also possible that the user is not on the move, but staying at the current location for a long time and just move around (e.g. home and office near to each other). To handle these problems, if the user trip information is available, then the QoS predictions CS also sends the e2eQoS predictions for the locations within a certain distance from the mobility path. If there is no user trip information available, then the e2eQoS predictions for all the locations within a certain radius of the current user location are sent. This behavior of the QoS predictions CS could be compared with the GPS navigation systems which cache the complete area map along the user mobility path. Herewith, the area map corresponds to the e2eQoS map.

The QoS prediction information is also probabilistic because it is derived from the historical data. So, if the device services' perception of the currently selected network degrades, there is a mechanism in the CR for the selection of alternate network. Furthermore, in cases of: i) unavailability of the predictions or ii) an unpredicted loss of connectivity and lack of further QoS predictions, the information provided by a Communication CS is used to make a (rough) handover decision to a new mobile network.

4.1.2 Context Reasoner

The *Context Reasoner* (CR) is an event driven component and is responsible for the selection of one of the available networks to handover to, by considering user preferences, device service requirements, device services' perception and its criticality level. To be able to make use of the available context information, CR considers the following set O of objectives for optimization for the selection of the network:

Objective 1: Maximize user's network preferences.
Objective 2: Consider user preferences for power consumption.
Objective 3: Maximize device services' throughput requirements.
Objective 4: Minimize device services' delay requirements.

Basic Analytic Hierarchy Process Method

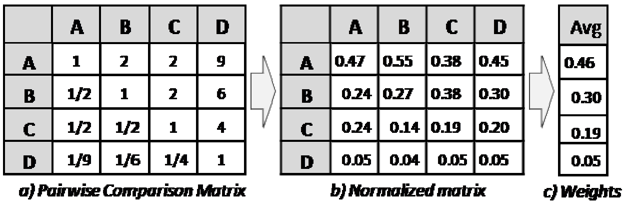
Since the problem of the selection of desired network consists of satisfying a number of objectives and there are a limited number of networks to be chosen from, we follow similar approach to [3] and [6] of using Analytic Hierarchy Process (AHP) method for the optimization. The ability of AHP to vary its weighting between each objective is useful for dealing with events as described further in the Section 3.2.3.2. AHP involves calculations using simple formulas and hence is expected to provide better computational performance than the alternative optimization techniques such as Genetic Algorithms. As described in [28], AHP is about dividing a problem into several sub-problems and later aggregating the solutions of these sub-problems into a conclusion. AHP method applied to our problem consists of the following three steps:

Step 1: *Decide relative importance of the optimization objectives.* The importance of an objective is decided by the weight assigned to it. These objective weights are always assigned such that their combined sum is 1. For example, if the value of user's power preference is `Yes`, then every objective is assigned weight $O_i = 0.25$ and $1 \le i \le 4$; otherwise the second objective is assigned weight 0 and the other objectives' weights are 0.33. The weights of the objectives also vary as per the type of events described in the Section 3.2.3.2.

Step 2: *Compute relative weight of each available network for each objective.* This step consists of the following sub-steps:

1. For each of the primary optimization objectives, assign an integer score between 1–9 for every available network depending on its position in the user preferences list, interface power requirements and predicted e2eQoS (throughput and delay available to the device service). The network with the best (worst) values in every category receives score 9 (1). E.g. four networks named A, B, C and D respectively may get score assigned 9, 6, 4 and 1 according to the user preferences list.
2. If n is the number of elements in a set of available networks N, for each objective, based on the network score S_i where $1\le i \le n$, calculate the pair-wise comparison matrix with values $P_{ij} = S_i/S_j$ for each i and j such that $1 \le (i, j) \le n$ and $i \le j$. P_{ij} is rounded off to the nearest integer. For $i > j$, $P_{ij} = 1/P_{ji}$. E.g. Continuing example above, $P_{AB} \Rightarrow Int\ (S_A/S_B + 0.50) \Rightarrow Int\ (9/4 + 0.50) \Rightarrow Int\ (2.75) = 2$.
3. For each optimization objective Oi, normalize each P_{ij} (divide by the sums of the columns) and average across rows to obtain the relative weights of the networks W_{no}.

For the above example, the pair-wise comparison matrix, weight matrix and relative weights according to the user preferences are shown in Figure 2.



	A	B	C	D
A	1	2	2	9
B	1/2	1	2	6
C	1/2	1/2	1	4
D	1/9	1/6	1/4	1

a) Pairwise Comparison Matrix

	A	B	C	D
A	0.47	0.55	0.38	0.45
B	0.24	0.27	0.38	0.30
C	0.24	0.14	0.19	0.20
D	0.05	0.04	0.05	0.05

b) Normalized matrix

Avg
0.46
0.30
0.19
0.05

c) Weights

Figure 2: AHP Calculation Example

Step 3: *Calculate the score for each network and select the network having the highest score.* The network score is the sum of relative network weights multiplied by the objective weight.

After selecting the network with the highest score, CR instructs the *Message Worker* and *Stream Worker* components to use the IP address of this network for the data transfer (thus completing the handover procedure).

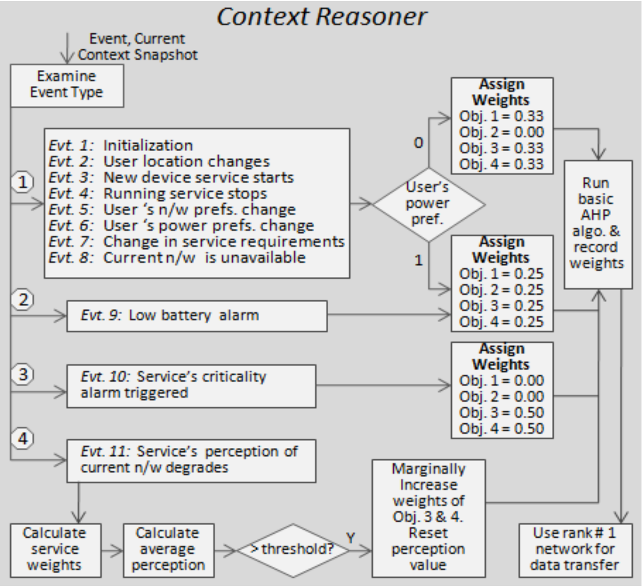

Context Reasoner
Event, Current Context Snapshot
Examine Event Type
Evt. 1: Initialization
Evt. 2: User location changes
Evt. 3: New device service starts
Evt. 4: Running service stops
Evt. 5: User 's n/w prefs. change
Evt. 6: User 's power prefs. change
Evt. 7: Change in service requirements
Evt. 8: Current n/w is unavailable
1
2
3
4
User's power pref.
0
1
Assign Weights Obj. 1 = 0.33 Obj. 2 = 0.00 Obj. 3 = 0.33 Obj. 4 = 0.33
Assign Weights Obj. 1 = 0.25 Obj. 2 = 0.25 Obj. 3 = 0.25 Obj. 4 = 0.25
Run basic AHP algo. & record weights
Evt. 9: Low battery alarm
Evt. 10: Service's criticality alarm triggered
Assign Weights Obj. 1 = 0.00 Obj. 2 = 0.00 Obj. 3 = 0.50 Obj. 4 = 0.50
Evt. 11: Service's perception of current n/w degrades
Marginally Increase weights of Obj. 3 & 4. Reset perception value
Calculate service weights
Calculate average perception
> threshold?
Y
Use rank # 1 network for data transfer

Figure 3: Context Reasoner Operation

Dealing with Various Events

In principle, CR deals with different events received along with the current context snapshot received from the CP by changing the weights of the optimization objectives. Figure 3 shows the list of events and the corresponding option. The names of events 1–10 and assigned objective weights are self-explanatory. We explain *Event 11* herewith.

Event 11: *Device service's perception of the network currently in use degrades.* To consider the user's device service priorities, on the receipt of this event, CR calculates the relative weight of each running device service according to user's service preferences (Similar to *Step 2* of the basic AHP algorithm). The overall perception is averaged over the sum of multiplication of the relative weight of each device service and the corresponding perception level. If the overall perception is beyond a certain threshold, then the weights of objectives 3 and 4 are increased marginally (and objectives 1 and 2 are decreased accordingly), the device service perception level is reset and the basic AHP algorithm is executed. This strategy ensures that the device services important to the user are given a fair treatment.

5 Technologies' choices for implementation

As described in the Section 3.1, MSP implementation is based on Jini surrogate architecture specification. The Message Worker and Stream Worker are java threads and use *Apache HTTPClient library* to send messages and transmit data to the surrogate host. CR converts the IP address of the selected network interface to the *InetAddress* and changes the *hostConfiguration* which is later used by the *HTTPClient* to open an HTTP connection. The communication CS implementation is based on the Network Abstraction Layer (NAL) reference implementation for Windows CE with the extensions to generate network resource descriptions in XML [26]. The current implementation of the QoS prediction engine using DBN is programmed in a Matlab environment; however we also work on deployment of the trained DBN model on the mobile device.

We are currently evaluating technical choices for the implementation of other elements. The CP, QoS prediction CS and surrogate host CS will use the Context Distribution Framework (CDF) for context exchange. CDF proposed in [29] takes a service oriented and *Quality of Context* driven approach for the distribution of context information within the mobile/fixed environment. The development of location and time CS could be based on GPS library developed in J2ME and available at [30] to interface with the GPS device. For developing the device CS, device service CS, user preferences CS, user trip information CS and surrogate host CS, MSP will provides necessary base classes developed in Java which could be extended to implement the desired functionality. For calculating the coordinates of the user mobility path, we are expecting to use a modified version of an open source street navigation solution named *Roadnav* [31] in the QoS predictions CS. The validation setup for the proposed architecture will an extension of the *System Under Test* (SUT) described in [25]. The SUT used in [25] consists of a Body Area Network (BAN), Qtek9090 PDA running Windows Mobile 2003 and a J2ME compliant JVM, MSP, communication CS and the GPRS, WLAN and USB network connections.

6 Conclusions and Future Work

This paper presents a context-aware middleware architecture for providing vertical handover support to nomadic mobile services which are hosted on a multi-homed mobile device to provide data to the clients located in the Internet. The proposed context information set for the mobile network selection includes service requirements, user preferences, device capabilities, interface power consumption, user mobility, service criticality and end-to-end QoS (e2eQoS) prediction information. Particularly, for providing near-accurate estimate of the e2eQoS and minimizing power required at the mobile to search all the available wireless networks, the QoS prediction context source located in the fixed network provides predictions on the availability of mobile networks and their estimated e2eQoS along the path that a users travels. Our middleware also provisions the mechanisms to handle missing and incomplete and probabilistic location (or other context) information and still be able to provide e2eQoS predictions. The context reasoner is an event-driven component which deals with different context change events by changing the weights of optimization objectives and further using the Analytic Hierarchy Process method for handover-decision making. We apply some of the concepts in our architecture from related work.

While some of the elements of the proposed architecture are already implemented, currently some are under implementation and we are evaluating technical choices for the rest of them. We propose the validation of the proposed architecture for the remote patient tele-monitoring service in the (mobile) m-health domain aiming to achieve a better perceived performance for healthcare professionals, while optimizing battery usage.

References

[1] Hui, S.Y. and K.H. Yeung, *Challenges in the Migration to 4G Mobile Systems.* IEEE Communications Magazine, December 2003. 41(12).

[2] Cunnigham, G., P. Perry, and L. Murphy, *Soft, Vertical Handover of Streamed Video*, in *Fifth IEE International Conference on 3G Mobile Communication Technologies, 2004*, October 2004: London, UK.

[3] Balasubramaniam, S. and J. Indulska, *Vertical handover supporting pervasive computing in future wireless networks.* Elsevier Computer Communications, March 2004. **27**(8): p. 708-719.

[4] Pawar, P., et al., *A Comparative Study of Nomadic Mobile Service Provisioning Approaches*, in *International Conference and Exhibition on Next Generation Mobile Applications, Services and Technologies (NGMAST 2007).* September, 2007: Cardiff, UK.

[5] Halteren, A.v. and P. Pawar. *Mobile Service Platform: A Middleware for Nomadic Mobile Service Provisioning.* in *2nd IEEE International Conference On Wireless and Mobile Computing, Networking and Communications (WiMob 2006).* June 2006. Montreal, Canada.

[6] Ahmed, T., K. Kyamakya, and M. Ludwig, *Architecture of a Context-Aware Vertical Handover Decision Model and Its Performance Analysis for GPRS – WiFi Handove*, in

11th IEEE Symposium on Computers and Communications (ISCC'06). 2006: Sardinia, Italy.

[7] McNair, J. and F. Zhu, *Vertical handoffs in fourth-generation multinetwork environments.* IEEE Wireless Communications, June 2004. 11(3): p. 8-15.

[8] 8. Vidales, P., R. Chakravorty, and C. Policroniades, *PROTON: A Policy-based Solution for Future 4G devices*, in *Fifth IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY'04)*. June 2004: New York, US.

[9] Murray, K., R. Mathur, and D. Pesch, *Intelligent Access and Mobility Management in Heterogeneous Wireless Networks using Policy*, in *ACM 1st International Workshop on Information and Communication technologies*. 2003: Dublin, Ireland. p. 181-186.

[10] Hasswa, A., N. Nasser, and H. Hassanein, *Generic Vertical Handoff Decision Function for Heterogeneous Wireless Networks*, in *2nd IEEE/IFIP International Conference on Wireless and Optical Communications Networks (WOCN '05)*. March 2005: Dubai, UAE.

[11] Sinderen, M.J.v., et al., *Supporting context-aware mobile applications: an infrastructure approach.* IEEE Communications Magazine, 2006. 44(9): p. 96-104.

[12] Pawar, P., et al. *Context-Aware Middleware Support for the Nomadic Mobile Services on Multi-homed Handheld Mobile Devices*. in *12th IEEE Symposium on Computers and Communications (ISCC 2007)*. 2007. Aveiro, Portugal.

[13] Wac, K., A.v. Halteren, and D. Konstantas. *QoS-predictions service: infrastructural support for proactive QoS- and context-aware mobile services*. in *International Workshop on Context-Aware Mobile Systems (CAMS) colocated with On The Move 2006*. October-November 2006. Montpellier, France.

[14] Akkari, N., S. Tohmé, and M. Doughan, *Applying Anticipated vertical handover (AVHO) in Next Generation Networks*, in *Fourth European Conference on Universal Multiservice Networks (ECUMN'07)*. February 2007: Toulouse, France. p. 311-319.

[15] Stevens-Navarro, E. and V.W.S. Wong, *Comparison between Vertical Handoff Decision Algorithms for Heterogeneous Wireless Networks*, in *IEEE 63rd Vehicular Technology Conference, 2006 (VTC 2006-Spring)*. May 2006: Melbourne, Australia. p. 947-951.

[16] Calvagna, A. and G.D. Modica, *Effects of the Vertical Handover Policy on the Performance of Internet Applications*, in *3rd European Conference on Universal Multiservice Networks (ECUMN'2004)*. October 2004: Porto, Portugal. p. 358-366.

[17] Joe, I. and S. Hong, *A Mobility-based Prediction Algorithm for Vertical Handover in Hybrid Wireless Networks*, in *2nd IEEE/IFIP International Workshop on Broadband Convergence Networks, 2007 (BcN '07)*. May 2007: Munich, Germany p. 1-5.

[18] 18. Hui-ling, J., Z. Zhao-yang, and L. Shi-ju, *A Power Threshold Based Policy for Vertical Handoff in Heterogeneous Networks*, in *International Conference on Wireless Communications, Networking and Mobile Computing, 2005 (WCNM 2005)*. September 2005: Wuhan, China. p. 1052-1055.

[19] Chen, W.-T. and Y.-Y. Shu, *Active Application Oriented Vertical Handoff in Next-Generation Wireless Networks*, in *IEEE Wireless Communications and Networking Conference 2005 (WCNC 2005)*. 2005: New Orleans, LA, USA. p. 1383-1388.

[20] Salamah, M., F. Tansu, and N. Khalil, *Buffering Requirements for Lossless Vertical Handoffs in Wireless Overlay Networks*, in *The 57th IEEE Semiannual Vehicular Technology Conference, 2003 (VTC 2003-Spring)*. April 2003: Jeju Island, Korea. p. 1984-1987.

[21] Chen, L.-J., et al., *USHA: a simple and practical seamless vertical handoff solution*, in *IEEE Consumer Communications and Networking Conference (CCNC'06)*. 2006: Las Vegas, USA.

[22] *Jini Technology Surrogate Architecture Specification*. October 2003 [cited; Available from: https://surrogate.dev.java.net/specs.html.

[23] Wang, Q., et al., *Hybrid User- and Network-Initiated Flow Handoff Support for Multihomed Mobile Hosts*, in *65th IEEE Vehicular Technology Conference, 2007 (VTC2007-Spring)*. April, 2007: Dublin, Ireland. p. 748-752.

[24] Jesus, V., et al., *Mobility with QoS Support for Multi-Interface Terminals: Combined User and Network Approach*, in *12th IEEE Symposium on Computers and Communications (ISCC 2007)*. July, 2007: Aveiro, Portugal. p. 325-332.

[25] Pawar, P., et al. *Performance Analysis of Nomadic Mobile Services on Multi-homed Handheld Devices*. in *2007 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS 2007)*. July 2007. San Diego, CA.

[26] Peddemors, A., I. Niemegeers, and H. Eertink. *An Extensible Network Resource Abstraction for Applications on Mobile Devices*. in *2nd IEEE/Create-Net/ICST International Conference on Communication System Software and Middleware (COMSWARE 2007)*. January 2007. Bangalore, India.

[27] Murphy, K., *The Bayes Net Toolbox for Matlab.* Computing Science and Statistics, 2001. 33.

[28] Saaty, T.L., *How to make a decision: The analytic hierarchy process.* European Journal Operation Research, 1990. 48: p. 9-26.

[29] Pawar, P., A.v. Halteren, and K. Sheikh. *Enabling Context-Aware Computing for the Nomadic Mobile User: A Service Oriented and Quality Driven Approach*. in *IEEE Wireless Communications & Networking Conference (WCNC 2007)*. 2007. hong Kong.

[30] Deriaz, M. *GPS Library in J2ME*. 2006 27 April 2006 [cited 2007 21 August 2007]; Available from: http://www.universal-locator.com/.

[31] Lynch, R. *Roadnav: Open Source Street Navigation Solution*. 2004 - 2007 30 June 2007 [cited 01 September 2007]; Available from: http://roadnav.sourceforge.net/.

Collaborative QoS-information Sharing for Mobile Service Users: A Web 2.0 Business Model proposal[1]

Katarzyna Wac, Richard Bults, Dimitri Konstantas, Hong Chen and
Bert-Jan van Beijnum

Abstract. Mobile service providers (MoSPs) emerge, propelled by ubiquitous availability of mobile devices and wireless communication infrastructures. MoSPs' customers satisfaction and consequently their revenues, largely depend on the quality of service (QoS) offered by wireless network providers (WNPs) at a particular location and time of a mobile service usage. This chapter presents a novel business method for the MoSP's QoS-assurance process. The method incorporates a location- and time-based QoS-predictions service facilitating the improvement of the WNP's selection process. We introduce and analyze business viability of QoSIS.net, an enterprise that provides the QoS-predictions service to MoSPs or directly to its customers (i.e. in B2B or B2C settings). QoSIS.net provides highly accurate QoS-predictions based on collaborative-sharing of QoS-information by its users. We argue that this business method can improve the MoSP's QoS-assurance process and consequently may increase its revenues, while creating revenues for QoSIS.net.

Keywords: Wireless and mobile services**,** Quality of Service, Web 2.0, collaborative information-sharing, QoS-predictions, data mining

Introduction

The last 15 years have been marked by the expansion, global adoption and seamless availability of the Internet with a multitude of its ubiquitous services. At the same time, a new era has undergone its preparation phase. Namely, the service users around the world, who have entered the digital era in the 1990s-early 2000s, are now entering the mobile era (Hansmann *et al.*, 2003). This era has been particularly propelled by miniaturization and personalization of communication devices, as well as the rapid expansion and adoption of mobile voice and data services and ubiquitous wireless communication infrastructures. In this era, ubiquitous mobile service providers (abbreviated through the document as *MoSPs*) bring to their customers their favorite existing Internet services and start offering, on a growing scale, a wide range of new mobile services. However, these MoSPs are fully aware that in order to gain customer acceptance, and secure own revenue, their mobile services must provide customers with a *quality of experience (QoE)* (ITU-T, 2007) comparable to the existing Internet-services (Afuah & Tucci, 2000). As part of customer QoE, a MoSP must at least assure meeting customer's (implicit or explicit) *quality of service* (*QoS*) requirements, expressed e.g. in terms

[1] A revised version of this paper is included as a chapter in the forthcoming book *Mobile and Ubiquitous Commerce: Advanced E-Business Methods*, M. Head, (Eds.), IGI Global publisher, vol. 4 (to appear in 2009).

of service performance, security level and monetary cost (ITU-T, 1993). However, to stay competitive, the MoSP should assure meeting the required customer's QoS and in a best-possible way meet his anticipated-QoE (Andersson *et al.*, 2006).

To achieve this goal, MoSPs require dependable wireless communication infrastructures supporting mobile service delivery to their customers, anytime, anywhere and anyhow. These infrastructures, and in particular their providers, i.e., *Wireless Network Providers* (*WNPs*), must be able to either match their users' (i.e. MoSPs) QoS-requirements, or provide detailed and precise information about their offered-QoS. This information would enable MoSPs to adapt their service delivery and assure meeting the required QoS-level of their mobile users (i.e. customers).

Nowadays, in almost every country, there co-exists a number of WNPs, operating different long-range wireless communication technologies. In particular, there exists at least one national *Mobile Network Operator (MNO)*, providing primarily mobile voice and data services over long-range wireless communication technologies (e.g. GSM/CDMA, GPRS/EDGE, UMTS/WCDMA/ HSDA). In co-existence with MNO's, WNPs like public WLAN providers emerge rapidly, especially in big cities. Moreover, new mobile devices support a multitude of long-range wireless technologies, as well as short-range wireless technologies (e.g. Bluetooth). Hence, communications means become ubiquitously available to mobile service users and MoSPs. This means that, at least in principle, a mobile service user and MoSP must be able to choose a WNP (and wireless technology) at any location and time, offering the QoS that meets user required- QoS thus meets his anticipated-QoE.

However, this scenario is far from reality today. The business strategy of existing WNPs, and particularly MNOs, is based on a user 'lock-in' (Buschken, 2004); the user can only choose from wireless networks (and therefore technologies) offered by 'his' WNP. Moreover, the information about QoS-offered by a WNP is based on marketing information providing numbers related to the network's theoretical performance; the real (i.e. objectively measured) QoS is unknown! Surprisingly, even mobile services provided by WNP's are based on assumptions regarding their offered-QoS. It is widely accepted practice that mobile service performance tests conduced by WNPs (and especially by MNOs) are based on 'drive-tests' (Gomez & Sanchez, 2005) and focus only on the availability of WNP's wireless communications technology at different locations. Out of these tests, the only information disclosed by WNPs to MoSPs and their customers are coverage maps. In effect, the WNPs offer their service at 'best-effort' level to MoSPs. As a result, it is not possible for a MoSP to select at a certain location and time a WNP (and technology), which offers QoS, that best assures meeting its customer's QoS-requirements. A MoSPs are constrained by above restrictions and can only provide mobile services to their customers at a 'best-effort' level.

We envision that the above restrictions on WNP selection will dissolve in the next decade of mobile service provisioning. Already we see *Mobile Virtual Network Operators (MVNOs)* appearing on the market, allowing their customers to choose a WNP from a (still small) selection of partner MNOs. With larger numbers of WNPs participating in a MVNO, a larger choice becomes available to a MoSP at any location and time. However, this does not solve entirely the problem of how the MoSP can make the selection of a WNP that best assures meeting the QoS-requirements of its mobile customer! Accurate information regarding QoS-offered by WNPs at a location and time is not available.

Because of the situation today, a MoSP that offers QoS-demanding mobile services, like mobile healthcare (MobiHealth, 2007) or mobile games (Digital Chocolate, 2008), face difficulties related to QoS-assurance to their customers and adherence to their anticipated-QoE. They are forced to provide their services at a ‘best-effort’ level; i.e. they can only base their customer’s QoS-assurance business process (part of QoS-operational management (TMF, 2004)) on assumptions about QoS provided by WNPs. These assumptions are usually derived from WNPs’ marketing (theoretical) QoS-information. However, these assumptions can be far from reality, which may significantly influence user-QoE and consequently influence MoSP’s revenues.

Based on the current situation of mobile service provisioning as presented above, in this chapter we propose a novel business method (i.e. a novel method of doing business) that incorporates location- and time-based QoS-predictions into MoSP’s QoS-assurance business process. Towards this end, firstly, we describe and analyze the viability of a new business enterprise: *Quality of Service Information Service Provider - QoSIS.net*, that offers an accurate QoS-predictions service to MoSPs and its customers. A QoS-prediction is a prediction of QoS-offered by a WNP (for a given technology) at the MoSP’s customer location and time. For QoSIS.net, we propose a novel business method to be employed in its operational (core) business process. The method is based on ‘users-collaborative-sharing’ of QoS-information, acquired when different MoSPs and their customers, use different WNPs at a particular location and time. The method’s novelty (as well as its major risk factor) lies in the fact that QoSIS.net can provide highly accurate QoS-predictions to MoSPs only based on a large volume of QoS-information acquired from its customers.

Secondly, we propose of a novel business method for MoSPs employing the QoSIS.net QoS-predictions service in its QoS-assurance business process. The method aims at using the QoS-predictions service to select a WNP (and technology) at a given MoSP’s customer location and time. It supports meeting of the mobile user’s QoS-requirements and improvement of his QoE beyond ‘best-effort’ level. The goal of the proposed method is to improve (e.g. in terms of efficiency and effectiveness) the MoSPs QoS-assurance process. We argue that the proposed business methods require trustful business inter-dependency between MoSPs and QoSIS.net, while it has a strong potential to bring mutual benefit to them - in terms of increase of MoSP revenues and creation of revenues to QoSIS.net.

2 Current trends

An issue of quality of service assurance in Internet-based-services and its relation with Internet performance has been indicated as a critical factor already in 1990s (ITU-T, 1993). QoS has been defined the as “collective effect of service performances which determine the (objective) degree of satisfaction of a user of the service”. It has been recognized, that QoS-offered by Internet influences the

QoS provided by Internet-based-services to its users, and hence these users' quality of experience (QoE). QoE has been defined as "the overall acceptability of an application or service, as perceived subjectively by the end-user" (ITU-T, 2007).

The Internet-based-services providers recognized early the necessity for QoS-assurance business processes to meet their users' anticipated-QoE requirements and secure their revenues. First solutions for the QoS-assurance business processes have been proposed particularly for providers of real-time multimedia services (Hutchison *et al.*, 1997; Shepherd *et al.*, 1996). These solutions, from the technical perspective, recommended use of a rigorous and complex (!) QoS management frameworks, employing functions like QoS negotiation and resource reservation (Andersen *et al.*, 2000; Xiao & Ni, 1999). Moreover, from the business perspective, these solutions required business contracts between Internet-based-services providers and network providers, i.e. Internet-providers (Afuah & Tucci, 2000). These solutions however contradicted with the 'open' nature of Internet and its services, because in effect they limited the Internet-based-services provider customer-base to the Internet-providers' base. In this situation, Internet-based-service providers skipped the proposed technological and business solutions and learned to deal with 'best-effort' quality offered by Internet, while being able to assure the QoS to their service users. The business methods, based on which they build their QoS-assurance process, relied on estimations of QoS-offered by Internet. This approach has been feasible due to at least two facts. Firstly, QoS-offered by Internet exhibits some regularities (claffy *et al.*, 1998), hence estimations could be valid for a longer period of time (e.g. months). Secondly, if necessary, providers could easily acquire offered-QoS-estimations via dedicated QoS monitoring, and all that for free and without an degradation of quality of their provided services (Michaut & Lepage, 2005)!

With the raise of mobile era, we are somehow repeating the history regarding QoS-provisioning. It has been indicated as a critical issue already in 1999 (Chalmers & Sloman, 1999); at the time when only basic voice and data services existed. It is recognized that the most critical factor in QoS assurance is related to user's mobility; a mobile user is exposed to access WNP different access points at different locations and times, or even to use different WNPs (over different wireless technologies) along his mobility path (Dekleva *et al.*, 2007). To deal with it, and to meet users' QoS-requirements, mobile service providers (MoSPs) are advised to employ a QoS management framework as a business method in their QoS-assurance business processes. From the business perspective, this solution requires a business relationship between a MoSP and a WNP supporting a mobile service delivery; which results in a WNP-centric business models. There is a lots of research supporting solutions of this type, for example (Han & Venkatasubramanian, 2006; Soh & Kim, 2003) propose MoSPs' QoS-assurance process to employ predictions of user mobility path acquired from MNO' access points (i.e. base stations). Similarly, MNOs work on new concepts like Universal Mobile Access, Generic Access Network and IP-Multimedia System (Cuevas *et al.*, 2006), striving to provide technological as well as contractual solution for MoSP QoS-assurance process tightly coupled with MNO's business processes.

Again, such approaches contradict the 'mobile' nature of services provided by a MoSP; a business relation with a WNP would limit its customer base and service usage area to the WNP's customer-based and its coverage-area. Despite this, the WNP-centric business models seem to be dominant today as presented in literature (Calvo *et al.*, 2004; Faber *et al.*, 2003; Robles *et al.*, 2002; Tan, 2004; Tsalgatidou & Pitoura, 2001).

However, we can observe that the individual MoSPs, not following the WNP-centric business models, emerge (Tan, 2004). The MobiHealth.com (MobiHealth, 2007) is an example of these in healthcare domain, while Digital Chocolate (Digital Chocolate, 2008) – in mobile gaming domain. Providers in both domains struggle with QoS-offered by WNPs to assurance of QoS to their users (Bults *et al.*, 2005; Busse *et al.*, 2004). Moreover, in parallel, we can observe rise of new WNPs or new long-range wireless technologies (like e.g. Ultra-Mobile-Wideband) being employed by existing WNPs, and that all striving to 4G vision of plentiful WNPs and wireless technologies being available for mobile users (De Vriendt *et al.*, 2002; Dekleva et al., 2007; Ortiz, 2007; Tachikawa, 2003). Yet, whenever a new WNP appears, it is a commonly used practice that it does not provide any information about its offered-QoS; it always starts with a 'best-effort' level offers (Gomez & Sanchez, 2005).

In light of these developments, and with respect to the fact that MoSP are very likely to be highly mobile and roaming between different WNPs, new trend appears in research on mobile business, focusing on user-centric business models. The aim is to enable for a MoSP to meet user QoS-requirements and his anticipated-QoE with use of any WNP at user given location and time. Towards this direction, there exists technically-oriented research proposals for e.g. advanced signaling between involved WNPs (Bless *et al.*, 2004; ITU-T, 2006) or new user-centric approaches for WNPs and MoSPs (Manner *et al.*, 2001). There are even large EU projects dedicated to new business methods for MNOs to support MoSPs, without locking-in their customers (Sanchez *et al.*, 2008). There exists also proposal of forming smart-business networks (van Heck & Vervest, 2007) to support special MoSPs users' QoS-requirements by building 'ad-hoc' business relationships between enterprises.

Still the question remains: how a MoSP can choose a WNP best matching his user's QoS-requirements at a given location and time. Methods used in Internet-based services do not work in mobile services; the QoS estimations vary per location and they change in time, moreover, dedicated QoS-monitoring is costly in terms of money and mobile device resources, e.g. battery and network throughput, which can in turn degrade the provided mobile services.

This situation supports our proposal for a MoSP QoS-assurance business process employing a QoS-predictions predictions service of QoSIS.net (details in the next Section). To our knowledge, this kind of solution has not been proposed yet in the literature. The novelty (as well as a risk factor) of QoSIS.net is in the fact that it provides pre-existing offline community of mobile service users with a complementary online service. This is particularly an idea behind all Web 2.0 services (Hoegg *et al.*, 2006; Martignoni & Stanoevska-Slabeva, 2007; O'Reilly, 2005; Pascu *et al.*, 2005), that emerge successfully nowadays in different domains and gain users' acceptance (e.g. Facebook, YouTube, Wikipedia). Particularly, Web 2.0 is coined by O'Reilly (2005) as "the philosophy of mutually maximizing collective knowledge and added value for each *participant* by formalized and dynamic sharing and creation of user generated content” (p.1) and its ‘mobile’ extension means that it is implemented as a mobile service. QoSIS.net employs a Mobile Web 2.0 paradigm, and its customers are such ‘participants’ contributing to community with acquired QoS-information. This chapter presents feasible business models for QoSIS.net, and the existing technical solutions supporting it service delivery.

3 QoSIS.net: Collaborative QoS-information sharing for mobile service users

In this and the upcoming sections we exploratively (rather than exhaustively) introduce and analyze the business viability of QoSIS.net along the business model framework given by (Hoegg et al., 2006); we consider its service, potential market-segments and customers, value chain, revenues-costs model and so on.

3.1 Features of a Value-add Service

QoSIS.net delivers a value-add QoS-predictions service to its customers: MoSPs and their customers. QoS-prediction is a prediction (along with its accuracy estimation) on QoS-offered by a WNP and wireless communication technology at a MoSP’s customer location and time. The QoS-predictions are used by a MoSP to select a WNP, to match the mobile service QoS-requirements and to facilitate an improvement of his QoE beyond ‘best-effort’ level. QoSIS.net offers a value-add service to MoSPs, because it is the responsibility of a MoSP how to use the information provided by QoSIS.net.

QoSIS.net provides its service based on user-collaborative QoS-information acquisition from MoSPs and their customers. QoSIS.net incorporates QoS-monitoring, -storage, -processing, -predictions engine and QoS-predictions dissemination functionality (Figure 1 (Wac, 2006) and (Pawar *et al.*, 2008)).

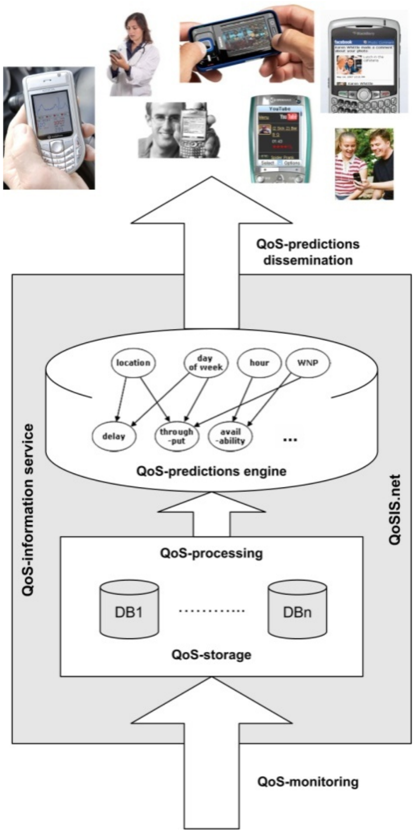


Figure 1. QoSIS.net high-level system architecture

QoS-monitoring acquires and stores QoS-information that concerns QoS observed (i.e. measured) by a MoSP and its customers when using a WNP and wireless communication technology at a particular location and time. A request for a QoS-predictions service, results in instantaneous processing of large quantities of QoS-information by prediction engine. The engine uses data mining techniques (e.g. Bayesian Networks (Heckerman, 1996)) to discover QoS-information patterns. The result is returned to the QoSIS.net customer (i.e. service user).

QoSIS.net provides its service users a private mobility-map (stored on the user's mobile device) showing all the visited locations so far combined with QoS-predictions information. The mobility-map is automatically updated, either at regular time-intervals or at the moment a prediction is requested. The information in the map allows the QoSIS.net service user to get historical QoS-predictions even if the user is out of coverage of any WNP.

One of the technological challenges for QoSIS.net is the fulfillment of customer expectations in terms of service speed, dependability, accuracy, interoperability, security, scalability and fault-tolerance ((Henricksen *et al.*, 2005), (ITU-T, 1993)). It is important to notice that QoSIS.net can use its own QoS-predictions services as value-add services in its QoS-assurance business process; i.e., to select a WNP (and technology) with use of which its prediction service is going to be delivered to its customers at a given location and time.

QoSIS.net employs a generic QoS-information model that includes (Chalmers & Sloman, 1999; ITU-T, 1993)):

a) WNP-related information – a provider name, wireless technology used (e.g. 3G/WLAN)
b) WNP performance - speed, delay/throughput, accuracy, dependability (incl. availability),
c) WNP monetary cost - e.g. per MB, per hour
d) WNP security features (e.g. authentication, confidentiality, integrity, non-repudiation)
e) MoSP services used and usage context i) mobile device used and ii) user location and time

Note that the mobile service (i.e. application)-specific user's QoE-related parameters like e.g., picture resolution, audio quality, AV rate/synchronization are outside of the scope of the QoS-information model, due to their application dependency. It is important to notice that the QoS-information does not have any notion of mobile user's identity, all information is anonymous.

3.2 Features of the Service Medium

After describing QoSIS.net service, the question arises what the service medium features are of importance for QoSIS.net service delivery to its customers. We emphasized that the QoSIS.net service is beneficial for mobile service users. Therefore, the features of the service medium depend on the user's context (e.g. WNPs available at a particular location) and may intentionally or accidentally disable or disturb QoSIS.net service delivery (Camponovo & Pigneur, 2003; Tsalgatidou & Pitoura, 2001). Hence, the user context can influence the QoSIS.net business processes as well as influence the QoE of QoSIS.net customers. Dealing with service medium issues needs to be taken care of in business contracts between QoSIS.net and its customers (see further sections for details). Examples of QoSIS.net service medium-related issues are:

a) user communication-autonomy, e.g. user can deliberately configures his mobile device to use one particular WNP and technology or can be unreachable due to out of WNP coverage or empty mobile device battery
b) ACID properties (atomicity, consistency, isolation, durability) of QoSIS.net's service transactions
c) vulnerability of the user's mobile device (in terms of possible device's damage or loss) and its limited storage, processing, communication and power capacity
d) WNP's wireless communication technology characteristics, e.g. asymmetrical throughput characteristics, variable delay characteristics and restrictions on volume of QoSIS.net service data exchange (may require use of lightweight protocols and QoS-information lossless compression).

3.3 Social Environment

There are several influences of the QoSIS.net services provision, rising from customer competition, legislation and social or ethical constraints. Firstly, competition amongst QoSIS.net customers (MoSPs) requires QoSIS.net to be a trustworthy enterprise. It should apply strong security mechanisms to prevent competitive customer information is disclosed to other QoSIS.net customers; e.g. any information regarding MoSP service usage statistics or customer base should be protected. Therefore, the business relationship between a MoSP and QoSIS.net is based on a strong trust relationship (Ratnasingam & Phan, 2003) and detailed business contracts. Similarly, WNPs' competitive market situation poses strong security requirements on QoSIS.net information. Any information regarding QoS provided by a WNP should not be altered in favor of this WNP. Moreover, in order to secure QoSIS.net revenues and its competitive advantage on the market, details of its QoS-information databases or QoS-predictions engine should not be disclosed to its customers.

An important social aspect of the provided service is related to the user-privacy consent. QoS-information acquired from QoSIS.net customers contains detailed location and time information of mobile service users. Therefore, it is required (at least in Europe) that a mobile service user is legally informed of the fact that this privacy sensitive information is acquired (Gorlach *et al.*, 2004), even if it is in anonymous form. It is the responsibility of a MoSP as the QoSIS.net customer to provide user-privacy informed-consent.

3.4 Generic Revenue-Costs Model

From the perspective of QoSIS.net, we can anticipate that QoSIS.net can get revenues by selling its stored QoS-information to other value-add service providers, or it can have costs related to supplying of QoS-information or external QoS-predictions engines.

Before we analyze in details possible revenues for QoSIS.net and its customers (in upcoming sections), we indicate possible (generic) costs related to QoSIS.net's service usage. Namely, from the QoSIS.net side, costs related to the setup and maintenance of QoS-information databases and enrichment of the QoS-processing service and QoS-predictions engine towards increased QoS-predictions accuracy. From the QoSIS.net's customer side (e.g. a MoSP user), these costs are related to use of customer's resources: computation, storage and communication capacity, as well as battery, especially on the mobile user's device. Moreover, due to the nature of provided service, it is required for QoSIS.net users to own a mobile device with one or multiple interfaces for a long-range wireless network to be used via different WNPs (or at least one WNP). This device must also have a GPS, or other location-determination sensor. At this point we would like to emphasize that technical realization of inter WNP (i.e. vertical) handovers is an ongoing research issue (Chen & Shu, 2005; Dekleva et al., 2007; Pawar et al., 2008) outside of the scope of this book chapter and discussed in more details elsewhere (Wac, 2008).

3.4 Market Description and Market Entry Strategy

QoSIS.net market contains of two market segments: a Business-to-Business (B2B) market segment containing MoSPs as its customers and a Business-to-Customer (B2C) market segment containing customers, who are mobile service end-users. In the B2B case, QoSIS.net offers a value-add service to MoSPs, which is invisible to the MoSP's customer. In the B2C case however, QoSIS.net offers its value-add service directly to a customer; i.e. a mobile customer, who is a mobile service user (of e.g. Facebook or VoIP) facilitates improvement of his own-QoE, by using QoSIS.net services. In addition, the mobile customer may also act as QoS-information provider for QoSIS.net helping it to improve its QoS-predictions service by enriching its information base.

We envision that the primary market segment is B2B (MoSPs) being for a QoSIS.net a mandatory stepping stone to become successful in the B2C market segment. Since QoSIS.net business thrives on large quantities of QoS-information for its QoS-predictions service, the question raises how to obtain this information. The B2B market entry strategy is to convince MoSPs that QoSIS.net adds an accurate location- and time-based QoS-predictions service to their infrastructure that improves the QoE of their customers. We are currently working on case study on the QoS-predictions service accuracy (Wac *et al.*, 2008b). QoSIS.net offers a value-add QoS-predictions service to its B2B customers. QoSIS.net B2B customers require information provided by the prediction service to assure meeting their customers QoS-requirements, and facilitate improvement of their customer-QoE. Consequently, MoSP's can increase own revenues. Once the QoSIS.net B2B market is substantial enough (i.e. sufficient location- and time-based QoS-information is available), the B2C market is targeted where the fundamental concept of Mobile Web 2.0 lies; i.e. services are created and used by customers.

We identify critical success factors for the QoSIS.net service. The first critical success factor is related to an initial market entry barrier. QoSIS.net must have a critical mass of customers and sufficient QoS-information to provide accurate QoS-predictions to its customers. The second critical success factor is related to creation of sustainable revenues to QoSIS.net and ensuring its competitive position in the market. It is necessary to have a critical mass of customer acquiring up-to-date QoS-information, based on which QoSIS.net is able to provide a highly accurate, and therefore highly competitive, QoS-predictions service.

4 B2B Market-Segment

In this section, we take a mobile healthcare SP – MobiHealth.com (MobiHealth, 2007) - as an existing MoSP. In this B2B setting, QoSIS.net is a value-add 3rd Party service provider (TMF, 2004) for MobiHealth.com and the assumption is that MobiHealth.com uses the QoSIS.net location- and time-based prediction service to improve its service delivery, while the prediction service itself remains invisible to the MobiHealth customer (end-user). The case, in which QoSIS.net is a value-add business partner for MobiHealth.com, due to space limitations, we consider elsewhere (Wac, 2008).

4.1 MobiHealth.com: a QoSIS.net customer

MobiHealth.com provides mobile health services (m-health services) for remote monitoring of a patient's health condition. These services are based on the MobiHealth Service Platform^{TM} that consists of a Body Area Network (BAN), Internet-based application-server and user Portal. Patients wear a BAN configured for monitoring physiological data relevant for the patient's disease (e.g. COPD, cardiac condition). The BAN uses WNP's wireless communications technology to continuously send patients' vital signs data to the application-server. A care professional at a healthcare centre uses the Portal to obtain (from the application-server) and display vital sign data (offline or real-time). To support patient mobility, the BAN supports handovers between different WNP and different wireless communications technologies. The QoS-requirements of the m-health services depend on the patient's disease. These requirements are defined (in most cases) in *qualitative* terms by the care professionals responsible for patient treatment, and mapped by MobiHealth specialists to *quantitative* requirements of the MobiHealth system. The MobiHealth business model can follow (Dijkstra *et al.*, 2006). MobiHealth can either be a healthcare centre or in a business relation with one. Moreover, insurance companies can reimburse (at least partially) the m-health services usage costs to the patients.

Further considerations on the MobiHealth system and MobiHealth.com business model are outside of the scope of this paper and is presented elsewhere (van Halteren *et al.*, 2004; Wac *et al.*, 2008a).

4.2 User Scenario

Sophie is a young COPD (chronic obstructive pulmonary disease) patient and is continuously remotely monitored with a MobiHealth COPD BAN to detect exacerbations (i.e. to cause a disease or its symptoms to become more severe). She does not have to visit a care professional at the hospital frequently, feels save being remotely monitored and being less limited in her active life. Her MobiHealth

COPD BAN always uses the most suitable WNP and wireless communications technology, available at her location (and time) and this process is completely transparent to her.

4.3 Features of QoSIS.net service vs. Mobile Web 2.0 paradigm

QoSIS.net offers anywhere-anytime-anyhow, accurate location-based QoS-predictions service via user-collaborative QoS-information sharing. As we said, the QoS-predictions service facilitates the choice of WNP by MobiHealth.com; however, it is responsibility of MobiHealth.com to use (or not) QoS-predictions to choose a WNP.

In the given scenario, QoSIS.net service is employed in the MobiHealth QoS-assurance process to facilitating provisioning of the m-health services by MobiHealth.com to its users (which is MobiHealth.com's core business) and increasing of its revenue. QoSIS.net's QoS-predictions service, i.e. QoSIS.net operational (core) business process, is based on the principle of collaborative content-creation and sharing amongst the MobiHealth.com users. We argue that from this perspective, QoSIS.net can be seen as a specific example of *stand-alone Mobile Web 2.0 Service Provider* (Hoegg et al., 2006; O'Reilly, 2005). The content generated and shared by MobiHealth.com users is the QoS-information acquired in QoS-monitoring service. The QoS-processing service analyzes this data, and includes automatic content manipulation, update, rating, and information quality annotations or enrichment, facilitating QoS-predictions service provided back to the MobiHealth.com users. In this sense the QoSIS.net is not a typical example of Mobile Web 2.0 (like e.g. YouTube) because the QoS-information providers are not actively (i.e. by taking an initiative) 'participating' as content provider-and-consumers in the service (MobiHealth.com is a producer and consumer of the QoS-information), moreover they do not have an access to the QoS-information acquired from and disseminated to them. We argue that implementing the QoSIS.net service based on the principle of a user-collaborative content-creation and sharing service can benefit from the fact that MobiHealth.com's users are very likely to be in close geographical location/in given time, e.g. in one city, and therefore they will collect overlapping QoS-information, which in turn will increase it QoS-predictions service accuracy.

4.4 Potential customers

MoSPs like MobiHealth.com are potential customers of QoSIS.net. These MoSPs are particularly mobile service providers in e.g. mobile information, education, entertainment, infotainment, and education or healthcare domains. QoSIS.net can also target niche markets – small MoSP with specific user QoS-requirements being mobile in very specific location area and time e.g. mobile workers inside and outside the buildings along widespread company fields far from the city.

Services provided by these types of MoSP require frequent data exchange with their mobile users and hence frequent use of WNPs. Data exchange can be continuous or in bursts. Because the success of MoSPs mobile service is related to the QoS-offered by a used WNP, the QoS-predictions service is an important value-add service for MoSPs, which can facilitate (proactive) choice of best WNP for their user anytime-anywhere-anyhow. Therefore, the QoSIS.net's value-add to MobiHealth.com's business lies in use of the QoS-predictions by MobiHealth.com in its QoS-assurance business process (TMF, 2004). Moreover, when using QoS-predictions, MobiHealth.com saves resources involved in its services delivery e.g. money while choosing cheaper WNP, or saving on service data delays/improved throughput, or in terms of saving user's device battery life.

The biggest incentive for MoSP to use QoSIS.net would be the fact that while using QoSIS.net, MobiHealth.com improves its user's-QoE and increases its revenues while creating revenues to QoSIS.net. The QoSIS.net fulfills the MoSP need of knowledge of QoS-offered by different WNPs at mobile user location and time. Without existence of QoSIS.net MoSPs can assure QoS to its users by a) carefully designing and operating its mobile services, assuming some minimal QoS-offered by WNP b) establishing a business relationship with (a set of) WNPs, which however would limit MoSP customer base and service usage area to WNP customer base and WNP coverage-area.

4.5 Value chain

In a QoSIS.net's value chain, we only present parties (and their relationships) that play roles in the context of a QoSIS.net's service interaction that provide value to MobiHealth.com as QoSIS.net's customer. The possible role of WNP in value chain is omitted in this section and discussed further in this chapter.

In the value chain (Figure 2) we distinguish QoSIS.net, MobiHealth.com and its user (e.g. Sophie). QoSIS.net is a 3rd Party SP, i.e., value-add SP for MobiHealth.com. Their *business relation* is defined in terms of contract, payment and service usage relationship. Particularly, QoSIS.net is a provider of services to MobiHealth.com (i.e. MobiHealth.com is a *subscriber*). These two business enterprises have a *contract relationship* (a SLA), i.e., a formal negotiated and agreed between them contract defining the terms and conditions for the delivery of the services, detailed services' specifications (along the agreed QoS-information model) as well as the *payment* specifications (e.g. monthly, post-paid) by MobiHealth.com to QoSIS.net.

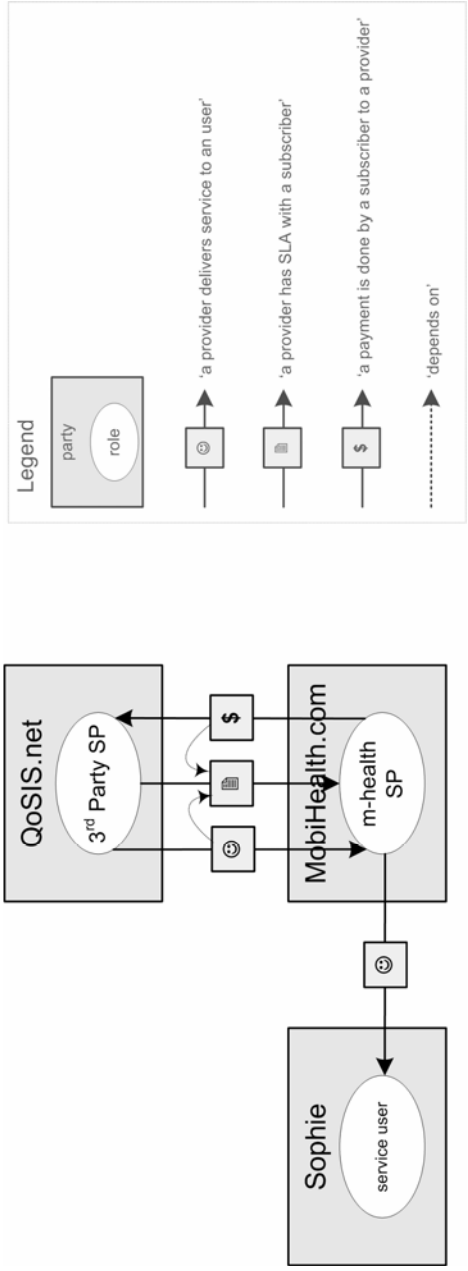

Legend
party
role
'a provider delivers service to an user'
'a provider has SLA with a subscriber'
$
'a payment is done by a subscriber to a provider'
'depends on'
QoSIS.net
3rd Party SP
$
MobiHealth.com
m-health SP
Sophie
service user

Fig.2. QoSIS.net as 3rd Party Service Provider

QoSIS.net's service *user* is not only MobiHealth.com, but particularly MobiHealth.com's users e.g. Sophie. QoSIS.net's services are seamlessly integrated in a MobiHealth.com's offer and transparent for Sophie. Anywhere she is, her BAN always uses (or handovers to) a WNP, which best assures meeting the

QoS-required by her m-health service. The WNP choice is made based on information provided by the QoS-predictions service and this choice is completely transparent to her. Moreover, along MobiHealth.com's services delivery, the QoS-monitoring and QoS-processing service are provided continuously, as they are based on the QoS-information about the QoS observed by MobiHealth.com using different WNPs.

4.6 The impact on existing value chains

The MobiHealth.com's business models will be influenced such that according to our scenario, the MobiHealth.com does not need to be in business relationship with WNP/MNO in order to be able to assure QoS to its mobile users. This will expand possible customer base of MobiHealth.com, not being limited to customer base of any given WNP/MNO (e.g. given region/country), but residing anywhere in the world, and using best WNP available there.

What is important and worth to emphasize once again is that due to the nature of service provided by QoSIS.net, a strong partnership trust (Ratnasingam & Phan, 2003) as well as technology trust (Ratnasingam *et al.*, 2002) is required between MobiHealth.com and QoSIS.net in order to assure the success of both businesses. The contracts between the enterprises need to be therefore defined in an innovative way.

4.7 B2B-specific revenues-costs model

We envision that QoSIS.net can charge its customers (MobiHealth.com), per transaction (i.e. per a single QoS-predictions service delivery) or it can introduce monthly (flat) subscription fee (anticipating particular QoS-predictions usage). Once QoSIS.net has a critical mass of customers, it can provide price differentiation. For example, the transaction fee can depend on a) number of WNPs available to mobile user at a given location and time (i.e. the price increases with number of WNPs, because the richer the choice, the higher probability that a MoSP can use WNP matching its user QoS-requirements, thus improving its user's-QoE), b) on the actual accuracy of QoS-predictions, where this accuracy can be checked against the QoS-information acquired along the acquired from user QoS-monitoring data or c) on the accuracy of QoS-predictions, which would be (on purpose) lower for lower accuracy predictions, and higher for higher-accuracy QoS-predictions. We argue that any pricing model proposed by QoSIS.net can be beneficial for QoSIS.net creating its revenues, as well as increasing MoSP's revenues, but it very much depends on the MoSPs application area and criticality of QoSIS.net service to MoSP's core business. We further consider different cases elsewhere (Wac, 2008).

QoSIS.net can have additional revenues by selling its QoS-information acquired in B2C scenario (described in next Section) to MobiHealth.com, and that in order to improve the accuracy of its QoS-predictions service.

4.8 Supporting services

From a business perspective, all QoSIS.net's services: QoS-monitoring, QoS-processing and QoS-predictions services are necessary to create a value to the QoSIS.net's service user. The QoS-information content producer and consumer is MobiHealth.com user. QoS-processing service needs to update its historical database upon each new QoS-information acquired by QoS-monitoring service. To support the QoS-monitoring service delivery, MobiHealth.com needs to obtain user-privacy informed-consent (see previous Sections) and instrument its services for acquisition of the QoS-information (along the QoS-information model agreed in contract between QoSIS.net and MobiHealth.com). Moreover, MoSP and QoSIS.net need to agree upon the QoS-monitoring service delivery: how often the data collected by a mobile MoSP's user is going to be acquired by QoSIS.net (e.g. depending on change of user location, time passed, or a WNP change). To use QoS-predictions service, MobiHealth.com needs also to instrument its services for using of the QoS-predictions in a WNP selection process.

Due to the nature of services delivered by QoSIS.net, partnership management is one of core activities of QoSIS.net, for example QoSIS.net needs to have supporting services for generating of service reports for MobiHealth.com.

4.9 Further evolvements of the business method in B2B scenario

QoSIS.net can have a dynamic SLA with MobiHealth.com, depending on actual QoS-requirements for m-health service delivery, or related to any other MobiHealth.com core-business-related objective. To tackle scalability requirement, QoSIS.net's can be a location-based service, i.e. can limit its scope of operation to particular city, region or country, hence limiting the scope of WNPs for which QoS-predictions can be provided; scoping can be dictated by need of limitation of QoS-information to be processed, or need for higher accuracy of QoS-predictions provided for a limited geographical-area. Furthermore, QoSIS.net can be limited in terms of WNPs and technologies, for which it acquires data. Moreover, in order to enhance the QoS-information base and improve the QoS-predictions accuracy, we envision that QoSIS.net customers - MoSPs whose users are likely to be in overlapping geographical-areas have a business relationship in which they agree to collaboratively share their QoS-information bases. This idea follows a vision of "smart-business-networks" (van Heck & Vervest, 2007), in which players initiate and maintain (short-term or long-term) business relationships for a purpose to deliver better services to their users; in our case all collaborating MoSPs could benefit from larger QoS-information base for QoS-predictions service provide to them.

5 B2C Market-Segment

In this scenario, QoSIS.net is an (additional) *Service Provider* (TMF, 2004) for a MoSP user (who is a QoSIS.net customer), responsible for usage of these two services accordingly to their purpose. For the purpose of this section, we take a social-networking (Mobile Web 2.0) SP – Facebook.com (Facebook, 2007) as an example of existing MoSP.

5.1 User Scenario

Eric is a student living in Amsterdam area. Nowadays he practices intensively his Spanish, while preparing for student-exchange stay in Madrid. He is a diligent student - following language evening-course and using lots of Internet-based-resources. He accesses them at PC, or (more often) at his new PDA - while waiting for/traveling by train or in a bus, finishing his course-homework. He particularly enjoys his Facebook.com social-networking space, where he meets his friends from Madrid! They sometimes chat online, post photos or a day-blog; all that in Spanish. He checks Facebook.com's marketplace for a students' room for his stay. Facebook.com is fun and excellent resource for learning about culture and practicing his language. He never gets bored!

To support his online activities, Eric's PDA by default used any available WNP in his location and time. He finds frustrating to loose the connection with his Facebook.com space, especially, when he is using it on the move.

To change it, recently he created his user-account on QoSIS.net website and downloaded from it a fancy application to his PDA. This application indicates him which WNP at his given location/time offers which service level for his PDA, and moreover it reconfigures his PDA such, it uses a WNP offering best service in terms of price/performance (and this process is completely transparent to him!). He likes the application's feel-and-look – it displays to him a mobility-map of his location and different colors correspond to (predicted) quality levels offered by different WNP. It is easy! - Green color is associated with very good quality, orange with good and red with bad quality. The application also indicates to him which WNP he is currently using (it's green, eventually orange). Using QoSIS.net already proved to work; he does not experience Facebook.com disconnections anymore while on the move; he enjoys even more his social-networking services in Spanish. Moreover, while planning a trip, for example for a weekend in city suburbs, he can always check with the QoSIS.net application (by putting location/time details by hand) what is the predicted quality offered by different WNP available there. Based on this knowledge he can prevent being disappointed that he cannot get 'online' there. Eric also knows that in case if there would be no WNP to be used, the QoSIS.net's application will start an audio/vibration alarm to him, and give him information about nearest location where a WNP can be used. As he lives in Amsterdam where there are plenty of available WNPs, this kind of alarm has not yet been raised to him.

To enable the QoSIS.net's application, the only thing he needed to do after downloading it on his PDA was a) to agree upon terms and conditions of service usage, b) to indicate that he is using Facebook.com as his primarily mobile service and c) to setup his WNP preferences by ranking quality criteria for a WNP choice: monetary cost (which for his students pocket is of priority!), performance and security. The service's terms and condition indicate that anywhere/anytime he is, QoSIS.net will always acquire from him data regarding the quality level offered by different WNP to his PDA, and that in anonymous form! This data is acquired assuring his privacy and it is later on used for predictions provided by QoSIS.net to him and other users. He knows that in this way he can ask for predictions for locations he never visited before, because there was always somebody using QoSIS.net who visited the place before him! From time to time, he just logs-in on the QoSIS.net website and checks his service usage statistics visualized in colorful easy-to-read maps: e.g. statistics of data he generated while visiting different locations along his busy days, as well as statistics of how often/when and where his PDA was requesting the update of mobility-maps, or what was the most frequent WNP he uses in which location and at which time. Moreover, on the QoSIS.net website he has already created his social community of family members and friends, who, following his invitation, also use QoSIS.net application. By collecting data regarding use of different WNPs in Amsterdam, his social community collectively improves accuracy of predictions provided by QoSIS.net to them. By default, Eric and all his community members can share their profile and statistics on the web; however, he knows that he can always easily adjust his privacy settings on the web and disable viewing his statistics by others - by the whole community or by particular person. Eric also knows that if he would be not satisfied with the services provided by QoSIS.net, he can at any point disable them from being used on his mobile and on the web. Using QoSIS.net, along usage of Facebook.com is also fun for him. He likes very much to share his service statistics with his community; he is proud because so far, by using his Facebook.com services on the move almost daily, he produces much more data than any other of his community members.

5.2 Features of QoSIS.net service vs. Mobile Web 2.0 paradigm

QoSIS.net is an example of an additional mobile service provider for Eric, a Facebook.com user. QoSIS.net provides QoS-predictions service to its users in a same way as in the B2B case. The only difference is that in B2C case, QoSIS.net, makes a choice which WNP to use at a given location and time (on behalf of a mobile user - based on his preferences). We argue that in the B2C scenario, QoSIS.net is a specific example of *stand-alone Mobile Web 2.0 Service Provider* (Hoegg et al., 2006) - QoSIS.net provides mobile service for its users based on a collaborative content-creation and sharing. The content is the QoS-information, acquired from users via QoS-monitoring service and disseminated to them via QoS-predictions service; the users are content producers and consumers. The QoSIS.net provides to its users also a web-based service, where users can see their service usage statistics (c.f. Eric), and shared them with their community. In this sense the QoSIS.net is not a typical example of Mobile Web 2.0 (like Facebook.com) because the QoS-information is not as ‘tangible’ as other multimedia data and users cannot see their contributions directly, they only see their service usage statistics visualized in maps. QoSIS.net cultivates the community culture for online and offline information exchange by means of set of formalized (IT) guidelines on its website e.g. via user’s privacy settings, chat or blog tools.

We argue that the idea of realizing of QoSIS.net operational business processes along the ‘mobile social-networking’ service that enables users to create own communities, has at least three advantages for QoSIS.net and its users. Primarily, each community member setup his/her own rules for privacy settings, i.e. whom to invite to his own community, and who can have access which statistics; non-community members can only see a user’s name. Secondly, by enabling users to invite each other to their communities, QoSIS.net can benefit from the fact that users in a community are very likely to be a group of closely related people (family or friends) in close geographical location/in given time. Therefore they collect overlapping QoS-information, which increases QoS-predictions service accuracy. Thirdly, belonging to such a community can be a trendy lifestyle choice (like for Eric) supporting his daily activities.

5.3 Potential customers

Mobile service users like Eric are potential customers of QoSIS.net. QoSIS.net can target small number of users living in little village of city suburbs with limited WNP choice, as well as big number of users living in city center with plentiful of available WNPs. Because their QoE while using their mobile services is related to the QoS-offered by a WNP used, they are interested in use of QoSIS.net’s value-add service in order to facilitate (proactive and automatic) choice of a WNP best for them anytime-anywhere-anyhow. The user grants responsibility to QoSIS.net on how to use QoS-predictions, i.e., when to demand new predictions (i.e. mobility-maps’ update) and when to choose another WNP.

From the users perspective, belonging to a QoSIS.net community can be a trendy lifestyle choice, for which they are willing to pay the price in terms of mobile device resources assigned for QoSIS.net execution (battery, capacity, storage etc.). The goal of community is to collectively empower it users in a WNP choice, fulfills a mobile service user’s need of knowledge of QoS-offered by different WNPs at a given location/time. However, what can be the most important objective for a mobile user driving him to use QoSIS.net, is the fact, that belonging to QoSIS.net community can also addresses his human belonging, gaining prestige, fulfilment and recognition needs (Hoegg et al., 2006). Moreover, an additional trustworthiness may arise in a user, if somebody who the user already knows is in; the user is very likely to accept invitation to QoSIS.net from his friends/family members, who is already in.

A QoSIS.net user saves resources involved in e.g. Facebook.com services delivery e.g. money while choosing WNP with cheaper tariffs, or in terms of saving mobile device’s battery life. The incentive for a mobile user to use QoSIS.net is an improvement of his QoE when using his mobile services. Without existence of QoSIS.net, Facebook.com mobile user like Eric would use any available WNP in her location and time and experience lower quality services.

5.4 Value chain

As in B2B scenario, we only present parties that play roles in the context of a QoSIS.net’s service interaction that provide value to Facebook.com user as

QoSIS.net's customer. Therefore, in the value chain we distinguish QoSIS.net, Facebook.com and Facebook.com's user – Eric (Figure 3). QoSIS.net is a value-add SP for Facebook.com user, and their *business relation* is defined in terms of contract and service usage relationship. Particularly, QoSIS.net is a provider of value-add services to Facebook.com user (i.e. a *subscriber*). These two parties have a *contract relationship* (a SLA), i.e., a formal contract defining the terms and conditions for the delivery of the services and detailed services' specifications. These terms and conditions need to be agreed upon the user, before the QoSIS.net application can be used by Eric, i.e., installed on his PDA. Moreover, it is important to notice that it is responsibility of Eric to make sure that by using QoSIS.net he does not violate SLA existing between his and the Facebook.com, especially with regards to the (default) WNP use or use of any mobile device resources, for which these two services would now compete.

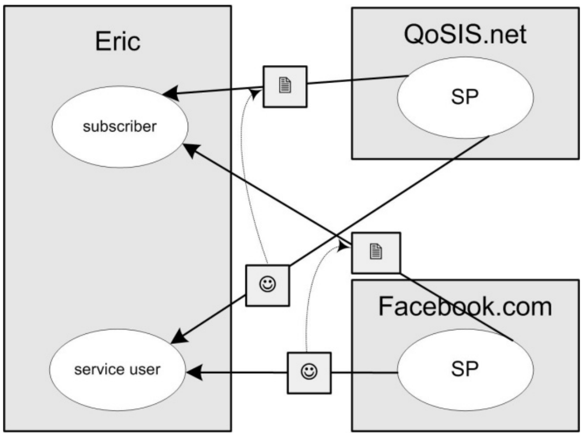


Fig.3. QoSIS.net as a Service Provider

QoSIS.net's service *user* is Facebook.com user – Eric. QoSIS.net's service is seamlessly and transparently delivered to him while on the move. Anywhere he is, his PDA always uses (or handovers to) a WNP, which best assures meeting the QoS-required by his social-networking service and his preferences like WNP usage monetary cost. The WNP choice is made based on information provided by the QoS-predictions service and this choice can be completely transparent to him. Moreover, along Facebook.com's services delivery, the QoS-monitoring and QoS-processing service are provided continuously, as they are based on the QoS-information about the QoS observed by Facebook.com services using different WNPs.

5.5 The impact on existing value chains

The Facebook.com does not need to be in business relationship with QoSIS.net, but it will profit from services provided by it - by means of increased user-QoE, which then can drive an increased Facebook.com's (or any other MoSPs') revenue.

5.6 B2C-specific revenues-cost model

We envision that QoSIS.net starts as a free service by mobile users for mobile users. The QoSIS.net can generate revenues from community-based (e.g. location-, or community-interests-based) advertising on its website (as it happens in case of Facebook or YouTube today). The QoSIS.net marketing costs would be little, as major advertisement would be "word-of-mouth" going from user to user, and from community to community. Additionally to maintenance costs of QoS-information base, new costs of management of community-databases are added.

5.7 Supporting services

As for the B2B scenario, from a business perspective, all QoSIS.net's service: QoS-monitoring, QoS-processing and QoS-predictions are necessary to create a value to the QoSIS.net's service user. Differently from B2B scenario, in here, to support the QoS-monitoring service delivery, QoSIS.net instruments user's mobile device for acquisition of QoS-information (using application similar to AcbTaskMan (AcbTaskMan, 2007) or CoSphere (Peddemors, 2008)). QoSIS.net need to decide upon the QoS-monitoring service delivery; how often the data collected by a mobile user is going to be acquired by QoSIS.net, e.g. depending on change of user location, time passed or a WNP change. QoSIS.net needs also to instrument user's mobile device for using of the QoS-predictions while choosing a WNP, and for an enforcing of this choice.

5.8 Further evolvements of the business method in B2C scenario

After having a critical mass of users, QoSIS.net can change its revenues model and can charge its users per a transaction (i.e. per a single QoS-predictions delivery) or it can introduce monthly subscription fee for premium-users (anticipating particular QoS-predictions usage). The transaction fee can be intentionally a) low for low-accuracy predictions or b) high for high-accuracy predictions. The accuracy in turn can be checked against the information acquired QoS-monitoring service. Transaction fee can also depend on number of WNPs, a user can choose from at a given location and time; the higher the number the higher the fee. It is also possible that user is offered e.g. two free transactions per day, and pays for any additional one.

QoSIS.net can also provide QoS-monitoring and QoS-predictions services as separate services, and reward a user of QoS-monitoring service (i.e. producing QoS-information) and charge a user for using QoS-predictions service (i.e. consuming QoS-information). This however brings to QoSIS.net a risk of not having enough contributing users; so far statistics for Web 2.0 indicate that only 1 % of service users (!) is willing to contribute and generate the content (Arthur, 2006).

Moreover, QoSIS.net can also have revenues from selling critical information like user profiles, statistical information about QoS-offered by different WNPs or most frequently used mobile services, to MoSPs and WNPs or any other interested parties. QoSIS.net can also launch an affiliation program, rewarding users who helped to acquire a new QoSIS.net customer via their website.

To tackle the scalability requirement, similarly to B2B scenario, we propose QoSIS.net to be a location-based service, i.e. limiting its scope of operation to particular city, region or country, hence limiting the scope of WNPs for which QoS-predictions can be provided. Scoping can be dictated by need of limitation of QoS-information to be processed, or need for higher accuracy of QoS-predictions provided for a limited geographical-area. Moreover, web-based communities build by QoSIS.net users can use different language or can have different representation of their data depending on privacy regulations in given country or their cultural background. In the case of different QoSIS.net' location-based instances, they can form a (short-term or a long-term) "smart-business-networks" (van Heck & Vervest, 2007) for a purpose of delivering better services to their users roaming in between locations areas belonging to different QoSIS.net; all QoSIS.net instances collaborating in the business network could benefit from larger QoS-information base for their QoS-predictions service.

The other evolvement of B2C we envision is related to the situation, where a user would like to use QoSIS.net, however, is not a frequent MoSP's service user (e.g. like Eric using his Facebook.com daily) and hence will not generate lots of 'real' mobile traffic. This kind of mobile user may be proposed using a QoSIS.net QoS-monitoring service as his mobile phone 'screen-saver'. Namely, at the moments when he would not use his mobile for a while, the QoS-monitoring service would take a role of active mobile service user - using a WNP (and imitating e.g. busty web-browsing) and acquiring QoS-monitoring information at a given location/time. In such a way, QoSIS.net would acquire QoS-information enriching its databases and such a mobile user could be paid for information generation. However, the critical issue would be related to mobile device resources usage for QoSIS.net QoS-monitoring service (battery, capacity, storage etc.).

6 Potential of QoSIS.net for WNPs/MNOs

The proposed B2C and B2B scenarios are different from existing WNP-centered (i.e. MNOs-centered) business models, as they aim to empower the MoSP and its mobile users to use best WNP, and not the one the one user is locked-in. The proposed B2B and B2C business scenarios have a possible impact on WNP's business models by means of unlocking the users from WNPs and disclosing publicly the information about their offered-QoS and thus increasing the competitiveness amongst WNPs. In our scenarios, the WNP is envisaged to have revenues coming from the QoS-information exchange (e.g. per MB), the WNP's role is assumed to be passive as a so-called 'bit-pipe'. It is however equally possible that a MoSP, as a QoSIS.net customer, is itself in a business relation with a WNP (or a set of WNP), who e.g. get additional revenues whenever MoSP user requests service delivery. In such settings, revenues of WNP would be increased along the improvement of MoSP's user-QoE and thus (expected) increase of mobile service use by this user.

We would like to emphasize that there exists a huge business potential for all parties in case where QoSIS.net QoS-information databases collected for a given WNP would be made available for this WNP. Namely, if WNP considers QoSIS.net as trustworthy business enterprise, it would analyze its offered-QoS (in terms of availability, speed, security level and monetary cost) at given locations and time and be willing to improve this QoS, in order to stay competitive amongst the WNPs. Moreover, once encouraged, a WNP can even provide QoSIS.net with some additional information regarding its network configuration, which would facilitate more accurate QoS-predictions service. We envision that improvements of the QoS-offered by WNP to MoSPs, as well as improvements in QoS-predictions service will improve mobile users'-QoE. This in turn will encourage MoSPs users to use even more their mobile services while on the move. This situation can be highly beneficial for all parties, as it has a potential to increase revenues to MoSPs, as well as to WNPs, and create revenues for QoSIS.net while improving a mobile service user's experience.

Regarding other possible business scenarios, it is also possible for WNP/MNO to have an active and dynamic role if: a) QoSIS.net's become a 3rd Party SP to WNP/MNO or b) WNP/MNO takes a role of QoSIS.net. In both cases QoSIS.net's service are provided i) to users of mobile services provided directly by WNP/MNO (i.e. being a MoSP) or ii) to MoSPs (and their users) that have business relationship with WNP/MNO. Due to space limitations, we do not further analyze these scenarios in this paper, however we consider them elsewhere (Wac, 2008).

7 Concluding Remarks and Future Trends

As we have presented in this chapter, mobile and ubiquitous service providers (MoSPs) emerge, struggling to provide their users with QoE at least comparable to one the user is familiar with from using Internet-based (fixed) services. All this happens because wireless communications' infrastructures, supporting delivery of these services, neither provide QoS guarantees, nor disclose information of their offered-QoS. To bridge the gap regarding the lack of information about QoS-offered by different WNPs in mobile users' location and time, in this chapter we propose business methods enabling firstly a creation of enterprise (QoSIS.net) providing such an information to MoSPs, and, secondly - usage of this information by MoSP in its QoS-assurance business process. We emphasize that the aims are to make MoSPs QoS-assurance process ubiquitous and competitive (i.e., efficient and effective) and to increase MoSP's user-QoE, hence to increase revenues to MoSP, while creating revenues to QoSIS.net. We currently attempt to implement the QoS-predictions service of QoSIS.net together with MobiHealth.com as a MoSP (Pawar et al., 2008; Wac et al., 2008b).

Future research opportunities within the domain of our topic relate to further evolvements of the proposed business methods for QoSIS.net and its customers. These research opportunities are related firstly to understanding the dependencies between WNP offered-QoS, the MoSP user QoS-requirements and his QoE, for MoSPs in different application domains. This can then serve, at least partially, as a basis for deriving detailed requirements posed on the QoSIS.net's QoS-predictions service by MoSPs (as customers and service users). These requirements can be expressed, for example in terms of service availability, accuracy and delay. Related to this, second research opportunity focuses on the efficient and effective market entry approach for QoSIS.net as an enterprise and its possible innovative marketing solutions; and this for overcoming its initial hurdle of attracting critical mass of users in order to be able to provide accurate QoS-predictions service and start generating revenue. Third research opportunity lies in understanding partnership-trust (Ratnasingam & Phan, 2003) required in the QoSIS.net's value chain, as well as challenges in QoSIS.net's customer management in B2B and B2C market segments. Fourthly, due to the nature of service provided by QoSIS.net, we indicate a need for research on trust in technology (Ratnasingam et al., 2002). This research investigates, on one hand, dependability features of architectural system design. On the other hand, this research investigates entries necessary in business contracts established between parties, in order that these contracts encompass business practices for possible technological scenarios endangering core business processes and revenues of involved parties. This will e.g. include research on security mechanisms employed in QoS-information exchange between parties.

Future research opportunities along the books theme relate particularly to research upon new competitive business methods that can be employed in existing management, operational (i.e. core) or supporting business processes of mobile and ubiquitous service providers as business enterprises. These business methods need necessarily aim in delighting their customer, while increasing their revenues. We propose that these business methods are based on emerging trend of short- and long-terms business inter-dependencies (i.e. "smart-business-networks") between different enterprises, bring into a value network different but complementary, expertise. This, on one hand, brings high risk, but on the other hand has a huge potential to substantially increase revenues of all of the involved parties, and that by increasing a customer experience anywhere-anytime-anyhow. Moreover we would like emphasize a risk, but also huge potential of employing of user-collaborative-content-sharing paradigm, i.e. Mobile Web 2.0, as a base for business methods employed in enterprise's core business processes. Its risk is mainly related to requirement of attracting a critical mass of contributing users, which may not always be easy. These methods would aim at creating new revenue streams from user-generated content-manipulation and enrichment. The enriched content could be then a part of enterprise service, consumed back by users. The Mobile Web 2.0-based methods however require careful research upon the content type to be generated and consumed by users. Therefore, answer for research questions like what is the pre-existing offline information possessed by users, which, if enabled to be manipulated and shared online amongst them, could empower them in some way?, as well as if this information violates in any way user's privacy?, and what is user's willingness to share this information online and with whom?, are critical for the success of the business method. Moreover, research upon design of the offered service, careful market analysis and management of the enterprise start-up phase, and so on, is necessary.

We envision that in order to fulfill the dream of novel successful services offered by mobile and ubiquitous commerce, and to increase their revenue, enterprises need to take a necessary risk and employ such novel business methods on a growing scale in their business processes.

References

[1] AcbTaskMan. (2007). AcbTaskMan software. Retrieved April 13, 2008, from www.acbpocketsoft.com

[2] Afuah, A., & Tucci, C. T. (2000). *Internet Business Models and Strategies: Text and Cases*: McGraw-Hill Higher Education.

[3] Andersen, D., Bansal, D., Curtis, D., Seshan, S., & Balakrishnan, H. (2000). *System Support for Bandwidth Management and Content Adaptation in Internet Applications*. Paper presented at the 4th Symposium on Operating Systems Design and Implementation (OSDI), San Diego, CA, USA.

[4] Andersson, C., Freeman, D., James, I., Johnston, A., & Ljung, S. (2006). *Mobile Media and Applications, From Concept to Cash: Successful Service Creation and Launch*. West Sussex, England: Wiley.

[5] Arthur, C. (2006). What is the 1% rule? Retrieved April 13, 2008, from http://technology.guardian.co.uk/weekly/story/0,1823959,00.html

[6] Bless, R., Hillebrand, J., Prehofer, C., & Zitterbart, M. (2004). Quality-of-Service Signaling for Next-Generation IP-Based Mobile Networks. *IEEE Communications Magazine, 42*(6), 72-79.

[7] Bults, R., Wac, K., van Halteren, A., Konstantas, D., & Nicola, V. (2005). *Goodput Analysis of 3G wireless networks supporting m-health services.* Paper presented at the 8th International Conference on Telecommunications (ConTEL05), Zagreb, Croatia.

[8] Buschken, J. (2004). *Higher Profits Through Customer Lock-In*: Thomson Texere.

[9] Busse, M., Lamparter, B., Mauve, M., & Effelsberg, W. (2004). *Lightweight QoS-support for networked mobile gaming.* Paper presented at the 3rd ACM SIGCOMM workshop on Network and system support for games (NetGames04), Portland, OR, US.

[10] Calvo, M., Rodríguez, C., & Dillinger, M. (2004). *Business models for reconfigurable communication systems.* Paper presented at the 13th IST Mobile & Wireless Communications Summit, Lyon, France.

[11] Camponovo, G., & Pigneur, Y. (2003). *Business model analysis applied to mobile business.* Paper presented at the 5th Intl Conference on Enterprise Information Systems (ICEIS03), Angers, FR.

[12] Chalmers, D., & Sloman, M. (1999). A survey of Quality of Service in mobile computing environments. *IEEE Communications Surveys and Tutorials, 2*(2).

[13] Chen, W. T., & Shu, Y. Y. (2005). *Active application oriented vertical handoff in next-generation wireless networks.* Paper presented at the Wireless Communications and Networking Conference (WCNC05).

[14] claffy, k., Miller, G., & Thompson, K. (1998). *The nature of the beast: recent traffic measurements from an Internet backbone.* Paper presented at the International Networking Conference (INET98), Geneva, Switzerland.

[15] Cuevas, A., Moreno, J. I., Vidales, P., & Einsiedler, H. (2006). The IMS Platform: A Solution for Next Generation Network Operators to Be More Than Bit Pipes. *IEEE Commun. Mag., Advances in Service Platform Technologies, 44*(8), 75-81.

[16] De Vriendt, J., Laine, P., Lerouge, C., & Xu, X. (2002). Mobile Network Evolution: A Revolution on the Move. *IEEE Commun. Mag., 40*(4), 104-111.

[17] Dekleva, S., Shim, J. P., Varshney, U., & Knoerzer, G. (2007). Evolution and emerging issues in mobile wireless networks. *Commun. ACM, 50*(6), 38-43.

[18] Digital Chocolate. (2008). Seize the Minute. Retrieved April 13, 2008, from www.digitalchocolate.com

[19] Dijkstra, S. J., Jurriens, J. A., & van der Mei, R. D. (2006). *A Business Model for Telemonitoring Services.* Paper presented at the High 14th Technology Small Firms Conference, University of Twente, NL.

[20] Faber, E., Ballon, P., Bouwman, H., Haaker, T., Rietkerk, O., & Steen, M. (2003). *Designing business models for mobile ICT services.* Paper presented at the 16th BLED Electronic Commerce Conf. - eTransformations, Bled, Slovenia.

[21] Facebook. (2007). A social utility that connects you with the people around you. Retrieved April 13, 2008, from www.facebook.com

[22] Gomez, G., & Sanchez, R. (2005). *End-to-End Quality of Service over Cellular Networks: Data Services Performance Optimization in 2G/3G*: John Wiley & Sons, Ltd.

[23] Gorlach, A., Heinemann, A., & Terpstra, W. (2004). *Survey on location privacy in pervasive computing.* Paper presented at the Workshop on Security and Privacy in Pervasive Computing (SPCC04) at PERVASIVE2004.

[24] Han, Q., & Venkatasubramanian, N. (2006). Information Collection Services for QoS-aware Mobile Applications. *IEEE Transactions on Mobile Computing, 5*(5), 518-535.

[25] Hansmann, U., Merk, L., Nicklous, M., & Stober, T. (2003). *Pervasive Computing: The Mobile World*: Springer.

[26] Heckerman, D. (1996). *A Tutorial on Learning With Bayesian Networks* (No. MSR-TR-95-06). Redmond, Washington, US: Microsoft Research.

[27] Henricksen, K., Indulska, J., McFadden, T., & Balasubramaniam, S. (2005). *Middleware for Distributed Context-Aware Systems.* Paper presented at the On the Move to Meaningful Internet Systems 2005, Agia Napa, Cyprus.

[28] Hoegg, R., Martignoni, R., Meckel, M., & Stanoevska-Slabeva, K. (2006). *Overview of business models for Web 2.0. communities.* Paper presented at the GeNeMe, Dresden, DE.

[29] Hutchison, D., Mauthe, A., & Yeadon, N. (1997). Quality-of-service architecture: Monitoring and control of multimedia communications. *Electronics & Communication Engineering Journal, 9*(3), 100.

[30] ITU-T. (1993). General aspects of Quality of Service and Network Performance in Digital Networks, including ISDNs (Vol. I.350): ITU.

[31] ITU-T. (2006). Framework for achieving end-to-end IP performance objectives (Vol. Y.1542): ITU.

[32] ITU-T. (2007). Vocabulary for performance and quality of service: Appendix I – Definition of Quality of Experience (QoE) (Vol. P.10/G.100): ITU-T.

[33] Manner, J., Burness, L., Hepworth, E., Lopez, A., & Mitjana, E. (2001). *Provision of QoS in heterogeneous wireless IP access networks.* Paper presented at the Intl Symposium on Personal, Indoor and Mobile Radio Communications.

[34] Martignoni, R., & Stanoevska-Slabeva, K. (2007). *Mobile Web 2.0.* Paper presented at the 20th BLED Electronic Commerce Conf. - eMergence, Bled, Slovenia.

[35] Michaut, F., & Lepage, F. (2005). Application-oriented network metrology: Metrics and active measurement tools. *IEEE Communications Surveys & Tutorials, 7*(2), 2-24.

[36] MobiHealth. (2007). Putting care in motion. Retrieved April, 13, 2008, from www.mobihealth.com

[37] O'Reilly, T. (2005). What is Web 2.0? Retrieved April 10, 2008, from www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[38] Ortiz, S. (2007). 4G Wireless Begins to Take Shape. *IEEE Computer, 40*(11), 18-21.

[39] Pascu, C., Osimo, D., Ulbrich, M., Turlea, G., & Burgelman, J. C. (2005). The potential disruptive impact of internet 2 based technologies. *First Monday - peer-reviewed Journal on the Internet.*

[40] Pawar, P., Wac, K., van Beijnum, B. J., Maret, P., van Halteren, A., & Hermens, H. (2008). *Context-Aware Middleware Architecture for Vertical Handover Support to Multi-homed Nomadic Mobile Services*. Paper presented at the 23rd Annual ACM Symposium on Applied Computing (ACMSAC08), Ceará, Brazil.

[41] Peddemors, A. (2008). CoSPhere NAL software. Retrieved April 13, 2008, from http://cosphere.telin.nl/nal

[42] Ratnasingam, P., Pavlou, P., & Tan, Y. (2002). *The Importance of Technology Trust for B2B Electronic Commerce*. Paper presented at the 15th BLED Electronic Commerce Conf. - eReality: Constructing the eEconomy, Bled, Slovenia.

[43] Ratnasingam, P., & Phan, D. (2003). Trading Partner Trust in B2B E-Commerce: A Case Study. *Information Systems Management, 20*(3), 39-50.

[44] Robles, T., Mitjana, E., & Ruiz, P. (2002). *Usage scenarios and business opportunities for systems beyond 3G*. Paper presented at the IST Mobile and Wireless Telecommunications Summit 2002, Thessaloniki, GR.

[45] Sanchez, A., Carro, B., & Wesner, S. (2008). Telco Services for End Customers: European Perspective. *IEEE Commun. Mag., 46*(2), 14-18.

[46] Shepherd, D., Scott, A., & Rodden, T. (1996). Quality-of-Service Support for Multimedia Applications. *IEEE MultiMedia, 03*(3), 78-82.

[47] Soh, W. S., & Kim, H. S. (2003). QoS Provisioning in Cellular Networks Based on Mobility Prediction Techniques. *IEEE Commun. Mag., 41*(1), 86-92.

[48] Tachikawa, K. (2003). A Perspective on the Evolution of Mobile Communications. *IEEE Commun. Mag., 41*(10), 66-73.

[49] Tan, S. (2004). *Evolution of mobile technology and business models (technical report)*: Center for Information and Communication Technologies, Lyngby, DK.

[50] TMF. (2004). *Shared Information/Data (SID) Model; GB922 Addendum 0 - SID Primer (ver.4.0)*. Morristown, New Jersey, US: TeleManagement Forum.

[51] Tsalgatidou, A., & Pitoura, E. (2001). Business models and transactions in mobile electronic commerce: requirements and properties. *Computer Networks, 37*(2), 221-236.

[52] van Halteren, A., Bults, R., Wac, K., Konstantas, D., Widya, I., Dokovsky, N., et al. (2004). Mobile Patient Monitoring: The MobiHealth System. *The Journal on Information Technology in Healthcare, 2*(5), 365-373.

[53] van Heck, E., & Vervest, P. (2007). Smart business networks: how the network wins. *Commun. ACM, 50*(6), 28-37.

[54] Wac, K. (2006). *QoS-predictions service: infrastructural support for proactive QoS- and context-aware mobile services*. Paper presented at the On the Move to Meaningful Internet Systems 2006: OTM Workshops, Intl Workshop on Context-Aware Mobile Systems (CAMS), Monpellier, France.

[55] Wac, K. (2008). *Application-level Performance Management for Mobile Services*. Unpublished doctoral dissertation, University of Geneva, Geneva, CH.

[56] Wac, K., Bults, R., & Broens, T. (2008a). A business model for mobile health monitoring services: MobiHealth case study (forthcoming). *Intl Journal of Electronic Commerce*.

[57] Wac, K., Hilario, M., Konstantas, D., & van Beijnum, B. J. (2008b). Data Mining on Application-level QoS Traces: the MobiHealth System Case Study (forthcoming). *IEEE Trans. on Mobile Computing*.

[58] Xiao, X., & Ni, L. M. (1999). Internet QoS: a big picture. *IEEE Network, 13*(2), 8-18.

Power- and Delay-Aware Mobile Application-Data Flow Adaptation: the MobiHealth System Case Study[1]

Katarzyna Wac, Mortaza Bargh, Arjan Peddemors, Pravin Pawar, Bert-Jan van Beijnum, Richard Bults

Abstract. Emerging healthcare applications rely on personal mobile devices to monitor patient vital signs and to send it to the hospitals-backend servers for further analysis. However, these devices have limited resources that must be used optimally in order to meet the requirements of healthcare applications end-users: healthcare professionals and their patients. This paper reports on a case study of a cardiac telemonitoring application delivered by the so-called MobiHealth system. This system relies on a commercial device with multiple (wireless) network interfaces (NI). Our study focuses on how the choice of a NI affects the end-to-end application's data delay (extremely important in case of patient's emergency) and the energy consumption of the device (relating to the service sustainability while a patient is mobile). Our results show the trade-off between the delay and battery savings achieved by various NI activation strategies in combination with application-data flow adaptation. For a given mobile device, our study shows a gain of 40-90% in battery savings, traded against the higher delays (therefore applicable mainly in non-emergency cases). The insights of our studies can be used for application-data flow adaptation aiming at battery saving and prolonging device's operation for mobile patients.

Keywords-mobile device connectivity management; energy efficiency; end-to-end delay; application adaptation; mobile healthcare

1 Introduction

The emergence of new wireless broadband networks and the increased diversity of miniaturized and personalized networked devices give rise to a variety of new mobile interactive applications in our daily life. Examples of these are, on one hand, applications supporting traditional users as information-*consumers*, e.g. news, leisure and entertainment content delivery. On the other hand, mobile users are no longer only passive information and content consumers, but on a growing scale they take the role of content *producers*. Examples of these applications are especially ones supporting social networking. However, another emerging application domain, in which a user acts as a content producer, is a mobile healthcare domain, where a mobile patient's vital signs can be telemonitored by his healthcare professional in the healthcare center. In this paper we focus on this application example.

The above mentioned applications are ultimately envisaged to be delivered to the user on the move: anywhere anytime and under different conditions, while

[1] This work is part of the Dutch Freeband AWARENESS project (http://awareness.freeband.nl, contract BSIK 03025)

fulfilling his *Quality of Service* (QoS) *requirements*. These requirements are, e.g., low application delays, long device battery life and seamless user mobility support along with low monetary cost of networks usage. However, as applications operate in a heterogeneous networking environment, consisting of a variety of wireless and wired networks owned by different parties, the QoS provided by this environment is one of the most critical factors influencing the assurance of the QoS provided by the application to the user. In this paper, the *QoS provided* by an *application* is defined as an *application-level throughput* (in kbps) and an *application-level delay* (in milliseconds).

There exists close relation between the provided application-level QoS and the provided network-level QoS. Particularly, the provided application-level throughput and delay depend respectively on throughput and data delay while using particular underlying (wireless) network over the given network interface (NI) on the mobile device. Moreover, the device battery life depends on a given application, given NI, and on how *application-data flow* is offered to this NI. Particularly, this flow is described in terms of its volume, i.e., size and rate of the data offered to the NI. By changing the size and the rate parameters we change volume of data to be sent; in such a way we can adapt the application-data flow to suit better the provided network-level QoS and to obtain better application-level QoS.

This paper focuses on 1) an choice of NI (as available on a mobile device) and its activation strategy (ON/OFF) and 2) an application-data flow adaptation, and relations of these two with a) a device's energy consumption and b) an application-data delay. In this paper we study the relation of these four parameters to the user's required QoS for a health telemonitoring application [1], and particularly, cardiac telemonitoring application delivered by the so-called MobiHealth system [2].

The rest of this paper is organized as follows. Section 2 provides a description of the MobiHealth system, while 3 - a mobile device's NI states and their selection criteria. Section 4 provides our measurement methodology for energy and delay measurements for a commercial mobile device used in the MobiHealth system. Section 5 analyzes the measurement results, based on which we defined NI activation strategies. R Section 6 discusses related work. Based on measurements results, in Section 7 we provide the conclusions and recommendations for the MobiHealth system usage and some future work areas.

2 The MobiHealth system

2.1 System Overview

The MobiHealth system is a distributed system for telemonitoring of a patient's health condition.

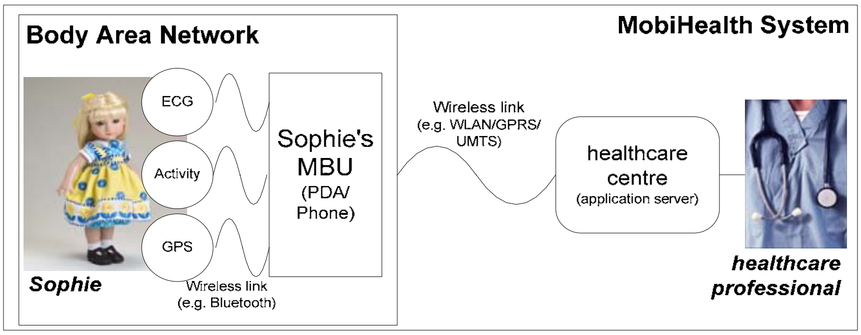

Body Area Network
MobiHealth System
ECG
Activity
GPS
Sophie's MBU (PDA/ Phone)
Wireless link (e.g. WLAN/GPRS/ UMTS)
healthcare centre (application server)
Sophie
Wireless link (e.g. Bluetooth)
healthcare professional

Figure. 1 MobiHealth system overview.

In the MobiHealth system (Fig. 1), a patient is wearing a *Body Area Network* (BAN), consisting of a sensor-set and a *Mobile Base Unit* (MBU). The sensor-set consists of specialized sensors monitoring the patient's vital signs, an alarm button, to be pressed by the patient in emergency, and a location-determination sensor (e.g. a GPS). The sensor-set is specific for a patient's health condition, e.g. respiration insufficiency, cardiac problems, and epilepsy. Emergency condition is defined individually for each patient, based on his/her health condition. It can be activated based on a) patient's pressing the alarm button or, based on the patient's vital signs analysis on b) the BAN or c) the backend-server.

The MBU is the central unit of a BAN, usually in the form of a mobile phone or PDA. It continuously collects sensor data, processes it (e.g. filters, shapes, correlates) and sends in real-time to a remote application *backend-server* located in a healthcare center, where it can be made available for e.g., medical decision support systems.

The BAN uses the intra-BAN communication network, e.g. Bluetooth (BT) to send data from the sensor-set to the MBU, and an extra-BAN communication network, e.g. WLAN or 2.5G/3G (i.e. GPRS/UMTS) for exchange of the application and control data between the MBU and the backend-server.

The application execution is supported by a proprietary *MSP-Interconnect protocol (MSP-IP)* [3]; a TCP/IP-stack-based protocol, facilitating application-data-plane and control-plane[2] data. The overall system architecture conforms the Jini Interconnect specifications as we presented in [4]. Detailed description of the MobiHealth system we presented in [2, 5].

2.2 Telemonitoring Application-Data Flow

In this paper, we consider the telemonitoring application for cardiac patients in a non-critical condition, i.e., with a small probability of an emergency. Hence we consider application-data flow adaptation cases separately for an emergency and non-emergency (Section 2.C).

[2] BAN control-plane data consists of the MBU management lifecycle and aliveness (Keep-Alive) messages

The sensor-set acquires patient's heartrate (HR), oxygen saturation (SO_2) and plethysmogram (pleth), an alarm button state, and a control-data. Sensors sampling frequency is 128Hz; each sample consists of a 5B of application-data. A unit of data that the application collects consists of 1s aggregated sensor-set data, so, in total of 640 B. Every unit of data is compressed (lossless) before being sent by the extra-BAN communication network. The data compression factor, i.e., the reduction in size relative to the uncompressed size, is 80-85 %. However, this factor strongly depends on the actual values of the measured vital signs; it decreases with increased variability in measured vital signs. The MSP-IP introduces a 10 B overhead per an aggregated and compressed data. Hence, the protocol stack overhead is 64 B for WLAN (MSP-IP/TCP/IP/Ethernet) and 58 B for GPRS (MSP-IP/TCP/IP/PPP). The resulting data unit is sent over the data-plane. The overall volume of data sent by the NI contains data-plane and control-plane[3] data; ~1.2-1.5 kbps[4].

2.3 QoS Requirements

The end-users of the telemonitoring applications are healthcare professionals are their patients. Only the former ones are in charge to define the QoS requirements [6]. These requirements are related to application-data exchange performance a) its reliability (lossless and error-free) and b) a minimum application-data delay (in case of a patient's emergency) from the sensor-set to the backend-server. The use of TCP/IP protocol stack and the use of local data storage in case when no network is available to send the data or a real-time sending is not required, ensure system recovery in case of data loss and encountered data-errors. Further study of application reliability is outside of scope of this paper.

Concerning the application-data delay requirement, we focus on the extra-BAN data delay, as a major contributor to the application-data delay in the MobiHealth system. Particularly the MobiHealth system performance is managed based on an *application-level Round Trip Response* (AppRTT)[5] times. The AppRTT is a time period it takes for a control message (i.e., a MBU Keep-Alive[6] [4]) originated from the MBU, to be bounced by the backend-server (without being processed there) and received back by the MBU. AppRTT strongly depends on the choice of the extra-BAN communication network, i.e. the NI choice at the MBU, and the volume of the application-data being sent. Moreover, The AppRTT reflects the delays provided by the underlying networks, as it is composed of the processing delays in the protocol stacks at the MBU and the backend-server side as well as an uplink (MBU to the backend-server) and downlink (backend-server to the MBU) network delays.

[3] of a negligible size comparing to the data-plane

[4] Data-plane calculation for compression factor of 80%: a) WLAN: [(640 *0.2)+64]*8 bps = 1536 bps b) GPRS: [(640 *0.2)+58]*8 bps = 1488 bps;

Data-plane calculation for compression factor of 85%: a) WLAN: [(640 *0.15)+64]*8 bps = 1280 bps b) GPRS: [(640 *0.15)+58]*8 bps = 1232 bps

[5] Reliable one-way delay measurements are only possible if the clocks of MBU and backend-server would be synchronized; it is hardly feasible in the operational system

[6] KeepAlive message of size of 41 B

The considered cardiac telemonitoring application's delay requirements strongly depend on the actual health condition of a patient. In emergency, patient vital signs data needs to be continuously sent (at a minimum possible delay) to the backend-server, where it is made available for a healthcare professional in real-time. For non-emergency, it is possible that the MBU acquires the application-data (for data-plane and control-plane), stores it locally, and sends it to the backend-server later (i.e., in bursts), e.g. when a cheap, high-throughput WLAN network is available. It is also possible that in non-emergency, the (real-time) BAN data is sent continuously to the backend-server along with the previously stored BAN data.

Another QoS requirement for MobiHealth is a maximum lifetime of the BAN. In this paper we focus on the MBU's power consumption for extra-BAN communication, as its contribution to the BAN's power consumption. We denote the MBU power consumption as $power_{MBU}$. It depends on the NI used for extra-BAN communication and the volume of the application-data being sent.

In our study we also consider an additional user's requirement resulting from the fact that a patient needs to use his MBU as a regular (WWAN) phone and needs to be WWAN-*reachable*, especially by his healthcare professional, for voice/data communication. Assurance of this requirement may not be favorable from the power perspective[7]; however in our study we consider this requirement.

We note additionally, that the MBU power consumption depends also on a user's mobility level and the MBU configuration parameters e.g. backlight level, other running applications, or MBU location with respect to the network's access point/base station (i.e., MBU's received signal strength). However, in our study, we assume that a patient wearing BAN is in a fixed location (i.e. not being mobile).

3 Network Interfaces Activation

3.1 Network Interface States

The existing wireless technologies accessible by commercial mobile devices can be divided into two categories: WWANs that provide a low-throughput and high-delay service over a wide geographic area (e.g. GPRS or UMTS) and WLANs that provide a high-throughput and low delay service over a narrow geographic area (e.g. WiFi) [7]. We consider a device NI state model for mobile devices with GPRS or/and UMTS as WWAN interface and WiFi as WLAN interface. A NI is in one of the following states: 1) OFF 2) ON-IDLE: an IP-idle state, where the mobile device has IP connectivity to the Internet. However it does not send/receive application level data-plane or control-plane IP packets (i.e., IP packets carrying application-data) or 3) ON-ACTIVE: an IP-active state, where mobile device is sending or receiving application level IP packets through this NI.

From the telemonitoring application perspective, application data can be send 1) via the WLAN and WWAN NIs 2) via the WLAN NI, while the WWAN NI is OFF or ON-IDLE or 3) via the WWAN NI, while the WLAN NI is OFF or ON-

[7] Additional power is consumed if WWAN is in ON-IDLE state without sending data

IDLE or 4) or being stored locally, while the WWAN (or WLAN) NI is OFF or ON-IDLE. Note that the NI that is used to send data could be in ON-ACTIVE state continuously or could alternate between the ON-ACTIVE and the ON-IDLE/OFF states (the latter implies data send in bursts).

3.2 Criteria for a Choice of a NI State

Based on the scope of our study and on the requirements posed by the MobiHealth users (Section 2.C), we conclude that a NI state (i.e., OFF/ON-IDLE/ON-ACTIVE) depends on criteria (i) application-data delay requirement posed by the current health condition of a patient (i.e., emergency or non-emergency) (ii) the $\text{power}_{\text{MBU}}$ consumption while using that NI and (iii) the provided AppRTT while using that NI.

4 Measurements

4.1 MobiHealth System Setup

The MobiHealth sensor-set is based on Mobi5-3e1as [8], with only the NONIN finger clip attached (for HR, SO_2 and pleth data). As a MBU we have used Qtek 9090 with Intel® PXA263 400 MHz processor (32b), 128 MB RAM, firmware version 1.31.00 WWE (from 13.12.2004), radio version 1.06.02, protocol version 1337.38 running Windows Mobile 2003 SE PocketPC OS edition version 4.21.1088. The Qtek's battery is of a standard type, rechargeable Li-ion Polymer of capacity of 1490 mAh (3.7V, model PH26B). The Qtek has a TFT touch screen display of size of 53x71 mm (214 x 320 pixels, 65K colors) and its backlight level was set to zero.

The MBU has the WWAN-GPRS (GSM 850/900/ 1800/1900 Hz, class 10: 4+1/3+2 slots) and WLAN-WiFi (IEEE 802.11b, with 'best-battery' setting in the OS) as NIs for extra-BAN communication. The BT NI is used continuously for intra-BAN communication for sensor-set data acquisition. The MBU uses GPRS network provided by Sunrise mobile operator (received signal strength of 100%) and WLAN provided by the University of Geneva, Switzerland (received signal strength of 50%), where the MBU was placed such that the received signal strength has been maximized along the measurements). The backend-server used is a standard high-performance server dedicated to MobiHealth telemonitoring services. The server was placed at Twente University, the Netherlands. The MobiHealth telemonitoring application software version is a release from 17 October 2007.

4.1.2 Power and delay measurements instrumentation

The MobiHealth system was configured such that during the execution of the telemonitoring application, we collected the measurements logs at the MBU and backend-server. To measure the energy consumption of the MBU, we logged the remaining battery percentage in 5 seconds intervals. For the purpose of delay measurements, the MBU was instructed to log the AppRTT in intervals of 10 seconds continuously along the telemonitoring application delivery.

To obtain high application-data flow volumes, not feasible in the current state of the MobiHealth system, and especially important for our measurements over (high-throughput) WLAN NI (cases 3 and 4); we have used the NetPerf application [9]. This application is generating TCP traffic and measuring a unidirectional throughput between the MBU and the backend-server. These measurements were done for the same conditions as the other measurements; however, the MobiHealth application was NOT running in the background of the NetPerf application. By using this application, we attempted to simulate a case where MBU sends previously stored patient vital signs data. In the NetPerf measurements we obtained only the power$_{MBU}$ values.

Along the measurements, we assumed the MobiHealth system to be in the steady-state representing the behavior of the system usage for a typical system user, i.e. a cardiac patient, whose vital signs are being monitored. Measurements have been done over a time span of two weeks, always at same location (our University of Geneva office) but at different hours.

4.2 Measurements Cases

For the purpose of our research, we considered various measurements cases based on the following two parameters: application-data flow and NIs states. Each case represents the combination of states of the MBU WLAN and GPRS NIs (and BT ON-ACTIVE for intra-BAN communication) during our experiments. These cases represent typical execution of a health telemonitoring application in the MobiHealth system, and they are:

0. WLAN OFF, GPRS OFF
1. WLAN OFF, GPRS ON-ACTIVE
2. WLAN ON-IDLE , GPRS ON-ACTIVE
3. WLAN ON-ACTIVE, GPRS OFF
4. WLAN ON-ACTIVE, GPRS ON-IDLE
5. WLAN OFF, GPRS ON-IDLE
6. WLAN ON-IDLE, GPRS OFF
7. WLAN ON-IDLE, GPRS ON-IDLE

Note: Theoretically, it is also possible to have the WLAN in ON-ACTIVE and GPRS in ON-ACTIVE state; however, because this case is not implemented yet in the MobiHealth system (would require substantial application changes) we have not included it in our study.

Case 0 represents application ‘base’ energy consumption, i.e. for intra-BAN communication and the MBU processing and local storage of application-data (no extra-BAN communication). Additionally to this case, cases 5-7 represent application ‘base’ energy consumption increased of energy consumption for maintaining one (or both) NI in an ON-IDLE state. These cases (5-7) represent continuous application execution and local data storage cases, i.e., no data being sent over a NI.

Along the measurements execution we discovered that the case 2, i.e. where GPRS is ON-ACTIVE and WLAN is ON-IDLE was not possible to be executed on the given MBU. It is because the Qtek 9090 is preconfigured such that, if both GPRS and WLAN are available, it will always send data over the WLAN NI rather than leaving the choice of the usage of the NI to the user.

The telemonitoring application-data flow represents the volume of application-data send over the NI that is in the ON-ACTIVE state. Note that our healthcare application produces 1.2-1.5 kbps of data at the NI, i.e., at the physical layer (Section 2.2). The measured application-data volumes are therefore:

- 1.2-1.5 kbps corresponding to the continuous application execution and real-time sending of application-data (used in emergency and non-emergency)
- 5.2 or 7.7 kbps corresponding to the continuous application execution and delayed data send (i.e., sending data in burst, where 4-6 seconds of patient vital signs data are sent together). This can be used only in non-emergency.

Due to the limited processing capacity of Qtek (i.e., experienced system crashes) it was not possible to increase application-data volume beyond 7.7 kbps in case 1 and beyond 5.2 kbps in cases 3 and 4. Therefore we obtained volumes of 1.2-1.5 kbps, 5.2 kbps and 7.7 kbps for case 1, and volumes of 1.2-1.5 kbps and 5.2 kbps for cases 3 and 4.

5 Measurements Results

5.1 MBU Power Consumption (power$_{MBU}$)

We have executed measurements for cases as given in Section 4.2. We measured MBU's remaining battery capacity in percents, and we transformed the results into the *normalized average power consumption values* indicating the decay rate of battery capacity over minutes. These normalized values facilitate comparison of the *relative energy cost* for WLAN and GPRS NIs in a particular device. Tab. 1 summarizes these normalized values for the Qtek device in different NIs states. We observed that in each experiment the remaining battery capacity decreased linearly with time; i.e., the normalized average power consumption values are constant.

The first 4 rows of Tab. 1 represent cases 0, 5-7, in which data was not sent, but locally stored at the device. Rows 1a, 3a and 4a correspond to cases of continuous application execution and sending of application-data in real-time. The other rows (1b, 1c, 3b, 3c, 4b, and 4c) correspond to cases of continuous application execution, but local data storage with delayed sending of data.

TABLE I. NI'S NORMALIZED AVERAGE POWER$_{MBU}$ VALUES

Case No.	**NI's Normalized Average Power$_{MBU}$ Values**	
	Measurement case (Note: BT ON-ACTIVE for all cases)	***Normalized power consumption [1/min]***

Case No.	NI’s Normalized Average Power$_{MBU}$ Values	
	Measurement case *(Note: BT ON-ACTIVE for all cases)*	*Normalized power consumption [1/min]*
0	WLAN OFF, GPRS OFF	0.00092
5	WLAN OFF, GPRS ON-IDLE	0.00487
6	WLAN ON-IDLE, GPRS OFF	0.00568
7	WLAN ON-IDLE, GPRS ON-IDLE	0.00963
1a	WLAN OFF, GPRS ON-ACTIVE (1.2-1.5 kbps)	0.00721
1b	WLAN OFF, GPRS ON-ACTIVE (5.2 kbps)	0.00874
1c	WLAN OFF, GPRS ON-ACTIVE (7.7 kbps)	0.00897
3a	WLAN ON-ACTIVE, GPRS OFF (1.2-1.5 kbps)	0.00873
3b	WLAN ON-ACTIVE, GPRS OFF (5.2 kbps)	0.00911
3c	WLAN ON-ACTIVE, GPRS OFF (NetPerf, 3.45 Mbps)	0.00982
4a	WLAN ON-ACTIVE, GPRS ON-IDLE (1.2-1.5 kbps)	0.00960
4b	WLAN ON-ACTIVE, GPRS ON-IDLE (5.2 kbps)	0.00974
4c	WLAN ON-ACTIVE, GPRS ON-IDLE (NetPerf, 3.95 Mbps)	0.00947

As we observe from Tab. 1, WLAN in ON-IDLE state consumes comparably the same energy as in ON-ACTIVE state (cases 4a, 7). This can be explained by the Qtek configuration, in which we did not instruct it to get into WLAN *power-save* mode when being ON-IDLE state. In this case, the WLAN NI continuously receives and processes all the data broadcasted between the Access Point and other WLAN devices.

Moreover, we conclude from the Tab.1, that from the power perspective, it is always better to use the GPRS NI and keep the WLAN OFF. If WLAN NI is used, it is always better to keep GPRS OFF.

5.2 Application-Data Delay (AppRTT)

We have executed measurements cases as given in Section 4.2 and observed, as we have previously expected, that AppRTT depends on the NI(s) used and the data volume being sent. Tab. 2 summarizes the results, with emphasis on the AppRTT mean value. Note that these results are reported only for the telemonitoring application execution, i.e., not for the cases when we have used the NetPerf.

TABLE II. NI’s AppRTT Values

AppRTT [ms] & case No.	NI’s AppRTT Values				
	mean	*stdev*	*min*	*max*	*med*
1a.	3739	2005	1979	20856	3318

AppRTT [ms] & case No.	**NI's AppRTT Values**				
	mean	***stdev***	***min***	***max***	***med***
1b.	5505	2627	2767	20702	4706
1c.	6693	3954	2322	28220	5589
3a.	2753	1769	530	23807	2706
3b.	3513	2863	587	36819	3290
4a.	1806	1082	556	15756	1553
4b.	2211	1084	379	13609	2204

As we observe, from the delay perspective, the best is, whenever possible, to use a WLAN ON-ACTIVE and keep GPRS ON-IDLE (cases 4a, 4b). If WLAN is not available and it is necessary to use GPRS, it is better to use lower data volumes (case 1a), or, if patient is not in an emergency, gather the data for a local storage, and send it at the maximum possible volume later over WLAN (case 4b). An interesting observation is that for the real-time application-data sending, the GPRS has higher delay but slightly lower delay variation (i.e. stdev value) comparing to the WLAN (case 1a of 53% vs. 3a of 64% of a mean value). Moreover, the WLAN has lower delay and delay variation when GPRS being ON-IDLE (case 4a) than when GPRS being OFF (case 3a). That may be related to the internal NI management of the mobile device used (the real reasons are unknown for us, and to the best of our knowledge similar results have not been published so far).

5.3 NI activation strategies

In this section, we define the *basic* MBU NI activation strategies as ones, in which sending of patient vital signs data is done through an ON-ACTIVE NI in a real-time, i.e., without application-data buffering. These strategies are S_{EM}, S_1 and S_2 defined correspondingly to cases 4a, 3a and 1a and are to be used in emergency, but can also be used in non-emergency. These strategies are ordered by their AppRTT in Tab. 3, with the most delay-efficient strategy S_{EM} (and hence most recommended in emergency) and the least efficient S_2. The power consumed by the strategy S_{EM} is considered as our reference point for comparing the power efficiency of other strategies. The power efficiency of strategy S_X is then defined by (power$_{MBU}$ (S_{EM}) – power$_{MBU}$ (S_X)) / power$_{MBU}$ (S_{EM}); the bigger the resulting value, the more efficient the strategy is. The last row in Tab. 3 indicates if the strategy fulfills the requirement of a user being reachable on his/her mobile device via the WWAN-GPRS network.

TABLE III. PERFORMANCE OF THE BASIC NI ACTIVATION STRATEGIES

Strategy	**Performance of the Basic NI Activation Strategies**		
	S_{EM} (4a)	***S_1 (3a)***	***S_2 (1a)***
AppRTT [ms]	1806	2753	3739
power efficiency	0	9	25

Strategy	Performance of the Basic NI Activation Strategies		
	S_{EM} (4a)	***S_1 (3a)***	***S_2 (1a)***
WWAN reachability	yes	no	yes

For cases where the larger AppRTTs are acceptable, i.e. in non-emergency, the MBU may adapt application-data flows by acquiring *n-1* (n>1) seconds of the patient vital signs data, temporarily storing this data, and sending it in a burst in the nth second (together with the nth second data sample) to the backend-server via a chosen NI. The entries in Tab. 2 and 3 for cases where data volumes achieve 5.2 kbps (1b, 3b, 4b) and 7.7 kbps (1c) make our basis to consider n=4 (thus achieving 5.2 kbps) and n=6 (thus achieving 7.7 kbps).

Tab. 4 summarizes the comparison results for three distinctive application-data flow adaptation cases extrapolated from measurements cases: 1b, 3b, 4b and 1c. The power efficiency of a strategy is again defined against the S_{EM}. The following relations hold in the Tab.:

AppRTT = (n-1)*1000 + measured AppRTT [ms]

Normalized power =
1/n [(n-1) power$_{MBU}$ (NI$_1$=ON-IDLE, NI$_2$=s)
+ power$_{MBU}$ (NI$_1$=ON-ACTIVE, NI$_2$=s)]

where NI$_1$ represents the NI through which the data is sent, while NI$_2$ is being in a state s.

TABLE IV. PERFORMANCE OF THE BASIC NI ACTIVATION STRATEGIES

Strategy	Performance of the Basic NI Activation Strategies			
	S4 (4b, n=4)	***S5 (3b, n=4)***	***S6 (1b, n=4)***	***S7 (4c, n=6)***
WLAN	alternates: ON-IDLE ↔ ON-ACTIVE	alternates: ON-IDLE ↔ ON-ACTIVE	OFF	OFF
GPRS	ON-IDLE	OFF	alternates: ON-IDLE ↔ ON-ACTIVE	alternates: ON-IDLE ↔ ON-ACTIVE
AppRTT [ms]	3000 + 2211	3000 + 3513	3000 + 5505	5000 + 6693
normalized power	0.00966	0.00654	0.00584	0.00555
power eff.	-0.6	32	39	42
WWAN reachability	yes	no	yes	yes

From the Tab. 4 we conclude that for a patient in non-emergency, strategies S_6 and S_7, where data is sent in burst through the GPRS NI, are more power efficient than those where data is sent through WLAN NI, however less AppRTT-efficient. The result for strategy S_4 shows that this strategy is a bit less power-efficient comparing to S_{EM}. This is due to the high power consumption of MBU for WLAN ON-IDLE state (as we explained for Tab. 1).

For the cases with larger bursts (i.e. larger n), we use the results for NetPerf measurements to extrapolate the efficiency, as presented in Tab. 5. Hereto, we estimate the maximum:

AppRTT ≈ (n-1) + C [s],

where C is a constant, with a slight dependency on n, in the order of a few seconds and approximately represents the time of n data samples. Similarly, the normalized power is computed as:

normalized average power ≈

1/n [(n-1) $\text{power}_{\text{MBU}}$ (WLAN=ON-IDLE, GPRS=s)

+ $\text{power}_{\text{MBU}}$ (WLAN=ON-ACTIVE, GPRS=s)]

where s is a given state of the GPRS NI. For large values of n, the normalized average power approaches the $\text{power}_{\text{MBU}}$ for (WLAN=ON-IDLE, GPRS=s) case.

Strategies S_8 and S_9 as defined in Tab. 5, disclose large difference for the WLAN NI alternating between ON-IDLE and ON-ACTIVE states, and GPRS being in ON-IDLE or OFF states. If n is large enough, one may switch the WLAN NI between OFF and ON-ACTIVE states[8] resulting in strategies S_{10} and S_{11}.

TABLE V. PERFORMANCE OF THE BASIC NI ACTIVATION STRATEGIES

Strategy	**Performance of the Basic NI Activation Strategies**			
	S8 (large n)	***S9 (large n)***	***S10 (large n)***	***S11 (large n)***
WLAN	ON-IDLE↔ ON-ACTIVE	ON-IDLE↔ ON-ACTIVE	OFF↔ ON-ACTIVE	OFF↔ ON-ACTIVE
GPRS	ON-IDLE	OFF	ON-IDLE	OFF
AppRTT [ms]	≈ n-1 + C	≈ n-1 + C	≈ n-1 + C	≈ n-1 + C
normalized power	≈ (n-1)/n 0.00963	≈ (n-1)/n 0.00568	≈ (n-1)/n 0.00487	≈ (n-1)/n 0.00092
power efficiency	-0.3	41	49	90
WWAN reachability	yes	no	yes	no

As can be seen from the Tab. 5, strategy S_{10} is slightly more power efficient than S_7 while it induces very large AppRTT. Only the power efficiency of strategy S_{11} is significantly higher with respect to that of the strategy S_7, but the drawback is that the mobile device is not WWAN-reachable. The results of Tab. 5, indicate that adapting patient vital signs data and sending it in large bursts (i.e. with a large n) is not power efficient enough to motivate having a very long AppRTT or being WWAN-unreachable.

6 Related Work

Related work on NI activation strategies is mainly theoretical, and moreover focuses mainly on applications in which mobile user acts as an occasional data consumer and does not produce data, as in the MobiHealth system. For example, authors of [10-13] consider NI strategies together with methods for local or proxy-based caching data for users of email application and web-services. The work reported in [14] reduced energy consumption by introducing a NI ON-IDLE stand-

[8] These NI state changes impose a fixed power penalty higher than that in case of strategies S8 and S9. This penalty is negligible as n increases.

by state, at which the mobile device is wakened-up if there is an incoming network event, e.g. a call. Considering the impact of applications on NI power consumption, the authors of [15] studied the WLAN NI energy consumption for different multimedia data streaming applications like Microsoft (Windows media), Real (Real media) and Apple (Quick Time) content. They considered only WLAN NI and downlink data streams. Similarly, but from the NI perspective, authors of [16] measured NI energy consumption of use/and alternating between BT and WLAN NIs for a multimedia content download. Furthermore, there exist general research frameworks, in which NI activation strategy is considered as one of multiple features. For example, the research reported in [17, 18] considered a simultaneous operation of NIs in multi-homed mobile hosts, and introduced a Basic Access Network to carry out signalling for network discovery, NI selection, inter-network handover, location updates, paging, authentication, authorization, and accounting. Authors tackled the NI activation strategy objective only theoretically. Similarly, the theoretical framework proposed in [19] focuses specifically on the WLAN NI activation strategy, based on the WLAN network availability, network state (throughput, delays and reliability), as well as application QoS requirements. Their NI strategy assumes that the UMTS network is always ON and available. However, they do not consider the NI energy consumption in their framework. Authors of [20] aimed to estimate WLAN network availability and conditions without powering a NI up - based only on historical data. They have simulated healthcare application by data for 3 leads ECG; however they neither include BT power consumption for sensor-set nor adapted application-data flow being sent by network (i.e. it was fixed at 5 minutes). And finally, in our previous work [21], we have studied the NI activation strategies based on its relative energy cost. We measured energy costs while sending dummy TCP packets over a given NI. However, the data range send was 25 kbps (GPRS) and 2 Mbps (WLAN), and the mobile devices, as well as measurements conditions were different, which made these results unusable for the MobiHealth case study presented in this paper. We emphasize the contribution of our research as an extensive case study of the existing system for telemonitoring of patient’s health conditions, based on which we provide extensive and valuable recommendations for the system users.

7 Conclusions and the MobiHealth System Recommendations

Based on our measurements, we derive some conclusions and recommendations for the MobiHealth system and its cardiac telemonitoring application, concerning the most efficient and effective NI activation strategies along the power and the delay QoS requirements. Particularly, we have observed that the GPRS and WLAN NIs have complementary power and delay profiles. For GPRS, there is lower energy cost to maintain connectivity and lower energy to send data, but higher delay. On the other hand, the energy cost of a WLAN data send can be higher, but delay is lower. Minimal power is used in strategies where data is stored and sent later it bursts (S_8-S_{11}), resulting in the highest delay (as they include long local storage time). Maximum power is used by S_{EM} (comparing to the other strategies where data is sent in real-time: S_1 and S_2), while the delay is minimal.

In an emergency case, the WLAN ON-ACTIVE and GPRS ON-IDLE NI activation strategy should be used, as it provides the system with the lowest patient vital signs data delay. However, if WLAN is not available, GPRS ON-ACTIVE and WLAN-OFF case should be used.

In non-emergency, when the user needs to be reachable, the data can be sent in burst and power needs to be optimized, we recommend the use of the WLAN ON-ACTIVE and GPRS ON-IDLE strategy. The recommended burst size is the one corresponding to n=4 seconds of patient vital signs data. However, if WLAN is not available, GPRS should be used with WLAN OFF with a recommended burst size corresponding to n=6 seconds patient vital signs data.

Another conclusion derived from our studies is that the device used as the MobiHealth's MBU – Qtek 9090 is not necessarily the best choice from the power efficiency perspective for GPRS/WLAN interfaces. As one of the future work areas, we recommend execution of measurements for other devices, and comparison of results between the studies.

Moreover, future work encompasses work on more elaborated NI activation strategies methods, e.g. those including multiple periodic application-data flows with different delay requirements (i.e. different delay defined per application-data flow). Moreover, the NI strategy should include network's monetary cost usage and a network's security level required by the MobiHealth users. Finally, we plan to extend our study of the power- and delay application-data flow adaptation from the user's stationary position to different mobility levels, where data is sent over the different WWAN networks (GPRS, UMTS, or HSPA) as available at a given user's geographical location and time.

References

[1] Tachakra, S., X. Wang, et al. Mobile e-Health: the Unwired Evolution of Telemedicine. *Telemedicine J and e-Health*, 2003, 9(3): 247-257.

[2] van Halteren, A., Bults, R., et al. Mobile Patient Monitoring: The MobiHealth System. *The Journal on Information Technology in Healthcare*, 2004, 2(5): 365-373.

[3] Dokovsky, N., A. van Halteren, et al. BANip: Enabling Remote Healthcare Monitoring with Body Area Networks. Intl Workshop on Scientific Engineering of Distributed Java Applications (FIJI03), 2003, Luxembourg, Springer Verlag.

[4] Pawar, P., van Beijnum, B. J., et al. Context-Aware Middleware Support for the Nomadic Mobile Services on Multi-homed Handheld Mobile Devices. *IEEE Symp. on Comp. & Comm.*, Portugal, IEEE, 2007

[5] Wac, K., Bults, R., et al. Mobile Health Care over 3G Networks: The MobiHealth Pilot System and Service. *Global Mobile Congress*, Shanghai, China, 2004.

[6] Broens, T., Huis in't Veld, R., et al. Determinants for successful telemedicine implementations: a literature study. *Journal for Telemedicine and Telecare*, 2007, 13(6): 303-309.

[7] Bernaschi, M., Cacace, F., and Iannello, G. Vertical Handoff Performance in Heterogeneous Networks. Intl Conf. on Parallel Processing Workshops (ICPPW04), 2004, Montreal, Canada.

[8] Twente Medical Systems Intl., www.tmsi.com, retrieved on 09/12/2007.

[9] Netperf homepage: www.netperf.org, retrieved on 25/11/2007.

[10] Flinn, J. and Satyanarayanan, M. Energy-aware adaptation for mobile applications. *ACM Symp. on Operating Systems Principles*, USA. ACM, New York, NY, 48-63., 1999.

[11] Armstrong, T., Trescases, O., Amza, C., and de Lara, E. Efficient and transparent dynamic content updates for mobile clients. *MobiSys'06*, Sweden. ACM, NY, US, 56-68, 2006.

[12] Lufei, H. and Shi, W. e-QoS: energy-aware QoS for application sessions across multiple protocol domains in mobile computing. *QShine'06*, Canada. v.191. ACM, New York, NY, 2006.

[13] Anand, M., Nightingale, E. B., and Flinn, J. Self-tuning wireless network power management. *Wirel. Netw.* 2005, 11(4), 451-469.

[14] Shih, E., Bahl, P., and Sinclair, M. J. Wake on wireless: an event driven energy saving strategy for battery operated devices. *MobiSys'02*, USA. ACM, New York, NY, 160-171, 2002.

[15] Chandra. S., Wireless network interface energy consumption. Implications for popular streaming formats, *Multimedia Systems*, Springer-Verlag, 9(2), pp. 185-201, 2003.

[16] Pering, T., Agarwal, Y., Gupta, R., and Want, R. *CoolSpots*: reducing the power consumption of wireless mobile devices with multiple radio interfaces. *MobiSys'06*, Sweden. ACM, US, 220-232, 2006

[17] Inoue, M. Mahmud, K., Murakami, H. Hasegawa, M. and Morikawa, M. Novel Out-of-Band Signaling for Seamless Interworking Between Heterogeneous Networks, *IEEE Wireless Comm.* 2004

[18] Wu, G., Mizuno, M. and Havinga, P. MIRAI Architecture for Heterogeneous Network, *IEEE Comm. Magazine*, Feb 2002.

[19] Song, Q. and Jamalipour. A. Network Selection in an Integrated Wireless LAN and UMTS Environment Using Mathematical Modeling and Computing Techniques. *IEEE Wireless Comm.*, 12(3), 2005, 42-48.

[20] Rahmati, A. and Zhong, L. Context-for-wireless: context-sensitive energy-efficient wireless data transfer. *MobiSys'07*, Puerto Rico. ACM, New York, NY, 165-178, 2007

[21] Bargh, M., A. Peddemors. Towards an Energy-Aware Network Activation Strategy for Multi-Homed Mobile Devices. *Intl Conf. on Pervasive Systems Computing*, USA, 2006

Hovering Information - Self-Organising Information that Finds its Own Storage[1]

Alfredo A. Villalba Castro, Giovanna Di Marzo Serugendo, and Dimitri Konstantas

Abstract. A piece of Hovering Information is a geo-localized information residing in a highly dynamic environment such as a mobile ad hoc network. This information is attached to a geographical point, called the anchor location, and to its vicinity area, called anchor area. A piece of hovering information is responsible for keeping itself alive, available and accessible to other devices within its anchor area. Hovering information uses mechanisms such as active hopping, replication and dissemination among mobile nodes to satisfy the above requirements. It does not rely on any central server. This paper presents the hovering information concept and discusses results of simulations performed for two algorithms aiming to ensure the availability of a piece of hovering information at its anchor area.

1 Introduction

Hovering information [7] is a concept characterising self-organising information responsible to find its own storage on top of a highly dynamic set of mobile devices. The main requirement of a single piece of hovering information is to keep itself stored at some specified location, which we call the anchor location, despite the unreliability of the device on which it is stored. Whenever the mobile device, on which the hovering information is currently stored, leaves the area around the specified storage location, the information has to hop - ”hover” - to another device.

Current approaches in this area (cf. Section 6) try to either define a virtual structured overlay network on top of this environment offering a stable virtual infrastructure, or propose a system-based approach offering services such as information dissemination and storage. In these approaches, the mobile nodes decide when and to whom the information is to be sent. Here we take the opposite view; it is the information that decides upon its own storage and dissemination. This opens up other possibilities, not available for traditional MANET services, such as different pieces of hovering information all moving towards the same location and (re-)constructing there a coherent larger information for a user, e.g. TV or video streaming on mobile phones.

Hovering information is a *self-organised* user-defined information which do not need a central server to exist. Individual pieces of hovering information each use local information, such as direction, position, power and storage capabilities of nearby mobile devices, in order to select the next appropriate location. Hovering information benefits from the storage space and communication capacities of the underlying

[1] This paper has been presented and included in the proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC’08), 11-13 June 2008, Taichung, Taiwan.

mobile devices. It is not residing in a centralized server, and is not bound to any mobile operator.

This paper presents the hovering information concept as well as a preliminary algorithm allowing single pieces of hovering information to get attracted to their respective anchor locations. A complete formal description of the hovering information model is described in [6].

Section 2 discusses potential applications of this concept. Section 3 presents the hovering information concept. Section 4 discusses the Attractor Point algorithm that we have designed where the information is ”attracted” by the anchor location and a general Broadcast algorithm we implemented in order to allow comparisons. Section 5 reports on simulation results related to availability and additional metrics such as number of messages exchanged or memory storage used. Finally Section 6 compares our approach to related works, and Section 7 discusses some future works.

2 Applications

When deployed over mobile devices, hovering information is an infrastructure free service that supports a large range of applications. Among others we can cite: *urban security* - users (citizens, policemen, security) post and retrieve comments or warnings related to dangers in their urban environment; *self-generative art* - users of a learning art experience centre provide collective inputs self-assembled together into a piece of art (painting, music, etc) generated by a computer according to some rules; *intravehicular networks* - drivers insert tags into the environment related to road conditions or accidents; *emergency scenarios* - emergency crew use hovering information to locate survivors or coordinate their work. More generally, hovering information is a technical way to support *stigmergy-based applications*. Stigmergy is an indirect communication mechanism among individual components of a self-organising system. Communication occurs through modification brought to local environment. The use of ant pheromone is a well known example of stigmergy. Users that communicate by placing hovering information at a geo-referenced position, which is later on retrieved by other users is also an example of stigmergy. The hovering information concept, using an infrastructure free storage media, naturally supports stigmergy-based applications that need to be deployed on an ad hoc manner (e.g. unmanned vehicles or robots).

3 Hovering Information Concept

3.1 Mobile Nodes and Hovering Information

Mobile nodes represent the storage and motion media exploited by pieces of hovering information. A *mobile node* n is defined as a tuple:

$$n = (id, loc, speed, dir, r_{comm}),$$

where id is its mobile node identifier, loc is its current location (a geographic location), $speed$ is its speed in m/s, dir is its current direction of movement (a geographic vector) and r_{comm} is its wireless communication range in meters.

A piece of hovering information is a piece of data whose main goal is to remain stored in an area centred at a specific location called the *anchor location*, and having a radius called the *anchor radius*. A *piece of hovering information* h is defined as a tuple:

$$h = (id, a, r, n, data, policies, size),$$

where id is its hovering information identifier, a is its anchor location (geographic coordinate), r is its anchor radius in meters, n is the mobile node where h is currently hosted (hosting node), $data$ is the data carried by h, $policies$ are the hovering policies of h and $size$ is the size of h in bytes. Policies stand for hovering policies stating how and when a piece of hovering information has to hover.

We consider that identifiers of pieces of hovering information are unique, but replicas (carrying same data and anchor information) are allowed on *different* mobile nodes.

We also consider that there is only one instance of a hovering information in a given node n, any other replica resides in another node.

Figure 1 shows a piece of hovering information (blue hexagon) and two mobile nodes (yellow circles). One of them hosts the hovering information whose anchor location, radius and area are also represented (blue circle). The anchor area is the disc whose center is the anchor location, and radius is the anchor radius. The communication range of the second mobile node is also showed.

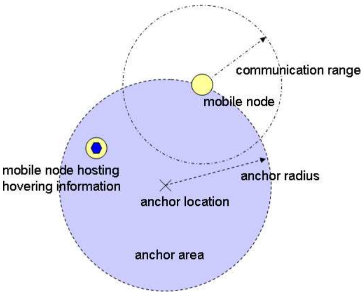


Fig. 1. Mobile Nodes and Hovering Information

A hovering information system is composed of mobile nodes and pieces of hovering information. A hovering information system at time t is a snapshot (at time t) of the status of the system, the system then evolves at each time tick t, $t+1$, etc. Mobile nodes can change location, new mobile nodes can join the system and others can leave. New pieces of hovering information can appear (with new identifiers), replicas may appear or disappear (same identifiers but located on other nodes), hovering information may disappear or change node.

Figure 2 shows two different pieces of hovering information h_1 (blue) and h_2 (green), having each a different anchor location and area. Two replicas of h_1 are currently located in the anchor area (in two different mobile nodes n_2 and n_4), while three replicas of h_2 are present in the anchor area of h_2 (in nodes n_2, n_3 and n_5). It may happen that a mobile device hosts replicas of different pieces of hovering information, as it is the case in the figure for the mobile node n_2 that is at the intersection of the two anchor areas. The arrows here also represent the communication range possibilities among the nodes.

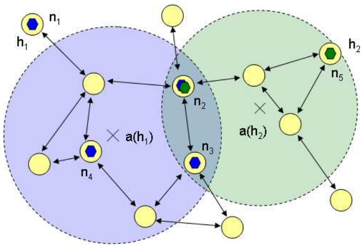


Fig. 2. Hovering Information System at time t

3.2 Properties - Requirements

Survivability. A hovering information h is alive at some time t if there is at least one node hosting a replica of this information. The survivability along a period of time is defined as the ratio between the amount of time during which the hovering information has been alive and the overall duration of the observation. The survivability of h between time 0 and time t is given by:

$$SV_H(h,t) = \frac{\sum_{\tau=0}^{t} sv_H(h,\tau)}{t},$$

where $sv_H(h,\tau)$ takes value 0 or 1 whether h is survival or not at time τ.

Availability. A hovering information h is available at some time t if there is at least a node in its anchor area hosting a replica of this information. The availability of a piece of hovering information along a period of time is defined as the rate between the amount of time along which this information has been available during this period and the overall time. The availability of h between time 0 and time t is given by:

$$AV_H(h,t) = \frac{\sum_{\tau=0}^{t} av_H(h,\tau)}{t},$$

where $av_H(h,\tau)$ takes value 0 or 1 whether h is available or not at time τ.

Accessibility. A hovering information is accessible by a node n at some time t if the node is able to get this information. In other words, if it exists a node m being in the communication range of the interested node n and which contains a replica of the piece of hovering information. The accessibility of a piece of hovering information h is the rate between the area covered by the hovering information's replicas and its anchor area. The accessibility of h between time 0 and time t is given by:

$$AC_H(h,t) = \frac{\sum_{\tau=0}^{t} ac_H(h,\tau)}{t},$$

where $ac_H(h,\tau)$ is the rate between the area covered by the hovering informations replicas and its anchor area. The interested reader can refer to [6] for a full set of definitions.

Let us notice that an available piece of hovering information is not necessarily accessible and vice-versa, an accessible piece of hovering information is not necessary available. Figure 3 shows different cases of survivability, availability and accessibility. In Figure 3(a), hovering information h (blue) is not available, since it is not physically present in the anchor area, however it is survival as there is a node hosting it. In Figure 3(b), hovering information h is now available as it is within its anchor area, however it is not accessible from node n_1 because of the scope of the communication range. Finally, in Figure 3(c), hovering information h is survival, available and accessible from node n_1.

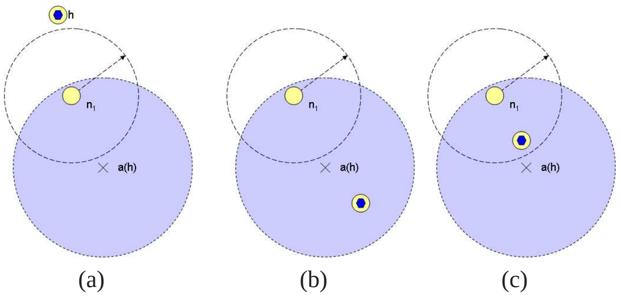


Fig. 3. Survivability, Availability and Accessibility

4 Algorithms for Hovering Information

4.1 Assumptions

We make the following assumptions in order to keep the problem simple while focusing on measuring availability and resource consumption. **Unlimited memory:** All mobile nodes have an unlimited amount of memory able to store any number of hovering information replicas. The proposed algorithms do not take into account remaining memory space or the size of the hovering information. **Unlimited energy:** All mobile nodes have an unlimited amount of energy. The proposed algorithms do not consider failure of nodes or impossibility of sending messages because of low level of energy. **Instantaneous processing:** Processing time of the algorithms in a mobile node is zero. We do not consider performance problems related to overloaded processors or execution time. **In-built geo-localization service:** Mobile nodes have an in-built geo-localization service such as GPS which provides the current position. We assume that this information is available to pieces of hovering information. **Neighbours discovering service:** Mobile nodes are able to get a list of their current neighbouring nodes at any time. This list contains the position, speed, and direction of the nodes. As for the other two services, this information is available to pieces of hovering information.

4.2 Safe, Risk and Relevant Areas

In this paper we consider that all pieces of hovering information have the same hovering policies: active replication and hovering in order to stay in the anchor area (for availability and accessibility reasons), hovering and caching when too far from the anchor area (survivability), and cleaning when too far from the anchor area to be meaningful (i.e. disappearance). The decision on whether to replicate itself or to hover depends on the current position of the mobile device in which the hovering information is currently stored. Therefore, we distinguish three different areas: safe area, risk area and relevant area.

A piece of hovering information located in the *safe area* can safely stay in the current mobile node, provided the conditions on the node permit this: power, memory, etc. This area is defined as the disc having as centre the anchor location and as radius the safe radius.

A piece of hovering information located in the *risk area* should actively seek a new location on a mobile node going into the direction of the safe area. It is in this area that the hovering information actively replicates itself in order to survive and stay available in the vicinity of the anchor location. This area is defined as the ring having as centre the anchor location and bound by the safe and risk radiuses.

The *relevant area* limits the scope of survivability of a piece of hovering information. This area is defined as the disk whose centre is the anchor location and whose radius is the relevant radius.

The *irrelevant area* is all the area outside the relevant area. A piece of hovering information located in the irrelevant area can disappear; it is relieved from survivability goals.

Figure 4 below depicts the different types of radiuses and areas discussed above centred at a specific anchor location a. The smallest disk represents the safe area, the blue area is the anchor area, the ring limited by the risk radius and the safe radius is the risk area, and finally the larger disk is the relevant area.

The values of these different radiuses are different for each piece of hovering information and are typically stored in the Policies field of the hovering information. In the following algorithms we consider that all pieces of hovering information have the same relevant, risk and safe radius.

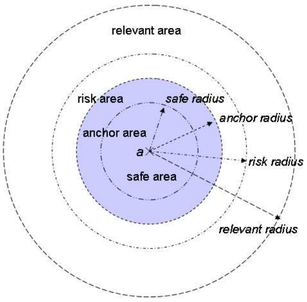


Fig. 4. Radiuses and areas

4.3 Replication

We describe two algorithms simulating two variants of replication policies: the Attractor Point and Broadcast algorithms. Both algorithms are triggered periodically each T_R seconds and only replicas of h being in the risk area are replicated onto some neighbouring nodes (nodes in communication range) which are selected according to the replication algorithm.

Attractor Point Algorithm

The anchor location of a piece of hovering information acts constantly as an attractor point to that piece of hovering information and to all its replicas. Replicas tend to stay as close as possible to their anchor area by replicating from one mobile node to the other.

Algorithm 1 Attractor Point Replication Algorithm

```
procedure REPLICATION
    $pos \leftarrow$ MY-POSITION
    $N \leftarrow$ MY-NEIGHBOURS
    $P \leftarrow$ POSITION$(N)$
    for all $replica \in REPLICAS$ do
        $anchor \leftarrow$ ANCHOR-LOCATION$(replica)$
        $dist \leftarrow$ DISTANCE$(pos, anchor)$
        if $(dist \geq r_{safe})$ and $(dist \leq r_{risk})$ then
            $D \leftarrow$ DISTANCE$(P, anchor)$
            $D' \leftarrow$ SORT$(D)$
            $M \leftarrow$ SELECT$(D', 1, k_R)$
            MULTICAST(info,M)
        end if
    end for
end procedure
```

Periodically and for each mobile node, the algorithm (see Algorithm 1) checks the position of the mobile node (line 2) as well as the list and position of all mobile nodes in communication range (lines 3 and 4). The algorithm then verifies whether there are some hovering information replicas being in the risk area that need to be replicated (line 8). The number of target nodes composing the multicast group is defined by the constant k_R. The distance between each mobile node in range and the anchor location is computed (line 9). The k_R mobile nodes with the shortest distance are chosen as the target nodes for the multicast (lines 10 and 11). The information part of the selected pieces of hovering information is then multicasted to the k_R mobile nodes, in communication range, closest to the anchor location (line 12). Figure 5 illustrates the behaviour of the Attractor Point algorithm. Consider a piece of hovering information h in the risk area. It replicates itself onto the nodes in communication range that are the closest to its anchor location. For a replication factor $k_R = 2$, nodes n_2 and n_3 receive a replica, while all the other nodes in range do not receive any replica.

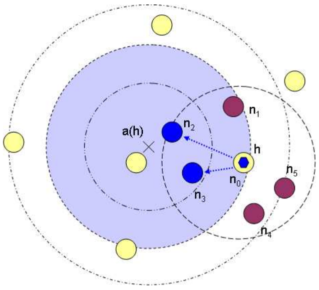


Fig. 5. Attractor Point Algorithm

Broadcast Algorithm

The Broadcast algorithm (see Algorithm 2) is triggered periodically (each T_R) for each mobile node. After checking the position of the mobile node (line 2); pieces of hovering information located in the risk area (line 6) are replicated and broadcasted onto all the nodes in communication range (line 7). We expect this algorithm to have the best performance in terms of availability but the worst in terms of network and memory resource consumption.

Figure 6 illustrates the behaviour of the Broadcast algorithm. Consider the piece of hovering information h in the risk area, it replicates itself onto all the nodes in communication range, nodes n_1 to n_5 (blue nodes).

4.4 Caching and Cleaning Modules

Each node is assumed to have an unlimited amount of memory. Therefore, when replicas are sent from one node to another, they are simply stored in the nodes mem-

Algorithm 2 Broadcast-based Replication Algorithm

```
procedure REPLICATION
    $pos \leftarrow$ MY-POSITION
    for all $replica \in REPLICAS$ do
        $anchor \leftarrow$ ANCHOR-LOCATION$(replica)$
        $dist \leftarrow$ DISTANCE$(pos, anchor)$
        if $(dist \geq r_{safe})$ and $(dist \leq r_{risk})$ then
            BROADCAST$(replica)$
        end if
    end for
end procedure
```

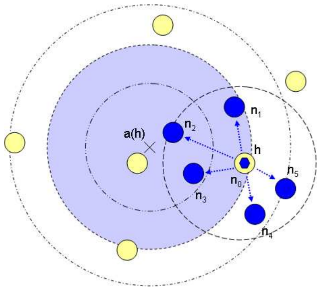


Fig. 6. Broadcast Algorithm

ory. However, if a node receives two or more replicas of the *same* piece of hovering information h, the first replica to arrive is stored in the memory, and any subsequent one is ignored. Therefore, at most one replica of each piece of hovering information is present in a given node n.

Periodically - each T_C seconds - and for each node, replicas that are too far from their anchor location are removed, i.e. those replicas that are in the irrelevant area. Although the amount of memory is unlimited and replicas could stay forever in the nodes' memory, we remove the replicas that are too far away from their anchor location, this represents the cases where the replica considers itself too far from the anchor area and not able to come back anymore. This avoids as well the situation were all nodes have a replica.

4.5 Metrics

In order to evaluate and compare the above algorithms, the following values have been measured.

Messages complexity. The message complexity at a given time t is the number of messages sent between time 0 and time t by all nodes n of the system ($\mathcal{N}_t$):

$$MSGS(t) = \sum_{\tau=0}^{t} \sum_{n \in \mathcal{N}_\tau} msgs_n(\tau),$$

where $msgs_n(\tau)$ represents the number of messages sent at time τ by node n.

Space complexity. While the complexity of messages relates to the number of sent messages, the space complexity measures the amount of memory used by a device while hosting a piece of hovering information. The space complexity measures the maximum amount of memory used by a single mobile node to store hovering information replicas.

$$MEM(t) = \max_{\tau=0}^{t}(\max_{n \in \mathcal{N}_\tau} mem_n(\tau)),$$

where $mem_n(\tau))$ is the number of pieces of hovering information in n at time τ.

We measure this value in order to know how much memory is consumed by our algorithms.

Concentration. The concentration of a given piece of hovering information h is defined as the rate between the number of replicas of h present in the anchor area and the total number of replicas of this hovering information in the whole environment.

5 Evaluation

We evaluated the behaviour of the two above described algorithms under different scenarios by varying the number of nodes. In these experiments, we considered only one piece of hovering information. For this given piece of hovering information h, we measured the availability of h, the corresponding message complexity, the corresponding space complexity and the concentration of h.

We performed simulations using the OMNet++ network simulator (distribution 3.3) and its Mobility Framework 2.0p2 (mobility module) to simulate nodes having WiFi-enabled communication interfaces.

5.1 Simulation Settings and Scenarios

The generic scenario consists of a surface of 500m x 500m with mobile nodes moving around following a Random Way Point mobility model with a speed varying from 1m/s to 10m/s without pause time. In this kind of mobility model, a node moves along a straight line with speed and direction changing randomly at some random time intervals. Before using this mobility model, the simulation has also been validated using two simple scenarios: in the first one nodes moved along a straight line at a constant speed following the same direction (one way road) and in the second one nodes moved along two opposite straight lines (double way road). Table 1 summarises the values used for the generic scenario.

Based on this generic scenario, we defined 10 specific scenarios with varying number of nodes: from 20 to 200 nodes, increasing the number of nodes by 20 each time. We have performed 20 runs for each scenario. One run lasts 3600 seconds of simulation time. All the results presented here are the average of the 20 runs for each scenario, and the errors bars represent a 95% confidence interval. All the simulations ran on a Pentium 1.7 GHz processor under Linux Mandriva OS.

Blackboard	500mx500m
Mobility Model	Random Way Point
Nodes speed	1m/s to 10 m/s
Communication range (MID)	121m
Replication time (T_R)	20s
Cleaning time (T_C)	60s
Anchor radius	50m
Min risk radius	30m
Max risk radius	70m
Relevant radius	200m

Table 1. Simulation settings

5.2 Results

Availability. Figure 7 shows, for each of the 10 scenarios, the average of the availability performance over the 20 runs (after one hour of simulation time). As expected, the Broadcast algorithm outperforms the Attractor Point algorithm. The results also indicate that the performances of the Attractor Point algorithm although lower are quite similar to those of the Broadcast algorithm. For both algorithms, we observe that an 80% of availability can be expected as soon as the number of mobile nodes in the environment reaches 120 nodes. This represents a density of 3.8 nodes per anchor area. The maximum availability value, nearly 95%, is reached by the Broadcast algorithm when the population of mobile nodes is 200, while the Attractor Point reaches 88% of availability for 180 nodes and above.

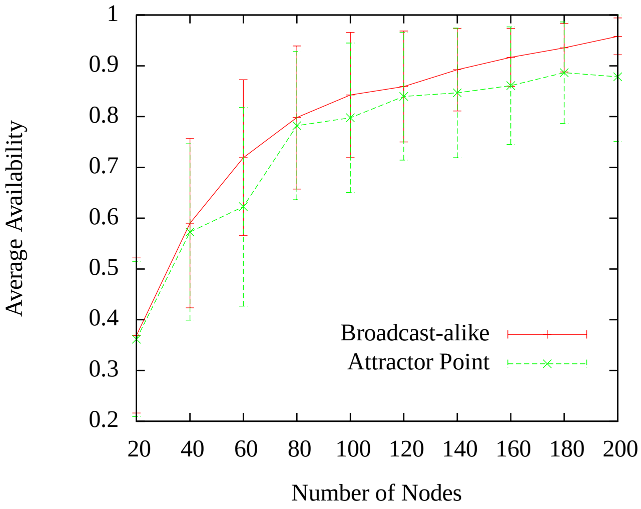


Fig. 7. Average availability after 1 hour of simulation

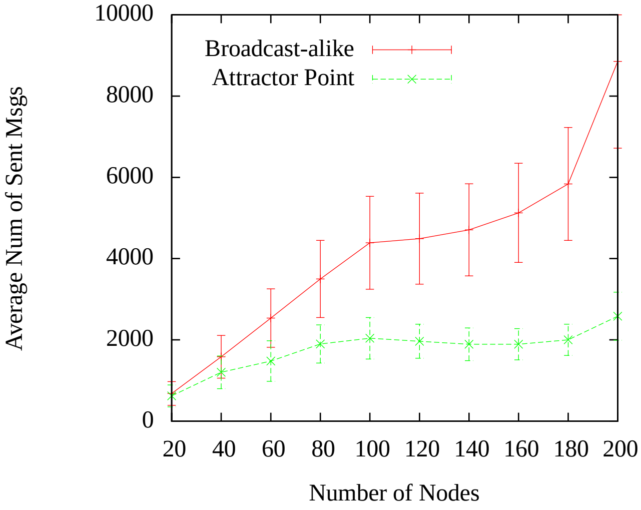


Fig. 8. Messages Complexity after 1 hour of simulation

Message Complexity. Figure 8 shows the average number of messages sent for each of the 10 different scenarios. As expected, the Broadcast algorithm sends a higher number of messages when compared to the Attractor Point algorithm. This phenomenon is amplified when the number of nodes increases. In the worst case (200 nodes), the number of sent messages, in average, by the Broadcast algorithm is four times higher than the number of messages sent when the Attractor Point algorithm is used. In the other cases, it is 2.5 times higher (100 to 180 nodes)

Space Complexity. Figure 9 shows in average the maximum number of replicas of a single piece of hovering information created during the simulations for each different scenario.

Again, we observe that the Broadcast algorithm creates more replicas than the Attractor Point. The curves for Space Complexity are very similar to those for Message Complexity (see Figure 8). This is explained as the number of sent messages is directly proportional to the number of existing replicas; since each replica can potentially send messages (replicate itself again).

Concentration. Figure 10 shows the concentration factor. We observe that the Attractor Point algorithm concentrates more replicas in the anchor area than the Broadcast algorithm. The concentration rate is above 14% for the Attractor Point algorithm when the number of mobile nodes is 80 or more. A maximal concentration rate of 8% is reached by the Broadcast algorithm. The Attractor Point concentrates 2 to 3 times more replicas than the Broadcast algorithm (depending on the number of nodes considered).

At the time of writing, additional simulations are running. They are aiming at computing the **accessibility** of hovering information under different anchor and communication radiuses.

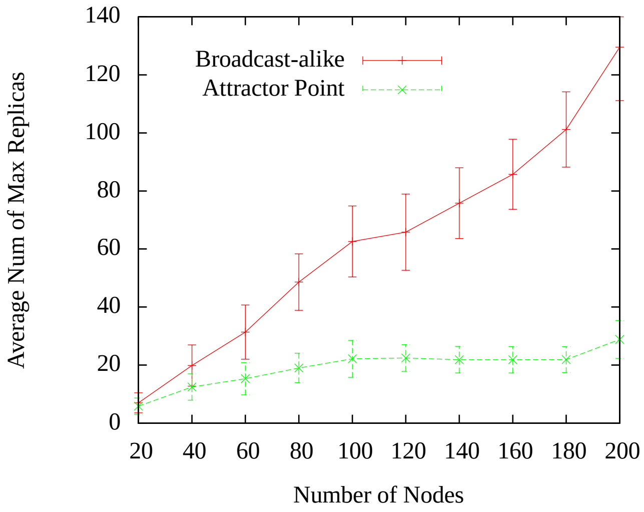


Fig. 9. Space Complexity

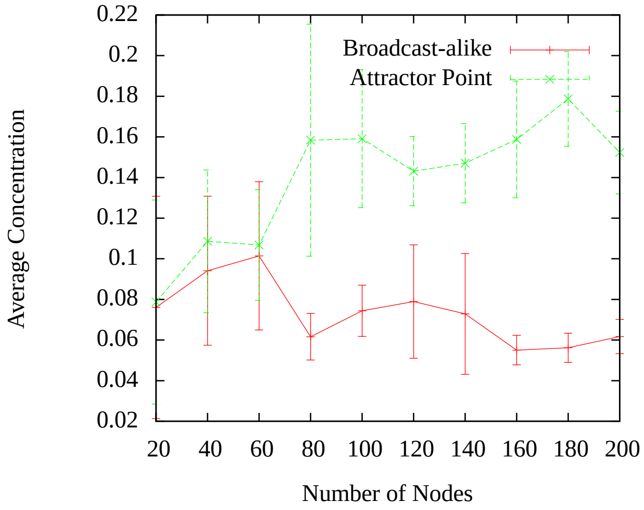


Fig. 10. Concentration

6 Related Works

The Virtual Infrastructure project [2, 3] defines virtual (fixed) nodes implemented on top of a MANET. This project proposes the notion of an *atomic memory cells*, implemented on top of a MANET, which ensure their persistency by replicating their state in neighbouring mobile devices. This notion has been extended to the idea of *virtual* *mobile nodes* which are state machines having a fixed location or a well-defined trajectory and whose content is also replicated among the nearby mobile devices. The motivation behind this project is the development of a virtual infrastructure on top of which it becomes easier to define distributed algorithms such as routing or leader election. Similarly, hovering information tries to benefit from the mobility of the underlying nodes, but the goal is different. The long term goal is to provide a hovering information service on top of which applications using self-organising user-defined pieces of information can be built.

GeOpps [4] proposes a geographical opportunistic routing algorithm over VANETs (Vehicular Ad Hoc Networks). The algorithm selects appropriate cars for routing some information from a point A to a point B. The choice of the next hop (i.e. the next car) is based on the distance between that cars trajectory and the final destination of the information to route. This work focuses on routing information to some geographical location; it does not consider the issue of keeping this information alive at the destination, while this is the main characteristics of hovering information.

The work proposed by [5] aims to disseminate traffic information in a network composed by infostations and cars. The system follows the publish/subscribe paradigm. Once a publisher creates some information, a replica is created and propagated all around where the information is relevant. Clusters are composed and replicas are removed or propagated to clusters where more subscribers and interested cars are situated. Replicas are also propagated to a randomly chosen car part of the cluster driving in the opposite direction to that of the current host in order to try to keep the information in its relevant area. While the idea is quite similar to that of hovering information, keeping information alive in its relevant area, this study does not consider the problem of having a limited amount of memory to be shared by many pieces of information or the problem of fragmentation of information. It also takes the view of the cars as the main active entities, and not the opposite view, where it is the information that decides where to go.

The Ad-Loc project [1] proposes an annotation location-aware infrastructure-free system. Notes stick to an area of relevance which can grow depending on the location of interested nodes. Information is periodically broadcasted to neighbouring nodes. Nodes are the active entities exchanging information. The size of the area of relevance grows as necessary in order to accommodate the needs of users potentially far from the central location. The information then becomes eventually available everywhere.

7 Conclusion

In this paper we discussed the notion of hovering information, defined and simulated the Attractor Point algorithm which intends to keep the information alive and available in its anchor area. This algorithm multicasts hovering information replicas to the nodes that are closer to the anchor location. The performances of this algorithm have been compared to those of a Broadcast version. The results show that the Broadcast algorithm outperforms the Attractor Point algorithm in terms of availability but only from a very small factor. The proposed Attractor Point algorithm is much less bandwidth and memory greedy than the Broadcast algorithm and achieves higher levels of concentration of data in the anchor area.

Considering that these results constitute a proof of concept of the hovering information paradigm, future works will concentrate on releasing the assumption of limited memory and in considering not only one piece of hovering information but multiple distinct pieces all hovering in the same environment. We intend as well to take into account the speed and direction of the nodes when choosing the nodes that will host replicas. We have tested the Attractor Point algorithm under a Random Way Point mobility model and under ideal wireless conditions. This is not characteristic of real world behaviour. We will apply the Attractor Point algorithm to scenarios following real mobility patterns (e.g. crowd mobility patterns in a shopping mall or traffic mobility patterns in a city) with real wireless conditions (e.g. channel interferences or physical obstacles).

References

1. D. J. Corbet and D. Cutting. Ad loc: Location-based infrastructure-free annotation. In *ICMU 2006*, London, England, Oct. 2006.
2. S. Dolev, S. Gilbert, L. Lahiani, N. A. Lynch, and T. Nolte. Timed virtual stationary automata for mobile networks. In *OPODIS*, pages 130–145, 2005.
3. S. Dolev, S. Gilbert, E. Schiller, A. A. Shvartsman, and J. Welch. Autonomous virtual mobile nodes. In *DIALM-POMC '05: Proceedings of the 2005 joint workshop on Foundations of mobile computing*, pages 62–69, New York, NY, USA, 2005. ACM Press.
4. I. Leontiadis and C. Mascolo. Geopps: Opportunistic geographical routing for vehicular networks. In *Proceedings of the IEEE Workshop on Autonomic and Opportunistic Communications. (Colocated with WOWMOM07)*, Helsinki, Finland, June 2007. IEEE Press.
5. I. Leontiadis and C. Mascolo. Opportunistic spatio-temporal dissemination system for vehicular networks. In *MobiOpp '07: Proceedings of the 1st international MobiSys workshop on Mobile opportunistic networking*, pages 39–46, New York, NY, USA, 2007. ACM Press.
6. A. Villalba, G. Di Marzo Serugendo, and D. Konstantas. Hovering information - self-organising information that finds its own storage. Technical Report BBKCS-07-07, School of Computer Science and Information Systems, Birkbeck, University of London, Nov 2007.
7. A. Villalba and D. Konstantas. Towards hovering information. In *Proceedings of the First European Conference on Smart Sensing and Context (EuroSSC 2006)*, pages 161–166, 2006.

Hovering Information - Infrastructure-Free Self-Organising Location-Aware Information Dissemination Service[1]

Alfredo A. Villalba Castro, Giovanna Di Marzo Serugendo, and Dimitri Konstantas

Abstract. This paper proposes a location-based service for disseminating geo-localised information generated by and aimed at mobile users. The service itself works in a self-organising manner. A piece of hovering information is attached to a geographical point, called the anchor location, and to its vicinity area, called the anchor area. It is responsible for keeping itself alive, available and accessible to other devices within its anchor area. Hovering information uses mechanisms such as active hopping, replication and dissemination among mobile nodes to satisfy the above requirements. It does not rely on any central server. Previous results involving a single piece of hovering information have shown the interest of the concept. This paper reports on a series of simulations involving multiple pieces of hovering information. Our goal is to investigate the scalability of the technique up to 200 pieces in a small geographic area. Two main replication algorithms for pieces of hovering information are compared, an Attraction Point algorithm and a Broadcast-based one. These replication algorithms are combined with two different caching policies, Location-based and Generation-based, for discarding hovering information pieces from mobile nodes buffer when memory is not enough.

1 Introduction

The last two decades were marked by a rapid evolution of computer and communication related technologies available to the end-users. From the 300MHz processors available in the late 80s, today we are above 4GHz, and from 56Kbps networks (modems), we have today home networks of more than 20Mbps. In a similar way mobile storage capacities available to end-users have gone up from 1MB diskettes to 8GB memory sticks. The average end user today has more wireless network, processing power and storage capacity available to a mobile device, like PDA or smartphone, than the semi-professional user of the late 80s. It is thus safe to assume that at the end of the 2010s the average user will have a mobile device with more than 4GHz processing power, 100Mbps wireless network connectivity and more than 1TB of local storage capacity, all supported by powerful power supply, allowing him to have a continuous high bandwidth network connection for periods longer than 24 hours. We can thus also anticipate that the mobile available storage capacity of the end-users will be an important percentage of the fixed storage capacity on the planet. Furthermore we can expect that, mobile devices like phones and PDAs will be equipped with high quality geo-localisation hardware (like GPS and Galileo chips).

Besides these technological advances, we have witnessed during the last few years a new direction taken by the end-users in the creation of information. With the

[1] This paper has been presented and included in the proceedings of 2nd ERCIM Workshop on eMobility in conjuntion with WWIC 2008, 28-30 May 2008, Tampere, Finland.

available communication means, end-users are changing their behaviour from information consumers to information producers. More and more information created by end-users is becoming available on the internet, as it is observed by the big success of sites like YouTube and FaceBook. We can thus anticipate that for the next decade end-users will keep on producing even more content, making it available to other users in different forms and under different means.

Considering the above predictions for the next decade, we can sketch a daily scenario for the average user of the late 2010s. The user is equipped with a mobile device through which he accesses location related information that was created by him, friends, or even strangers. All this information is stored primarily into his mobile device and may become available to other users without passing by a mobile operator or a centralized server, instead setting up a mobile ad hoc network and taking advantage of the vast amounts of mobile storage capacities available. In this direction, we have defined the concept of Hovering Information, which provides a promising solution by creating the basis and models for large-scale, location related information management. Instead of accessing location-based information via a wireless network operator, the Hovering Information concept will allow mobile devices to disseminate the information among them, thus relieving the load of the wireless networks. Applications such as stigmergy-based systems, traffic management over vehicle networks or mounting distributed information systems on disaster areas could be implemented using hovering information.

This paper presents the hovering information concept, a replication algorithm allowing single pieces of hovering information to get attracted to their respective geographical related locations, called anchor locations; a broadcast-based algorithm for comparing network and memory usage performances; and two mobile nodes caching policies, Location-based and Generation-based, for discarding pieces of hovering information when memory space is full. A complete formal description of the hovering information model is described in [14].

2 Hovering Information Concept

This section describes the main concepts of a hovering information system: mobile nodes, hovering information; and three main dependability requirements of hovering information: survivability, availability and accessibility.

2.1 Mobile Nodes and Hovering Information

Mobile nodes represent the storage and motion media exploited by pieces of hovering information. A *mobile node* n is defined as a tuple:

$$n = (id, loc, speed, dir, r_{comm}, buff),$$

where id is its mobile node identifier, loc is its current location (a geographic location), $speed$ is its current speed in m/s, dir is its current direction of movement (a geographic vector), r_{comm} is its wireless communication range in meters and $buff$ is its buffer (having a limited size) aimed to store replicas of the pieces of hovering information.

A piece of hovering information is a piece of data whose main goal is to remain stored in an area centred at a specific location called the *anchor location*, and having a radius called the *anchor radius*. A *piece of hovering information* h is defined as a tuple:

$$h = (id, a, r, n, data, policies, size),$$

where id is its hovering information identifier, a is its anchor location (geographic coordinate), r is its anchor radius in meters, n is the mobile node where h is currently hosted (hosting node), $data$ is the data carried by h, $policies$ are the hovering policies of h and $size$ is the size of h in bytes. Policies stand for hovering policies stating how and when a piece of hovering information has to hover.

We consider that identifiers of pieces of hovering information are unique, but replicas (carrying same data and anchor information) are allowed on *different* mobile nodes. We also consider that there is only one instance of a hovering information in a given node n, any other replica resides in another node.

A hovering information system is composed of mobile nodes and pieces of hovering information. A hovering information system at time t is a snapshot (at time t) of the status of the system. We denote by $\mathcal{N}_t$ the set of mobile nodes at time t. Mobile nodes can change location, new mobile nodes can join the system and others can leave. New pieces of hovering information can appear (with new identifiers), replicas may appear or disappear (same identifiers but located on other nodes), hovering information may disappear or change node.

Figure 1 shows two different pieces of hovering information h_1 (blue) and h_2 (green), having each a different anchor location and area. Three replicas of h_1 are currently located in the anchor area (in three different mobile nodes n_2, n_3 and n_4), while two replicas of h_2 are present in the anchor area of h_2 (in nodes n_2 and n_5). It may happen that a mobile node hosts replicas of different pieces of hovering information, as it is the case in the figure for the mobile node n_2 that is at the intersection of the two anchor areas. The arrows here also represent the communication range possibilities among the nodes.

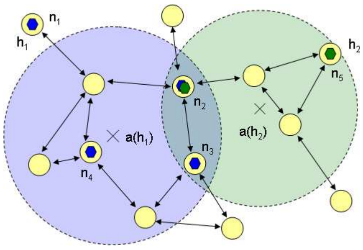


Fig. 1. Hovering Information System at time t

2.2 Properties - Requirements

Survivability. A hovering information h is alive at some time t if there is at least one node hosting a replica of this information. The survivability along a period of time is defined as the ratio between the amount of time during which the hovering information has been alive and the overall duration of the observation.

Availability. A hovering information h is available at some time t if there is at least a node in its anchor area hosting a replica of this information. The availability of a piece of hovering information along a period of time is defined as the rate between the amount of time along which this information has been available during this period and the overall time.

Accessibility. A hovering information is accessible by a node n at some time t if the node is able to get this information. In other words, if it exists a node m being in the communication range of the interested node n and which contains a replica of the piece of hovering information. The accessibility of a piece of hovering information h is the rate between the area covered by the hovering information's replicas and its anchor area.

The interested reader can refer to [14] for a full set of definitions.

3 Algorithms for Hovering Information

In [15] we studied the performances in terms of availability of a hovering information system containing many replicas of *only one* piece of hovering information. We assumed that each node had a buffer with an unlimited amount of memory for storing replicas. Therefore, the proposed replication algorithms should not have had to cope with buffer overflows problems when a new incoming replica arrived. In this paper, we drop the assumption of unlimited memory and we study a hovering information system containing *multiple* (distinct) hovering information and their respective replicas. Instead of keeping an unlimited buffer size for storing replicas, we limit the size of the buffer. The need for caching policies becomes then important when it is time to insert a new incoming replica.

In this paper, we propose and study two different caching policies: Location-Based Caching (LBC) and Generation-Based Caching (GBC). The LBC policy decision to erase a replica is based on the proximity of that replica to its anchor location and on the portion of the surface of the anchor area covered by the communication area of the node hosting it. The GBC policy takes the decision of removing a replica based on the generation of the replica, removing replicas having the oldest generations.

Assumptions. We make the following assumptions in order to keep the problem simple while focusing on measuring availability and resource consumption. *Uniform size:* All pieces of hovering information have the same size and the caching algorithms do not take in consideration the size as a criteria when removing a replica. *Unlimited energy:* All mobile nodes have an unlimited amount of energy. The proposed algorithms do not consider failure of nodes or impossibility of sending messages because of low level of energy. *In-built geo-localization service:* Mobile nodes have an in-built geo-localization service such as GPS which provides the current position. We assume that this information is available to pieces of hovering information. *Neighbours discovering service:* Mobile nodes are able to get a list of their current neighbouring nodes at any time. This list contains the position, speed, and direction of the nodes. As for the other two services, this information is available to pieces of hovering information.

3.1 Safe, Risk and Relevant Areas

In this paper we consider that all pieces of hovering information have the same hovering policies: active replication and hovering in order to stay in the anchor area (for survivability, availability and accessibility reasons), caching when there is no free space to store a replica, and cleaning when too far from the anchor area to be meaningful (i.e. disappearance). The decision on whether to replicate itself or to hover depends on the current position of the mobile node in which the hovering information is currently stored. Therefore, we distinguish three different areas: safe area, risk area and relevant area.

A piece of hovering information located in the *safe area* can safely stay in the current mobile node, provided the conditions on the node permit this: power, memory, etc. This area is defined as the disc having as centre the anchor location and as radius the safe radius (r_{safe}).

A piece of hovering information located in the *risk area* should actively seek a new location on a mobile node going into the direction of the safe area. It is in this area that the hovering information actively replicates itself in order to survive and stay available in the vicinity of the anchor location. This area is defined as the ring having as centre the anchor location and bound by the safe and risk radii (r_{risk}).

The *relevant area* limits the scope of survivability of a piece of hovering information. This area is defined as the disc whose centre is the anchor location and whose radius is the relevant radius (r_{rele}). The *irrelevant area* is all the area outside the relevant area. A piece of hovering information located in the irrelevant area can disappear; it is relieved from survivability goals.

All these radii cope with the following inequality (where r is the anchor radius):

$$r_{safe} < r < r_{risk} < r_{rele}$$

The values of these different radii are different for each piece of hovering information and are typically stored in the Policies field of the hovering information. In the following algorithms we consider that all pieces of hovering information have the same relevant, risk and safe radius.

3.2 Replication

A piece of hovering information h has to replicate itself onto other nodes in order to stay alive, available and accessible. We describe two replication algorithms simulating two variants of the replication policies: the Attractor Point and Broadcast-based algorithms. Both algorithms are triggered periodically - each T_R (replication time) seconds - and only replicas of h being in the risk area are replicated onto some neighbouring nodes (nodes in communication range) which are selected according to the replication algorithm.

When replicas consider themselves too far from their anchor area and not able to come back anymore, the cleaning mechanism periodically - each T_C (cleaning time) seconds - and for each node, removes the replicas that are too far from their anchor location, i.e. those replicas that are in the irrelevant area. This avoids as well the situation where all nodes have a replica.

Attractor Point Algorithm (AP)

The anchor location of a piece of hovering information acts constantly as an attractor point to that piece of hovering information and to all its replicas. Replicas tend to stay as close as possible to their anchor area by jumping from one mobile node to other. The number of target nodes composing the multicast group that will receive a replica is defined by the constant k_R (replication factor).

Figure 2(a) illustrates the behaviour of the Attractor Point algorithm. Consider a piece of hovering information h in the risk area. It replicates itself onto the nodes in communication range that are the closest to its anchor location. For a replication factor $k_R = 2$, nodes n_2 and n_3 receive a replica, while all the other nodes in range do not receive any replica.

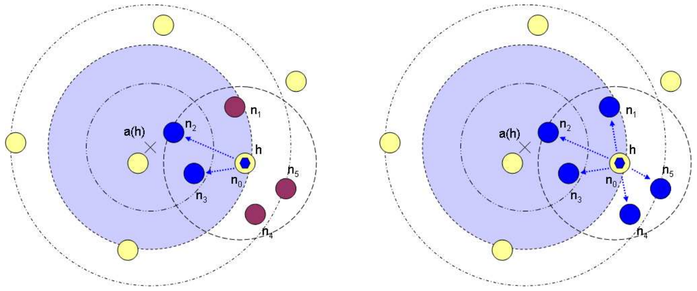
Broadcast-based Algorithm (BB)

The Broadcast-based algorithm is triggered periodically (each T_R) for each mobile node. After checking the position of the mobile node pieces of hovering information located in the risk area are replicated and broadcasted onto *all* the nodes in communication range. We expect this algorithm to have the best performance in terms of availability but the worst in terms of network and memory resource consumption.

Figure 2(b) illustrates the behaviour of the Broadcast-based algorithm. Consider the piece of hovering information h in the risk area, it replicates itself onto all the nodes in communication range, nodes n_1 to n_5 (blue nodes).

3.3 Caching

In this paper we assume that nodes have a limited amount of memory to store the pieces of hovering information (replicas). As the number of distinct hovering information increases, so will be the total number of replicas. The buffer of nodes will get full at some point and some replicas should have to be removed in order to store new ones. We present two different caching policies: Location-Based and Generation-Based Caching. We compare these caching techniques with a simpler one which only ignores the incoming replicas as soon as there is no free space in the mobile device buffer.



(a) Attractor Point Algorithm (b) Broadcast-based Algorithm

Fig. 2. Replication Algorithms

Besides these caching algorithms, it is important to notice that we only consider the position and the generation of replicas. We do not take into consideration caching policies such as the priority, the time-to-live or the replicas size (since all replicas considered in this paper have the same size).

Location-Based Caching (LBC)

At each node, this caching policy decides to remove a previously stored replica from the node's full buffer, and to replace it by the new incoming replica based on their respective location relevance value. We define the location relevance value of a replica, being this replica already stored in the node's buffer or being a new incoming replica, to its anchor location and area as it follows:

$$relevance = \alpha * area + \beta * proximity,$$

where $area$ is the normalised estimation of the overlapping area of the nodes' communication range area and the replica's anchor area, $proximity$ is the normalised proximity value between the current position of the node and the anchor location of the replica, α and β are real coefficients having values between 0 and 1 and $\alpha + \beta = 1$.

Each time a new incoming replica arrives, the least location relevant replica is chosen from all the replicas stored in buffer of the node. The location relevance of the incoming replica is computed and compared to that of the least location relevant replica, whatever the original hovering information they refer to. The least location relevant replicas is removed from the buffer and replaced by the incoming replica if the latter has a greater location relevance value. Otherwise, the incoming replica is just discarded. In this way, the location-based caching algorithm will tend to remove replicas being too far from their anchor location or being hosted in a node covering only a small part of their anchor area.

Generation-Based Caching (GBC)

We define the generation of a replica in the following way: the first replica created (normally by the user or user application) of a piece of hovering information has a generation 0, when this replica replicates itself then it creates new replicas having generation 1, and so on. The generation of a replica gives us an idea of the number of replicas existing as the process of replication follows an exponential growth. The generation-based caching algorithm tends to remove replicas having a high generation number as they are likely more replicas leaving around than a replica having a lower generation number.

Each time a new incoming replica arrives, the oldest replica (the one having the highest generation value) is chosen from all the replicas stored in the buffer of the node. The generation of the incoming replica is retrieved and compared to that of the oldest replica, whatever the original hovering information they refer to. The oldest replica is removed from the buffer and replaced by the incoming replica if the latter has a smaller generation value. Otherwise, the incoming replica is just discarded.

4 Evaluation

We evaluated the behaviour of the above described replication and caching algorithms under different scenarios by varying the number of pieces of hovering information, each having many replicas of its. We also considered nodes having a limited buffer size to store the different replicas. We have measured the average availability, message complexity, replication complexity, overflows and erased replicas over the total number of pieces of hovering information existing in the system.

We performed simulations using the OMNet++ network simulator (distribution 3.3) and its Mobility Framework 2.0p2 (mobility module) to simulate nodes having a simplified WiFi-enabled communication interfaces (not dealing with channel interferences) with a communication range of 121m.

4.1 Simulation Settings and Scenarios

The generic scenario consists of a surface of 500m x 500m with mobile nodes moving around following a Random Way Point mobility model with a speed varying from 1m/s to 10m/s without pause time. In this kind of mobility model, a node moves along a straight line with speed and direction changing randomly at some random time intervals.

In the generic scenario, pieces of hovering information have an anchor radius (r) of 50m, a safe radius (r_{safe}) of 30m, a risk radius (r_{risk}) of 70m, a relevance radius (r_{rele}) of 200m, and a replication factor of 4 (k_R).

Each node triggers the replication algorithm every 10 seconds (T_R) and the cleaning algorithm every 60 seconds (T_C). Each node has a buffer having a capacity to store 20 different replicas. The caching algorithm is constantly listening for the arrival of new replicas.

Based on this generic scenario, we defined 5 specific scenarios with varying number of pieces of hovering information: from 40 to 200 nodes, increasing the number of pieces by 40. Each of this scenarios has been investigated with different replication and caching algorithms, and with a different number of nodes.

We have performed 20 runs for each of the above scenarios. One run lasts 3'600 simulated seconds. All the results presented here are the average of the 20 runs for each scenario, and the errors bars represent a 95% confidence interval. All the simulations ran on a Linux cluster of 32 computation nodes (Sun V60x dual Intel Xeon 2.8GHz, 2GB RAM).

4.2 Results

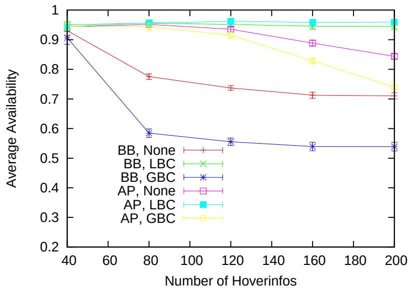


Fig. 3. Availability - 200 Nodes

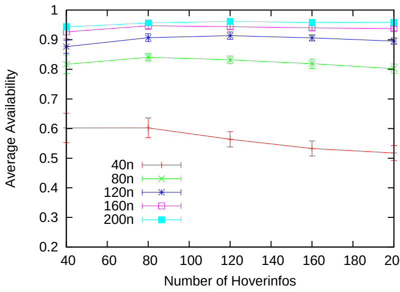


Fig. 4. Availability - AP with LBC

Figure 3 shows the average availability for the AP and BB algorithms using LBC or GBC caching policies, or without using any (None). For this experiment we used 200 nodes. We observe that both algorithms using LBC outperform the cases using GBC or no caching policy (None). We can also notice that the BB algorithm gets worse availability performances compared to AP. This is explained by the nature of the BB algorithm which tends to overload the system with an exponential growing number of replicas. As the buffer size at each node has a limited size of 20 replicas, the overloading causes the buffer resources to become badly shared by the pieces of hovering information, in particular regarding their individual target anchor location. Consequently the availability becomes lower. On the other hand, the AP algorithm controls the replication process by limiting the number of replicas and by focusing on the anchor location. When it is combined with the LBC caching policy, the system becomes scalable keeping high availability rates (around 95%) as the number of pieces of hovering information increases.

Figure 4 depicts the average availability of the AP replication algorithm using the LBC caching policy, under several number of nodes. For a number of nodes above 120, we notice that the availability is high enough (above 85%) and it keeps quite stable as the number of pieces of hovering information increases. We confirm from this that the AP with LBC algorithms are scalable in terms of absorption of hovering information (number of distinct pieces of hovering information), since during the experiments with 120 nodes and more, up to 200 distinct hovering information pieces have been accommodated into the system with an availability above 85%. In the case when the number of nodes is 40, we can observe that the absorption limit, for this configuration, has been reached as the availability starts decreasing after 80 pieces of hovering information.

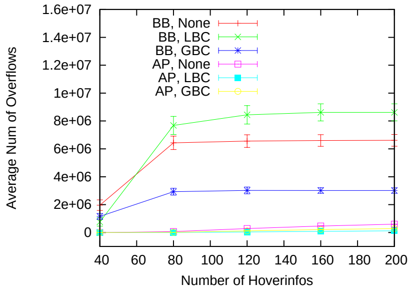


Fig. 5. Overflows - 200 Nodes

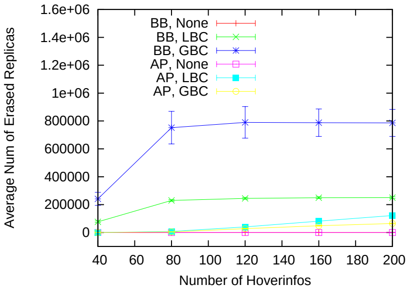


Fig. 6. Erased Replicas - 200 Nodes

Figure 5 shows the average number of overflows generated by the different replication and caching algorithms each time a new incoming replica arrives and there is no free space to store it. We observe that the AP algorithm produces at least ten times less overflows than the BB algorithm. We also observe that the number of overflows of the BB algorithm starts growing fast and then it gets stable after for 80 pieces of hovering information or more, this is again a consequence of the overloading of the system which tends to lose pieces of hovering information by not distributing the buffer resources in a fair way.

Figure 6 illustrates the average number of erased replicas from the buffer of the nodes after an overflow. Again, the BB erases around 10 times much more replicas than the AP algorithm to store a new incoming replica. We also notice that the BB with GBC tends to erase too many replicas compared to the other ones; this means that the generation-based caching policy combined with the exponential replication behaviour of BB is not a good differentiation factor for caching replicas since this combination of algorithms tends to insert and erase replicas permanently.

Figure 7 shows the average number of sent messages for the AP algorithm using the LBC policy under several numbers of nodes. We notice that the message complexity does not grow exponentially. Instead, it grows linearly of even logarithmically with the number of pieces of hovering information. This is a very important issue as the feasibility of these algorithms depends strongly on the messages complexity, especially when we will need to deal with the interference channel for even more realistic network interfaces.

Finally, figure 8 shows the average of the maximal number of replicas having existed in the system for the AP using the LBC. It is interesting to see that it decreases as the number of pieces of hovering information increases. It means that the

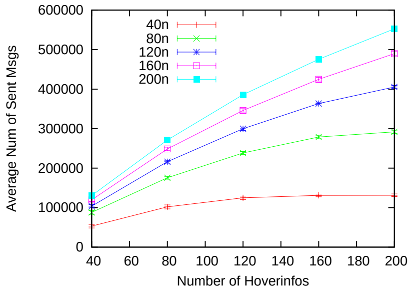


Fig. 7. Messages Complexity - Attractor Point with Location-Based Caching

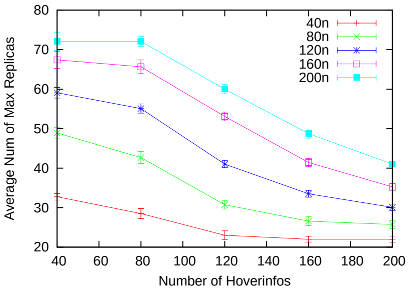


Fig. 8. Replication Complexity - Attractor Point with Location-Based Caching

buffer resources are evenly shared among the different pieces of hovering information, while the availability still remains at high levels (see Figure 4). We conclude from this, that the AP with LBC succeeds to distribute the network resource in a fair way among all the pieces of hovering information, and that we probably observe an emergent (not coded) load-balancing of the memory allocated to the different pieces of hovering information.

5 Related Works

To our knowledge, while there exist other information annotation/dissemination works closely related to the hovering information concept, they do not take the approach of offering a generic infrastructure-free location-aware information dissemination service as hovering information does. The Hovering Data Clouds (HDC) concept [16, 8], which is part of the AutoNomos project, is applied to the specific design of a distributed infrastructure-free car traffic congestion information system. Although HDCs are defined as information entities having properties similar to hovering information, the described algorithms do not consider them as an independent service but as part of the traffic congestion algorithms. The hovering information dissemination service is thought as a service independent from the applications using it. The Ad-Loc system [1] is an infrastructure-free location-aware annotation system and shares similarities to hovering information. However, this approach does not focus on: studying properties such as the critical number of nodes or the absorption limits; or dealing with self-organizing algorithms allowing the information to adapt its behaviour according to the network saturation, the buffers' size, the mobility pattern of nodes or the number of replicas.

The opportunistic spatio-temporal dissemination service over MANETs [11] is a car traffic centred application and does not encompass ideas such as recombination of information. Similarly, works like Epcast [13] and Gossip [3] aim to disseminate information based on epidemic and gossiping spreading models. These works provide an interesting starting point for replication algorithms, but do not offer a solution for ensuring the persistency of the information.

In the domain of location-driven routing over MANETs, we can mention works such as GeoOpss [10], search and query propagation over social networks like PeopleNet [12] and collaborative services such as collaborative backup of the MoSAIC project [9, 2].

Finally, the virtual infrastructure project [4, 5, 6, 7] aims to set up a set of virtual nodes having a well-know structure and trajectory over a mobile ad hoc network. These virtual nodes are equipped with a clocked automaton machine which will permit to implement distributed algorithms such as leader election, routing, atomic memory, motion coordination, etc. This approach works on offering a structured abstraction layer of virtual nodes. Hovering information takes a different approach where each piece of hovering information is an autonomous entity responsible for its own survivability exploiting the dynamics of the overlay network to this aim.

6 Conclusion

In this paper we discussed the notion of hovering information, we defined and simulated the Attractor Point algorithm which intends to keep the information alive and available in its anchor area. This algorithm multicasts hovering information replicas to the nodes that are closer to the anchor location. The performances of this algorithm have been compared to those of a broadcast version.

We have also defined and simulated two different caching polices, the Location-Based Caching and the Generation-Based Caching. Their performances have been compared under a scenario containing multiple pieces of hovering information and nodes having a limited amount of memory.

Results show that the Atrractor Point algorithm with the Location-Based Caching policy is scalable in terms of the number of pieces of hovering information that the system can support (absorption limits). They also show the emergence of a load-balancing property of the buffer usage which stores replicas in an optimal way as the number of pieces of hovering information increases.

Concerning future work, we have tested the algorithms under a Random Way Point mobility model and under ideal wireless conditions. This is not characteristic of real world behaviour. We will apply the different algorithms to scenarios following real mobility patterns (e.g. crowd mobility patterns in a shopping mall or traffic mobility patterns in a city) with real wireless conditions (e.g. channel interferences or physical obstacles).

References

1. D. J. Corbet and D. Cutting. Ad loc: Location-based infrastructure-free annotation. In *ICMU 2006*, London, England, Oct. 2006.
2. L. Courts, M.-O. Killijian, D. Powell, and M. Roy. Sauvegarde coopprative entre pairs pour dispositifs mobiles. In *UbiMob '05: Proceedings of the 2nd French-speaking conference on Mobility and uibquity computing*, pages 97–104, New York, NY, USA, 2005. ACM Press.

3. A. Datta, S. Quarteroni, and K. Aberer. Autonomous gossiping: A self-organizing epidemic algorithm for selective information dissemination in mobile ad-hoc networks. In *IC-SNW'04, International Conference on Semantics of a Networked World*, LNCS, pages 126–143, 2004.
4. S. Dolev, S. Gilbert, L. Lahiani, N. A. Lynch, and T. Nolte. Timed virtual stationary automata for mobile networks. In *OPODIS*, pages 130–145, 2005.
5. S. Dolev, S. Gilbert, N. A. Lynch, E. Schiller, A. A. Shvartsman, and J. L. Welch. Virtual mobile nodes for mobile ad hoc networks. In *DISC*, 2004.
6. S. Dolev, S. Gilbert, N. A. Lynch, A. A. Shvartsman, and J. Welch. Geoquorums: Implementing atomic memory in mobile ad hoc networks. In *DISC*, 2003.
7. S. Dolev, S. Gilbert, E. Schiller, A. A. Shvartsman, and J. Welch. Autonomous virtual mobile nodes. In *DIALM-POMC '05: Proceedings of the 2005 joint workshop on Foundations of mobile computing*, pages 62–69, New York, NY, USA, 2005. ACM Press.
8. S. P. Fekete, , C. Schmidt, A. Wegener, and S. Fischer. Hovering data clouds for recognizing traffic jams. In *Proceedings 2nd International Symposium on Leveraging Applications of Formal Methods, Verification and Validation (IEEE-ISOLA)*, pages 213–218, 2006.
9. M.-O. Killijian, D. Powell, M. Banâtre, P. Couderc, and Y. Roudier. Collaborative backup for dependable mobile applications. In *MPAC '04: Proceedings of the 2nd workshop on Middleware for pervasive and ad-hoc computing*, pages 146–149, New York, NY, USA, 2004. ACM Press.
10. I. Leontiadis and C. Mascolo. Geopps: Opportunistic geographical routing for vehicular networks. In *Proceedings of the IEEE Workshop on Autonomic and Opportunistic Communications. (Colocated with WOWMOM07)*, Helsinki, Finland, June 2007. IEEE Press.
11. I. Leontiadis and C. Mascolo. Opportunistic spatio-temporal dissemination system for vehicular networks. In *MobiOpp '07: Proceedings of the 1st international MobiSys workshop on Mobile opportunistic networking*, pages 39–46, New York, NY, USA, 2007. ACM Press.
12. M. Motani, V. Srinivasan, and P. S. Nuggehalli. Peoplenet: engineering a wireless virtual social network. In *MobiCom '05: Proceedings of the 11th annual international conference on Mobile computing and networking*, pages 243–257, New York, NY, USA, 2005. ACM Press.
13. S. Scellato, C. Mascolo, M. Musolesi, and V. Latora. Epcast: Controlled dissemination in human-based wireless networks by means of epidemic spreading models. *CoRR*, abs/0711.2780, 2007.
14. A. Villalba Castro, G. Di Marzo Serugendo, and D. Konstantas. Hovering information - self-organising information that finds its own storage. Technical Report BBKCS-07-07, School of Computer Science and Information Systems, Birkbeck, University of London, Nov 2007.
15. A. Villalba Castro, G. Di Marzo Serugendo, and D. Konstantas. Hovering information - self-organising information that finds its own storage. In *IEEE International Conference on Sensors, Ubiquitous and Trust Computing (SUTC'08)*, 2008.
16. A. Wegener, E. M. Schiller, H. Hellbrck, S. P. Fekete, and S. Fischer. Hovering data clouds: A decentralized and self-organizing information system. In *International Workshop on Self-Organizing Systems*, pages 243 – 247, 2006.

PlayMancer[1]: A European Serious Gaming 3D Environment

Elias Kalapanidas, Hikari Watanabe, Costas Davarakis, Hannes Kaufmann, Fernando Fernandez Aranda, Tony Lam, Todor Ganchev and Dimitri Konstantas

Abstract. Serious games are about to enter the medical sector to give people with behavioural or addictive disorders the ability to use them as part of health promotion and disease prevention. The PlayMancer framework will support physical rehabilitations and psycho-education programs thru a modular multiplayer networked 3D game based on the Universally Accessible Games (UA games) guidelines.

1 Introduction

The potential of games for entertainment and learning has been demonstrated thoroughly from both research and market. Unfortunately, the investments committed to entertainment dwarfs what is committed for more serious purposes. Furthermore, game development has become more complex, expensive, and burdened with a long development cycle. This creates barriers to independent game developers, and inhibits the introduction of innovative games, or new game genres, i.e. serious games, or games accessible to communities with special needs. The aim of the PlayMancer project is to implement a framework and a platform for serious games. PlayMancer will implement a new Serious Game environment, by augmenting existing 3D gaming engines with new possibilities. The objectives of the project are four-fold:

- To construct a next generation networked gaming environment, mainly augmenting the gaming experience with innovative information and communication technologies (ICT) modes of interaction between the player and the game world,
- to allow for a shorter and most cost-effective game production chain, by enabling techniques for procedural content creation based on generative modelling, and thus reduce the cost of offering a full-fledged pre-designed gaming world,
- to evolve the principles of Universally Accessible Games (UA games) [1] for application into 3D-based games, following a design for all philosophy, with the

[1] The work described in this paper is supported by the Commission of the European Union in the frame of the ICT FP7 research program under project PlayMancer FP7-215839.

ultimate goal of designing games to be equally challenging to players of different abilities and

- to evaluate the proposed framework and gaming infrastructure by developing and testing a series of serious games modules as applied to two application domains: physical rehabilitation, and therapeutic support and lifestyle management programs for behavioural and addictive disorders.

The driving applications for the project will be physical rehabilitation and lifestyle related disorders. Physical rehabilitation will drive platform requirements for supporting the development of UA games and the integration of low cost player motion tracking and bio-feedback devices. Games scenarios from the lifestyle related disorder will implicate platform requirements for emotion recognition of states such as boredom, depression, anxiety and associated cognitive responses. Due to the modular nature of the envisioned PlayMancer gaming platform architecture and the commitment to Design-for-All philosophy, the project results could be generalised to other serious games applications and user communities.

2 Concept

Serious games (SGs) or persuasive games are computer and video games used as educational technology or as a vehicle for presenting or promoting a point of view. They can be similar to educational games, but are often intended for an audience outside of primary or secondary education. Serious games can be of any genre and many of them can be considered a kind of edutainment. The serious games are intended to provide an engaging, self-reinforcing context in which to motivate and educate the players towards non-game events or processes, including business operations, training, marketing and advertisement. Serious games can be compelling, educative, provocative, disruptive and inspirational.

The PlayMancer project is conceived to take advantage of the current market momentum towards a next generation gaming platform. After the mass-market adoption of 3D graphics acceleration cards due to the recent game technology advancement, the PlayMancer consortium anticipates a similar trend to happen with new interaction modes in the near future. The game core (the PlayMancer platform) will be built by encompassing an Augmented Reality (AR) 3D game-like world. AR has been used so far for scientific applications with success, but due to the fact that these applications are very specialized, AR constitutes a very limited market and as a consequence the average cost of AR products are exaggeratedly high for an average game user. However, we believe that reducing the games production cycle will accelerate create an economy of scale that will drive down the cost and increase the availability of games, games components and technology. Thus it is expected that the cost of AR technology will be considerably dropped, if future game platforms massively exploit them, the same way that 3D games did for 3D acceleration card costs.

In remote constrained physical spaces (distributed playgrounds), multiple players will use an accessible navigation interface through as many information channels as possible: speech, emotional cues, simple gestures, tactile devices, haptic sensitive devices, 3D shapes and volumes. Using a full-scale spoken dialogue system that includes speech recognition and understanding, speaker recognition, emotion detection, user modelling, context awareness, dialogue flow modelling, natural language generation, text-to-speech (TTS), etc, would enable advanced human-machine interaction and eventually higher level of user satisfaction. According to the game plot, the players will have to navigate through the virtual universe, interact with AI-motivated digital characters, with each other, make plans, or just watch the narration of the story. The games will be structured around micro-goals, which will form missions and quests. By combining different interaction patterns, the players will be able to achieve these goals. An indicative scenario including such a micro-goal follows: A player's avatar has dropped off his wheel chair. In order to sit again on the chair, the player should: be calm (self-controlled), move to the chair, grab the chair and then lift himself and put himself on the chair. The game will judge calmness by getting and processing input from bio-feedback regarding pulse and heart rate, and then each of the motion actions based on gait monitoring and motion tracking.

3 General description

The project will follow a user-centred design cycle throughout its development engaging the different user groups in all the stages of design. For determining the proper interaction mechanisms that users would like, two iterations will conduct definition, specification, implementation, integration and utilization on a set of different interface configurations, enabling different setups of the available modes and information sources: spoken dialogue interaction, recognition of motion gestures, recognition of emotional states based on speech, bio-feedback. User-centred design implies:

- Early focus on users, tasks and environment,
- active involvement of users,
- an appropriate allocation of function between user and system,
- incorporation of user-derived feedback into system design,
- iterative design whereby a prototype is designed tested and modified.

Project iterations (lifecycles) have been designed according to ISO 13407 (Human-Centred Design Processes for Interactive Systems).

3.1 PlayMancer system architecture and system components

PlayMancer will exploit open-source 3D game engines, such as OGRE [2] and CrystalSpace [3], and develop code and tools in order to deliver a new advanced computer game platform. The overall architecture will follow the layers in figure 1:

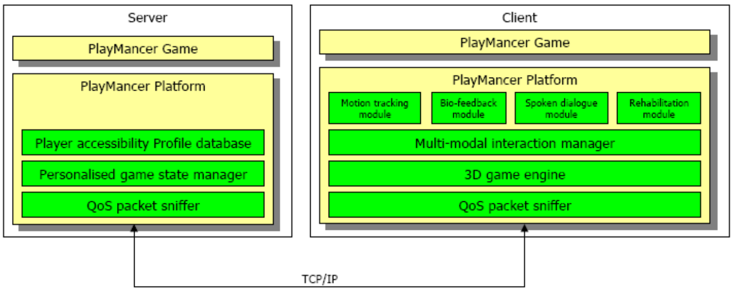


Fig. 1. Deployment of PlayMancer components over the network

3.2 Motion capture systems

In order to enable full body interaction in networked environments, an affordable motion capture system will be developed. Marker-based optical motion capture has become a de facto standard capture technology in the movie and entertainment industry in the previous ten years. Thereby one or more actors wearing retro-reflective markers attached to them, are tracked by a number of cameras (minimum six). All motions are computed in real time (with millimetre accuracy) and are available for further processing e.g. recording, analyzing, motion transfer to a virtual character and more.

The hardware platform will be based on TUW's work [4] of an existing accurate, fast and affordable infrared-optical tracking system. Choosing commodity hardware over custom-built components has always been a reliable cost-minimization strategy.

Due to lower costs there are a number of fields that would dramatically benefit from affordable motion-capture including rehabilitation clinics (e.g. gait analysis, stroke patient therapy and many more) and independent biomedical researchers in many fields. Even veterinary clinics could use accessible motion-tracking systems to examine animal gaits and behaviours for diagnosis.

3.3 Bio-feedback interaction

In recent years bio-feedback has become increasingly important as a non-classical user interface. Especially in medical applications where users' biosignals are of vital importance various sensors have been integrated and used for feedback to patients and medical personnel.

Within the PlayMancer game biosignals will give important indications on a patient's medical condition, his motivation, excitement and engagement. These user input signals will be integrated into the platform, providing user interaction in a non-classical way. The game itself will respond to these signals and provide feedback accordingly. In order to analyze multiple biosignals and to research which sensor modalities can best be utilized within the project, a mobile biosignal acquisition device g.MOBIlab (from g.tec) will be acquired. It comes with four

EEG/EOG, two ECG/EMG channels, four digital channels and two analog inputs which can be used for other sensors. This allows the investigation of brain-, heart-, muscle-activity, eye movement, respiration, galvanic skin response, pulse and other body signals.

Wireless connection to the game client provides sufficient flexibility to analyze biosignals in the different PlayMancer game scenarios.

3.4 Gait rehabilitation infrastructure

Gait therapy might be necessary after neurological injuries of patients with movement disorders caused by stroke, spinal cord injury and traumatic brain injury, multiple sclerosis or Parkinson’s disease. Other causes can be orthopaedic injuries such as fractures in the leg or foot. In Germany for instance 250.000 people suffer from stroke every year. Modern concepts of motor learning favour a task specific training, i.e. to relearn walking, the patient has to walk repetitively in a correct manner.

Up to date a small number of robotic treadmill devices are available on the market (e.g. Lokomat, HapticWalker,…) which are used for gait therapy in rehabilitation clinics throughout Europe. One disadvantage of existing devices is that they provide no kind of visual feedback to the patient. Feedback, especially positive feedback is very important to encourage and motivate patients to continue training. The motion capture module of the PlayMancer platform will be used to capture and conduct gait motion analysis. A PlayMancer game will be developed to provide important positive visual feedback to the patient. Bio-feedback monitoring will be used in addition to interface with the game.

3.5 QoS packet sniffing and adaptive state synchronisation

The traditional game topology of networked games is client-server, with most to all sensitive data kept server-side. An architecture commonly used for such applications is a client-server object replication system. In this system, specific objects and data members are transmitted on creation or changes. This can happen explicitly, with a call to serialize and replicate, or implicitly, where data is polled every update cycle and compared to the value it had the last update cycle. This is a good approach that is both straightforward and easy to understand. It fits into existing single player systems. With implementations that replicate objects automatically, scripts can modify data without knowledge of the underlying system and still achieve network synchronization. However, it is less efficient than a hand-tweaked system, sending data independent of context and often redundantly. It is also less secure, creating a strong coupling between game-code and network activity. As game-code changes are made, especially by programmers that aren't familiar with network security, vulnerabilities arise. Aiming at improving the efficiency of network usage in networked games, PlayMancer will introduce a technique for adaptive game state synchronization that exploits a quality-of-service packet sniffing component.

3.6 Dialogue systems

The increasing use of spoken dialogue systems raises the need for more effective and user friendly interaction between human and machine. Most of the dialogue systems implemented in the past years do not take advantage of the knowledge of the emotional state of the user. However, detection of the emotional state of the user can be proved valuable for the dialogue manager. For instance, retrieving information about the user's emotion can provide feedback to the dialogue flow manager so as to resolve problematic situations.

Further to exploiting knowledge and experience from past research results [5] [6] [7], PlayMancer is going to develop a full-scale multimodal dialogue system that is aware about the identity of speakers, their emotional state, the specific context, and that is capable of natural spoken speech understanding and generation. Specifically, we will investigate new research directions related to the multimodal dialogue interaction and to develop novel algorithms for fusion of heterogeneous data streams. We will invest efforts in designing a next-generation dialogue flow management algorithms which would enable dynamic compilation of the dialogue flow depending on various indicators. Finally, an advanced user modelling, context awareness and behaviour analysis techniques will be employed to enhance the intelligence of the PlayMancer game platform. The final multimodal dialogue-enabled interaction manager will interface the PlayMancer game engine and their complementary use would enable advanced human-machine interaction and eventually higher level of user satisfaction.

Due the multidisciplinary effort that is required for creation of the PlayMancer multimodal interaction interface and to the new application area, we anticipate that we will push beyond state in the area of multimodal interaction, multimodal dialogue systems, intelligent computing and learning. Specifically, employing novel techniques such as argumentation based reasoning and hierarchical temporal memory [8] we will address game-strategy and human-behaviour analysis problems, which up to now have not been solved.

4 Game modules

While designing the game scenarios, PlayMancer will consume effort into creating the main characteristics of flow experience [9]. This is a state where the following components are identified relative to the player interaction with the game:

- A challenging activity requiring skill,
- a merging of action and awareness,
- clear goals,
- direct, immediate feedback,
- concentration on the task at hand,
- a sense of control,
- a loss of self-consciousness,
- an altered sense of time.

Development of the PlayMancer platform will be driven by serious games scenario from two application domains: therapeutic support programs for lifestyle related disorders, and physical rehabilitation. These domains provide some unique opportunities and challenges to move the networking and platform requirements beyond the current state of the art and provide a basis for developing a generic platform that could support application of serious games in other application domains. Specifically, the focus on physical rehabilitation will drive platform requirements for supporting the development of UA games and the integration of low cost player motion tracking and gesture recognition devices. Games scenarios from the lifestyle related disorder could implicate a platform requirements for multi-modal emotion measurement, domain rather or task-oriented dialogue, measuring emotion states such as boredom, depression, anxiety and associated cognitive responses. While guided online therapeutic support programs are developing at a rapid rate [10] [11] with promising initial results, several major problems remain: The difficulty of communicating and working with important cognitive concepts, and the problem of motivation and compliance, i.e. working with the program consistently. The goal of PlayMancer is to develop serious games to illustrate difficult to explain concepts, to provide sensory and emotional feedback, and to provide more enjoyable exercises that would increase compliance. These games are a replacement. They should be used initially as a tool for health care professionals. Specific game scenario will be defined in PlayMancer, to deal with problems, cognitive traits and risk factors common to a variety of lifestyle related behavioural disorders. These scenarios will be developed within the PlayMancer platform and will be evaluated in pilot trials to determine user acceptance and efficacy in improving cognitive believes and lifestyle improvement. The user requirements and game play scenarios will be developed to insure anonymity, privacy and confidentiality. These requirements will no doubt require that no personal data of any kind will be stored on the platform, in some cases, the games scenario would be only be played by an individual interacting with a game objects, or with a restricted group of authorised users.

These games are intended for use as part of the health promotion and disease prevention process, and not as a replacement. The games modules could be evaluated as supplementary modules (seen as vertical components in the modular framework in figure 2) based on the SALUT platform for delivery of guided self-management programs for behavioural and addictive disorders. This platform was developed by NetUnion within the FP5 research project SALUT and evaluated by hospitals and clinics in Sweden, Germany, Switzerland, Spain (University Hospital of Bellvitge), Holland, and Austria. All evaluations were conducted with approval from the ethical committee of the partner institution.

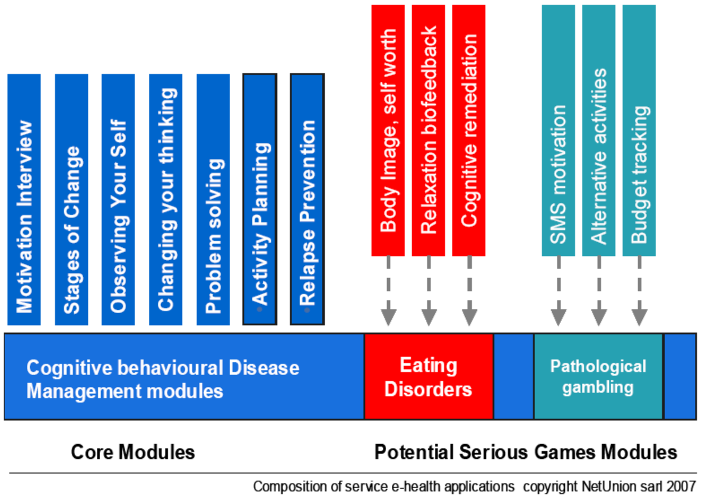


Fig. 2. Game modules for gambling addiction and eating disorders treatment

The games modules could also be integrated, dynamically, based on user need and interest using a dynamic composition of service model for games delivery. Special configurations will be tailored for each user group. Scoring and feedback will also be configured according to needs and requirements for different user groups. All players will receive some standard scoring, while end user with special needs will be also be provided additional information, available only to the user, such as charts showing progress towards a therapeutic objective over time. After the implementation of these game modules in the 3D platform of PlayMancer, experimental field trials will be conducted in order to evaluate the effectiveness of applying these modules to adult players. Evaluation of the games module will be conducted by the University Hospital of Bellvitge, in Barcelona, supported by NetUnion in Lausanne. All evaluation studies will be conducted in compliance with all national and European ethical standards and guidelines. Both partners have extensive experience in working with online therapeutic support programs with populations with special needs within the SALUT project (cited above) and other national and international research projects. Approval from ethical committee, and informed consent for patients will be part of any studied protocol within PlayMancer. The consortium will conform to all applicable laws and regulations regarding experiments with human subjects.

- The subjects will be selected among the adult population who can give consent. Most eating disorder (bulimia and binge eating disorders) and pathological gamblers are adult patients.
- The serious games will only be made available to an adult population having been fully informed about the purpose of the games and the application of certain rules of conduct.

– Small populations or closed communities will be selected to take part into the evaluation field trials. Most of the participants will have to be screened or will require individual interviews.

The project started in November 2007 and at the time of the writing of this paper it has reached the state of user requirement identification and a first definition of functional requirements. We expect to have a first fast prototype by January 2009 and start trials in May 2009.

References

[1] Grammenos, D., Savidis, A., Stephanidis C.: UA-Chess: A Universally Accessible Board Game. Proceedings of the 3rd International Conference on Universal Access in Human-Computer Interaction. G. Salvendy (ed.). Las Vegas, USA (2005). Lawrence Erlbaum

[2] OGRE (Object-Oriented Graphics Rendering Engine), http://www.ogre3d.org

[3] CrystalSpace realtime 3D graphics SDK, http://www.crystalspace3d.org

[4] Pintaric T., Kaufmann H.: Affordable Infrared-Optical Pose-Tracking for Virtual and Augmented Reality. Proceedings of Trends and Issues in Tracking for Virtual Environments workshop, IEEE VR 2007, Charlotte, USA (2007)

[5] Lee, C. M., Narayanan, S.S.: Towards detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing, (2005) Vol 13, No. 2, 293–303

[6] Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A.: Prosody based automatic detection of annoyance and frustration in human computer dialog. Proceedings of ICSLP, Denver (2002) 2037–2040

[7] Liscombe, J., Riccardi, G., Hakkani-Tür, D.: Using context to improve emotion detection in spoken dialog systems. Proceedings of Interspeech (2005), 1845–1848

[8] Hawkins, J. and George, D.: Hierarchical Temporal Memory: Concepts, Theory, and Terminology. Numenta Inc, (2006)

[9] Csikszentmihalayi, M.: Flow: The Psychology of Optimal Experience. Harper Perennial, London (1990)

[10] Fernández-Aranda, F., Martínez, C., Núñez,A., Jiménez-Murcia, S.: Nuevas tecnologías en el tratamiento de los trastornos de la alimentación. En J. Vallejo (Ed.). Update Psiquiatría. Ed. Masson (2005) 105–118

[11] Carrard, I., Rouget, P., Fernandez-Aranda, F., Volkart, A.C., Damoiseau, M., Lam, T.: Evaluation and deployment of evidence based patient self-management support program for bulimia nervosa. International Journal of Medical Informatics (2006) 75, 101–109

User Experience and Emotion-Aware Business Network Service Selection[1]

Jean-Marc Seigneur and Xavier Titi

On the 10th of September 2008, the European Commission launched its Future Internet Research and Experimentation (FIRE) initiative [1]. We envision the Future Internet as being able to infer the user experience quality of the network services it provides and take into account these user emotions at time of selection of these network services. As a first step towards this vision, we are investigating appropriate mechanisms for mobile network selection based on Quality of Experience (QoE) as part of the consortium who won funding from the EU for the 3 million Euros plus PERIMETER project [2]. We stress that it is important to make the difference between QoE and Quality of Service (QoS). The ITU-T in its E800 recommendation [3] defines QoS as follows: "the collective effect of service performances, which determines the degree of satisfaction of service users". However, until now, QoS is mostly based on technical network results rather than what the users really perceive of the service, for example, in terms of usability, accessibility, retainability, reliability, efficiency... QoE corresponds to that forgotten side of the results. In the same line of thought, the idea of having an Always Best Connected (ABC) connection seems more vision than reality. This may because the current state-of-the-art solutions, such as IETF Mobile IPv6 (MIP) or the emerging Host Identity Protocol (HIP), mainly focus on mobility management, instead of again considering additional user related issues such as user preferences, associated cost, access-network, operator reputation, and trust and application related issues like QoS and failure recovery in conjunction with mobility. Another explanation could be the different meaning associated by telecom operators and users with the word *best*. Unless telecom operators can directly benefit from allowing a user to switch to another operator, operators have an incentive to bind the user to their networks or service provisioning. In contrast, for end-users ABC means saving money by switching to the lowest cost operator.

PERIMETER's main objective is to establish a new paradigm of user-centricity for advanced networking. In contrast to network-centric approaches, user-centric strategies could achieve true seamless mobility. Putting the user at the centre rather than the telecom operator enables the user to control his or her identity, preferences and credentials, and so seamless mobility is streamlined, enabling mobile users to be ABC in the multiple-access multiple-operator networks of the Future Internet. In addition to mechanisms for QoE selection, as depicted in Figure 1, PERIMETER will provide innovative implementation of protocols for fast authentication, authorisation and accounting based on privacy-preserving digital

[1] This work is sponsored by the European Union, under the FP7-ICT-2007-2-224024 PERIMETER project.

identity models. Supplementary QoE statistics will drive session level content adaptation processes, thus requiring session level QoE signalling mechanisms between mobile terminals and application/service providers, to support continuously changing network conditions and user preferences. All these PERIMETER mechanisms will be designed to be independent from the underlying networking technology and service provider, so that fast, inter-technology handovers will be possible.

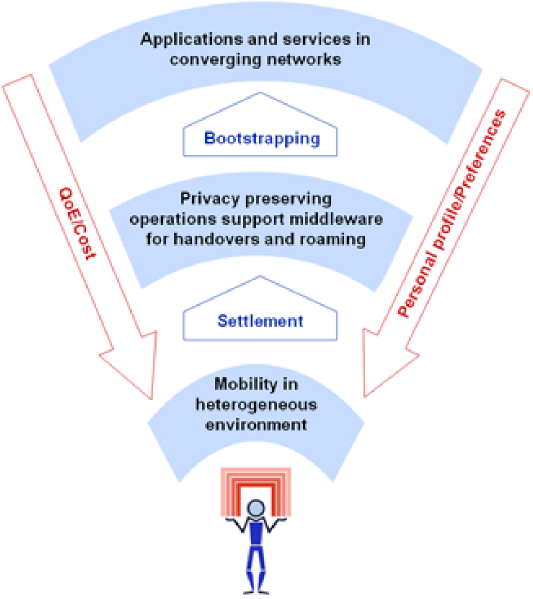

Applications and services in converging networks
Bootstrapping
Privacy preserving operations support middleware for handovers and roaming
Settlement
Mobility in heterogeneous environment
QoE/Cost
Personal profile/Preferences

Fig. 1. High-level View of PERIMETER

The users will propagate their QoE results in a decentralised overlay of information that will not be controlled by the telecom operators. In this way, the competition between the telecom operators will be more transparent to the users and we expect that it will end up in a very competitive Future Internet in Europe. To avoid fraud and security issues, distributed and self-organizing methods of QoE aggregation will require incorporation of trust and reputation algorithms and mechanisms in mobile terminals. Based on our experience in designing attack-resistant decentralised computational trust and identity management [4], we will research and deliver the required computational trust building blocks for PERIMETER decentralised QoE aggregation.

References

1. EU FIRE, http://cordis.europa.eu/fp7/ict/fire/
2. EU-funded PERIMETER project, http://www.ict-perimeter.eu/
3. ITU-T E800 draft, http://www.itu.int/md/T05-SG02-080506-TD-WP2-0118/en
4. Seigneur, J.-M., “Trust, Security and Privacy in Global Computing”, PhD Thesis, Trinity College Dublin, 2005.

Activity Report Summary of the ASGroup

During the period September 2007-September 2008 the AS Group continued its research in three major areas of mobile systems: Quality of Service, Trust and Identity Management, and novel paradigm for mobile information management. The group as a living and active entity, show a slight mutation in personnel, acquired new projects, participated in conference organization and published papers in journals and conferences. In this report summary we present the main points of the group status and research activities.

From the point of view of personnel in September 2008 the group was composed of 8 persons, headed by Prof. Dimitri Konstantas, (who is since July 2007 Vice-Dean of the Social and Economic Science Faculty sharing his time between research and administration). In April 2008, M. Michel Deriaz, completed his PhD and left group becoming the R&D director (Europe) at Organix IT. On the other hand, two new PhD students joined the group, M. Dejan Munjin and M. Xavier Titi.

During the last 12 months the group gained a series of projects from the European Commission (7th FP) and the Swiss Federal Government. In addition our group participates in the COST action IC0703 TMA – *Traffic Monitoring and Analysis: theory, techniques, tools and applications for the future networks*.

Project name	***Support from***	***Start***	***End***	***Total Funding***
Geo-Tags	OFFT CTI	1-5-2007	30-4-2009	72.000 CHF
PlayMancer	EU 7th FP	1-11-2007	30-10-2010	220.000 EUR
PERIMETER	EU 7th FP	1-5-2008	30-4-2011	270.000 EUR
QoS prediction Service	SER	1-10-2008	30-9-2001	170.000 CHF
TMA	COST	1-2-2008	31-1-2011	- 0 -

From the point of view of recognition of research results, the group published one PhD thesis and 20 papers in international journals, conferences and book chapters (in addition to this technical repot which is composed of early or revised versions of some the published papers, as well as unpublished papers - A full list of the published work of the groups is given at the end of the activity report). In addition another 8 publications are in the pipeline to appear in 2009.

Our research, nevertheless, is not performed on an isolated island. We have established active collaboration with a number of research institutes in Europe, exchanging students and researchers, and working together towards common targets. This is reflected in the publications of the group, many of which are written in collaboration with researchers in other institutes. The most important institutes we collaborate with are the University of Twente in the Netherlands and the School of Computer Science and Information Systems, Birkbeck College of London University in the United Kingdom.

List of Publications (September 2007 – September 2008)

Journals

1. Val Jones, Aart van Halteren, Dimitri Konstantas, Ing Widya, Richard Bults, *An application of augmented MDA for the extended healthcare enterprise*, International Journal of Business Process Integration and Management, Vol 2, No 3, October 2007
2. E. Grandgirard, C. Gertosio and J.-M. Seigneur, *Trust Engines to Preserve the Quality of an Operational Decision System*, in International Journal of Factory Automation, Robotics and Soft Computing/, ISSN 1828 – 6984, 2007.
3. J. Abendroth and J.-M. Seigneur, *Leveraging the Trusted Platform Module for More Trustworthy P2P File Sharing Peer Software*, in Transactions on Communications Journal, WSEAS, 2007.
4. Giovanna Di Marzo Serugendo,Alfredo A. Villalba Castro and Dimitri Konstantas, *Hovering Information* in Works in Progress – Activity Based Computing, IEEE Pervasive Computing Journal, April-June 2008.

PhD Thesis

5. Michel Deriaz, *GeoVTag, Trusting Vtrtual Tags*, PhD Thesis, University of Geneva, Faculty of Social and Economic Sceinces, Thesis No 665, April 2008.

Conferences

6. Katarzyna Wac, Mortaza Bargh, Pravin Pawar, Bert-Jan van Beijnum, Arjan Peddemors, Richard Bults, *Power- and Delay-Aware Mobile Application-Data Flow Adaptation: the MobiHealth System Case Study*, 10th IEEE International Conference on e-Health Networking, Applications & Services (HealthCom 2008), July 2008, Singapore, publisher: IEEE Press
7. Pravin Pawar, Katarzyna Wac, Bert-Jan van Beijnum, Pierre Maret, Aart van Halteren, Hermie Hermens*, Context-Aware Middleware Architecture for Vertical Handover Support to Multi-homed Nomadic Mobile Services*, proceedings of the 23rd Annual ACM Symposium on Applied Computing (ACMSAC08), Ceará, Brazil, publisher: ACM Press
8. Alfredo A. Villalba Castro, Giovanna Di Marzo Serugendo and Dimitri Konstantas, *Hovering Information - Self-Organising Information Using Location-Based Caching Policies*, 2nd ERCIM Workshop on eMobility (in conjunction with WWIC 2008), May 30, 2008 - Tampere, Finland.
9. A. Villalba Castro, G. Di Marzo Serugendo, D. Konstantas, *Hovering Information - Self-Organising Information that Finds its Own Storage*, IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC'08), Taichung, Taiwan, June 2008

10. L. Ragia, M. Deriaz and J.-M. Seigneur, *Mobile Location Based Services for Trusted Information in Disaster Management*, Proceedings of the 17th International Conference on Information Systems Development IEEE, 2008.
11. E. Grandgirard, C. Gertosio and J.-M. Seigneur, *Trust Engines to Optimize Semi-Automated Industrial Production Planning*, Proceedings of the fourth International Symposium on Industrial Electronics, IEEE, 2007.
12. T. El Maliki and J.-M. Seigneur, *A Survey of User-centric Identity Management Technologies*, Proceedings of the SECURWARE International Conference, IARIA, 2007.
13. T. El Maliki, Y. Durukan and J.-M. Seigneur *EasIDeploy: A Usable and Convenient Identity Management Platform with Strong Authentication, PKI and Biometry,* Proceedings of the 2nd Conference on Advances in Computer Security and Forensics, 2007.
14. O. Powell, J.-M. Seigneur and L. Moraru, Trustworthily Forwarding Sensor Networks Information to the Internet, Proceedings of the SECURWARE International Conference, IARIA, 2007.
15. Michel Deriaz, The Uncertainty of Truth, proceedings of the Sixth Annual Conference on Privacy, Security and Trust, PST2008, October 1-3, 2008, Delta Fredericton, Fredericton, New Brunswick, Canada
16. Elias Kalapanidas, Hikari Watanabe, Costas Davarakis, Hannes Kaufmann, Fernando Fernandez Aranda, Tony Lam, Todor Ganchev and Dimitri Konstantas, *PlayMancer: A European Serious Gaming 3D Environment*, Proceedings of the 2nd International Workshop on e-health Services and Terchnologies - EHST 2008 in Conjustion with ICSOFT 2008, Porto-Portugal, July 2008, pp 51-59 ISBN 978-989-8111-56-2

Book Chapters

17. Jean-Marc Seigneur, *Social Trust of Virtual Identities*, book chapter in *Computing with Social Trust and Reputation*", ISBN 978-1-84800-355-2, Springer, 2008.
18. Jean-Marc Seigneur, *AmbiTrust? Immutable and Context-Aware Trust Fusion* book chapter in "*Trust Management in Virtual Environment*", ISBN 81-314-1254-1, Icfai University Press, 2008.
19. J.-M. Seigneur, L. Moraru and O. Powell, *Survivability of Sensors with Key and Trust Management*, book chapter in "Handbook of Research on Wireless Security", ISBN 978-1-59904-899-4, IGI Global, 2008.
20. J.-M. Seigneur and C. D. Jensen *User-Centric Identity, Trust and Privacy*,, book chapter in *Trust in E-services: Technologies, Practices and Challenges*, ISBN 978-1599042077, Idea Group Publishing, 2007.
21. Dimitri Konstantas, An Overview of Wearable and Implantable Medical Sensors in the IMIA Yearbook of Medical Informatics 2007, Geissbuhler A, Haux R, Kulikowski C, editors, IMIA and Schattauer GmbH

To appear in 2009

Books

1. "Collaborative Security Technologies for Data Assurance Management: Building Systematic Trust", J.-M. Seigneur and A. Slagell, co-editor of the book, IGI Global, (to appear in 2009)

Book Chapters

2. Katarzyna Wac, Richard Bults, Bert-Jan van Beijnum, Hong Chen, Dimitri Konstantas, *Toward Mobile Web 2.0-based business methods: Collaborative QoS-information sharing for mobile service users*, book chapter in *Mobile and Ubiquitous Commerce: Advanced E-Business Methods*, M. Head, (Eds.), IGI Global publisher, vol. 4 (to appear in 2009)
3. Jean Marc Seigneur, "*Reputation Management Services*", book chapter in *Computer And Information Security Handbook*, Elsevier (To appear in 2009)
4. Tewfik El Maliki and Jean-Marc Seigneur, *Identity Management Services* , book chapter of "*Computer And Information Security Handbook*", Elsevier, (to appear in 2009).
5. J.-M. Seigneur, Engineering Reputation Services with Collaborative Information Systems., book chapter in *Collaborative Security Technologies for Data Assurance Management: Building Systematic Trust*", IGI Global (to appear in 2009).
6. J.-M. Seigneur and P. Dondio, *Trust and Reputation for Successful Software Self-organisation*, book chapter in *Self-Organising Software*, Springer, (to appear in 2009).
7. J.-M. Seigneur, G. Lenzini and B. Hulsebosch "Adaptive Trust Management", book chapter in "Self-Organising Software", Springer (to appear in 2009)

Conferences

8. Pravin Pawar, Bert-Jan van Beijnum, Katarzyna Wac, Hermie Hermens, Dimitri Konstantas *Towards Location Based QoS-Aware Network Selection Mechanism for the Nomadic Mobile Services*, to be presented in 6th Annual IEEE Consumer Communications & Networking Conference - IEEE CCNC 2009, 10 - 13 January 2009 in Las Vegas, Nevada