Aspects of personal information valuation in web browsing and mobile communication

Juan Pablo Carrascal Ruiz

TESI DOCTORAL UPF / 2014

Directors de la tesi

Prof. Dr. Josep Blat,
Department of Information and Communication Technologies
Dr. Rodrigo de Oliveira,
Google, Inc.

Universitat
Pompeu Fabra
Barcelona

A Belen, Edwin e Ina.

Acknowledgements

Agradezco a mis supervisores Josep Blat y Rodrigo de Oliveira por mostrarme el camino.

A Karen Church, Mauro Cherubini, Beverly Harrison y Ernesto Arroyo por su apoyo en diversas etapas de este mismo camino.

A a mis padres Belén y Edwin por apoyarme y estar siempre ahí.

A Ina por ser tan maravillosa y por darme calma y besos.

A mi familia por estar siempre pendientes.

A mis amigos de Barcelona por los buenos momentos: Andrea, Gullit y Joaquim; Jose y Magda; Tálaga y Julie; Jose Zapata; Mónica e Isaac; Natalia.

A los chicos de la uni y demás proyectos, utópicos y de los otros: Nina, Guiller y Penélope (vamos, castores, vamos); Alberto González; Sergio Bromberg, Yuji, Olivier, Davinia, Valeria, Susan, Ping, Fabien.

A los compañeros de Telefónica, por las buenas charlas y enseñanzas: Vijay, Jose San Pedro, Martin, Matylda, Marc, Pol, Antonio, Nuria, Andrea, Montse y Coral.

To the guys at Yahoo! also for the interesting chats, fun and support: Cheng I, Cheng II, Marius, Peter, Maia, Jingyu, Vera, Beverly, Jofish, Kent, Henriette, Frank, Christian, Senaka.

A Lydia, Vanessa, Jana y las chicas de la secretaría por su paciencia.

Y a Spacebarman por la cordura.

Abstract

The goal of this dissertation is to provide insights into how web and mobile phone call users assign value to their personal information. Two user studies are presented. The first one used a refined Experience Sampling Method to obtain the the monetary valuation of Personally Identifiable Information (PII) of web users. We observed how the value of different types of PII compare, and reflected upon how these values relate to users' concerns about privacy and monetization of their PII. The second study focuses on annotation of mobile phone calls, a process that involves selecting the information that the user considers to be the most important from the call. We found the factors that mainly influence the need of the caller to take notes. We also observed how the annotation needs and behaviors change over time. Finally, we evaluated different annotation techniques, including annotations made by context-independent observers. Our findings provide important insights for the development of automatic annotation tools, and also suggest the potential of crowdsourcing for finding noteworthy information from mobile phone calls.

Resumen

El objetivo de esta tesis es entender cómo los usuarios de las web y de telefonía móvil asignan valor a su información personal. La tesis describe dos estudios. En el primero se utiliza un Método de Muestreo de Experiencias refinado para obtener la valoración económica de la Información Personal Identificable (PII) de usuarios web. Observamos los valores de diferentes tipos de PII de manera comparativa, y reflexionamos sobre cómo estos valores se relacionan las preocupaciones de los usuarios sobre la privacidad y la monetización de su PII. El segundo estudio se enfoca en la anotación de llamadas telefónicas móviles, un proceso que involucra la selección de información que el usuario considera ser la más importante dentro de una llamada. Encontramos los factores que principalmente influencian la necesidad del llamante de tomar notas. También observamos cómo los comportamientos y necesidades de anotación cambian a través del tiempo. Finalmente, evaluamos diferentes técnicas de anotación, incluyendo anotaciones hechas por observadores independientes del contexto. Nuestros hallazgos incluyen ideas importantes para el desarrollo de herramientas automáticas de anotación, y además sugieren que existe potencial en el uso de *crowdsourcing* para encontrar información notable en llamadas telefónicas móviles.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *"If you want to keep a secret, you must also hide it from yourself."*
> *—George Orwell, 1984*

The increasing connectivity in almost every context of our lives has made possible to access and collect huge amounts of data from individuals. This ever increasing availability of information is leading to the appearance of the so-called Big Data [45]. ICT-based companies that have the adequate technical resources, such as ISPs, telecommunication companies and web-based service providers, have found a source of monetary income without precedent in these huge repositories of aggregated data. But for users of the technology, the perspective is different. As individuals, their personal data collections might be smaller, but still highly valuable.

Additionally, mobile phones have become one of the most pervasive technologies today[1]. The applications of smartphones have exploded in recent years [33], and the number of mobile connections has outnumbered the total world population[2]. Such abundance of devices means that humans are increasingly connected, using Internet services with more frequency, and making a large amount of phone calls.

Technology users—both desktop and mobile—are constantly receiving information. Email, phone calls, blogs, social media, shared files, represent a constant flow of data that is hard to parse and organize. These users also generate information, as a product of their interaction with technology; location, inertial sensor data from sport and health-related activities, and metadata are just a few examples. And while a part of all this information can be ephemeral, maybe being just a support for social needs, another part is important for our lives, and worth keeping.

An inherent feature of digital data—in particular personal data—, is that as important as it can be, it is very difficult to control. We want to keep data that is private for ourselves only, but frequently we need to share it, as it is a fundamental element of transactions with the very same technological systems that allow us to collect it. What is the point of taking photos if we cannot share them with our friends and family? What is the point in having a credit card nowadays if we cannot use it to buy what we want online? Some important problems then arise. From all those data, what is important and what is not? How can we assign value to different pieces of our data? How can we take the right decisions in information-based transactions based on that valuation? How can we discriminate and organize our data so we can use it in a reasonable way?

[1] According to the International Data Corporation (IDC), in the first quarter of 2014, a total of 281.5 million smartphones were shipped worldwide. (press release: http://www.idc.com/getdoc.jsp?containerId=prUS24823414).

[2] Compare http://www.census.gov/popclock/ and https://gsmaintelligence.com/

These problems go beyond—but are also related to—personal information management, stepping into the grounds of psychology, economics, or more precisely, behavioural economics applied to personal data. While managing our information is a hard task, understanding its value in all applicable contexts is not easy either. And that issue influences the basic concerns of Personal Information Management [34].

This thesis aims to shed light on important aspects related to the rationale behind valuation of personal information, as we discuss them next.

1.1 Main Research Questions

This dissertation presents answers to four main research questions, which are framed in the context summarized above. These questions are:

- *Question 1: What monetary value do Internet users assign to their personal information while being online? What are the perceptions of the users with respect to the monetization of their personal information by Internet-based companies?* When information is personal, it has a clear—personal—value for our lives. But the Internet-based companies have discovered that personal information can be also valuable as it can be economically exploited. This implies a value conversion. While this conversion is clear for companies as an essential part of their business model, it is not so clear from the users' perspective. With this research question we want to shed some light on the web users' valuation of their private information, as well as their opinion on the monetization of their information.

- *Question 2: What factors influence the mobile phone users' need of taking notes?* The note taking process involves a process of information valuation, though in this case there is no conversion; the value of information remains in the personal domain. But even without conversion, personal information is valuable in our daily lives for such diverse uses as communicating, keeping track of activities, setting up personal meetings and appointments, entertainment etc. And still, we go through a personal valuation process when we need to determine what pieces of information should be preserved for later use and what is not that important. Areas such as personal information management and personal informatics have deal with such personal economics problems. In this thesis, by using phone call annotation as a use case, we advance the understanding of personal information valuation in the relevant scenario of mobile communication.

- *Question 3: Is valuation of personal information consistent or variable across time? How this variation manifests in the study case of mobile phone call annotation?* Continuing with the case of annotation on mobile phone calls, we aimed at understanding if the information deemed to be important—annotated information—remained constant across time, and if not, what changes appear and how they unfold. Understanding the dynamics aspect of the note taking activity helps to understand how personal information value can change over time. Additionally, our study on this topic helps to inform the eventual design of a system that could perform automatic annotation of mobile conversations.

- *Question 4: Is personal information—in the case of mobile call transcripts—value "unanimous"? How does valuation change from the perspective of a neutral or context-free observer?* We might consider that the value that we assign to our personal information is clearly defined. However, while answering the previous research question, we found that in the case of mobile phone call annotation, time influences such valuation. While conducting a comparative evaluation of annotation methods, we introduced the context-free annotator: a person that take notes while having no relationship with the original information. This figure helped us to see that value of information—in the context of note taking—is highly subjective.

It would be to ambitious to pretend to cover all aspects related to valuation of personal information with these four questions. However, answering them is an important contribution towards understanding the topic. The contributions will be discussed through the this dissertation, and specially in the conclusions.

But first we will discuss how these questions are framed in the main related research areas, as well as some aspects of the methods that have inspired the research herein presented.

1.2 Research areas and some methodological aspects

The work presented in this dissertation can be articulated mainly with three research areas: Human-Computer Interaction, Behavioural Economics and Personal Information Management (PIM). It is also framed into two relevant interaction scenarios, web browsing, and mobile communication. Next we discuss how the topics of this thesis relate to the aforementioned research areas, and our methodological approach.

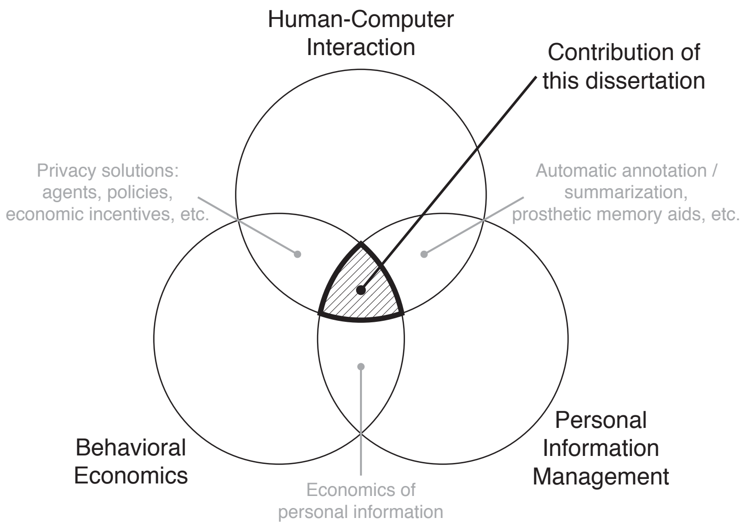


Figure 1.1: Main research areas related to this dissertation.

Economics of online personal information

Some of the most important business models of the digital economy rely on the exploitation of Personally Identifiable Information (PII) [41]. When users browse the web they inevitably share their PII. It can be done intentionally, e.g.by sharing information in social networks [32], or unintentionally, in the form of traces (e.g. cookies, logs, etc.) left when visiting different websites [36, 40]. The PII is then exploited in a number of ways, such as in targeted advertising.

Unfortunately, this process tends to be obscured by the Internet companies. This is also helped by the users' lack of information, bounded rationality, and systematic deviations from rationality [3] that prevent them to perform better privacy-related decisions. Privacy concerns are easily manipulated by external factors. It has been observed that the wording employed when asking users about privacy influences the resulting answers [9]. These findings mean that users' cognitive biases make it difficult for them to adopt the right strategies and solutions to protect their privacy. Indeed, privacy policies are rarely read [26]; privacy solutions are not adopted [2] while simple interface design elements lower users' privacy concerns [27]; actual behaviour easily deviates from reported concerns [4]; and actual users' knowledge diverges from reality [27]. In other words, when looking at privacy protection as an economics problem, Internet users' behaviour is far from that of a rational agent.

As Acquisti explains [1], users are in a position of asymmetry with respect to PII-exploiting companies: users are losing control of their information and there is no way to know what companies do with it. This might be changing in more recent times, as an increase of media coverage [16, 48, 46, 51, 52, 50, 56, 38, 55] of Internet privacy threats is increasing the awareness of Internet users to this issue. But still it is very difficult for users to perform an adequate cost-benefit analysis for transactions involving their PII during Internet browsing.

While studies [47] have showed the potential for "transactional privacy" for Internet users, their monetary valuation of personal information is not clear. Understanding this valuation can help Internet users to take more informed decisions when sharing information online, as well as to inform the creation of incentives for the adoption of privacy solutions. Chapter 2 of this dissertation sheds some light on this important issue.

Refined Experience Sampling Method for PII valuation

As already indicated, asking users about privacy-related topics is problematic due to a number of psychological biases. In order to overcome such issues, we implemented a refined Experience Sampling Method [14] (rESM) for extracting the valuation of PII while being online. This method, originally employed in psychology [35, 17], consists in probing the subjects, usually at random intervals, during their daily routine. This minimizes retrospective recall while allowing to obtain responses with a high ecological validity. It has been increasingly used in HCI research increasingly in recent years [15, 20, 21, 19]. Our rESM implementation used a Firefox plugin that presented pop-up questions to participants during their normal web browsing activity. Additionally a reverse second-price auction was implemented, with a two-fold goal in mind: it helped keeping the interest in participation—since participants could actually earn money by participating—, while at the same time it promoted the honesty of the information valuation—an inherent property of the chosen auction type.

We think that applying this methodology to privacy-related studies opens some doors and provide ideas for further research. In fact, it has actually already inspired new studies in the study of economics of private information [49].

Note taking and mobile phones

Taking notes helps to preserve selected pieces of information for later access. It is done in a wide amount of contexts, such as work meetings, in the classroom, while reading a text or during phone calls. There are multiple strategies for annotation, but they can be summarized in two categories: *derived*, or the composition of new material by re-synthesizing fragments of the original document as a new piece [44]; or *explicit*, the selection of the passages that are considered to be the most important by means of adding some type of marker to the document [23]. As Geyer [23] mentions, it is also possible to annotate *online*, synchronously with the activity—e.g. taking notes while a conversation is happening—, or *offline*, after the activity has happened—e.g. adding markers or highlights to a transcript. In all cases, it implies assigning value to the different pieces of information that constitute the original document—or meeting, or conversation—and preserve those which are of higher importance.

Annotation has been studied in general contexts [37, 31, 5], as a relevant topic for Personal Information Management (PIM). It has been studied from a number of perspectives, commonly taking into account the psychological factors that influence the note taking process. Piolat et al. [44] points to the cognitive load that it involves. A study by Bernstein et al. [5] mentions a dichotomy between the stated perception of organization of several participants' notes—fundamental for the retrieval of information—and their actual organization. This directly influences the process of accessing the information *a posteriori*, something that was widely discussed by Lansdale [34] when studying wider implications for the design of PIM tools.

Note taking acts as a memory aid for the creator of the notes [23] or as a form of prosthetic memory [29]. Given the influence of time on human memory, it has also been taken into account when studying the note-taking process. For instance, Kalnikaite and Whittaker [29] observed that the recall value and retrieval efficiency of notes is high only in the short term, and decrease with time. Lin et al. [37] studied the lifecycle of "micro notes". Whittaker et al. [54] studied users' *a posteriori* access frequency to their meeting records.

On the other hand, voice calls are still not widely explored as a source of non-ephemeral data, in consequence most information exchanged during them ends up being lost. Most of the time phone users have tackled this problem with simple tools such as taking notes with a pen and a paper, though this is not always possible when the user's hands are busy [42] or when such tools are simply not available.

One could consider a phone call to be similar to a two-person meeting. In that sense, related research can be found in relation to annotation of work meetings, a topic that has been widely explored by scholars [53, 28, 7, 8, 23, 25]. There are indeed similarities between both contexts, as both deal with voice data and the real time interaction between people. However there are fundamental differences between phone calls and meetings. The former are less structured and less planned than the latter, and mobile phone calls are frequently made while being on mobile contexts. Therefore, most of the findings related to work meetings can not be applied to mobile phone conversations, and more research should be done in order to inform the future design of automatic phone conversation annotation tools.

A longitudinal study on mobile phone call annotation

In recent years, there has been an increasing interest in conducting longitudinal research in HCI [30]. In this type research, data is collected at at least two different points across time [22], as opposed to cross-sectional research, where only one point in time is considered [39]. The main reason that justifies longitudinal studies is that they allow to observe how an observed phenomenon or behaviour unfolds across time. As Karapanos et al. mention [30], this is important in HCI studies as the “shift to a more experience-oriented design led to a strong emphasis on the temporal dynamics of interactive product use”.

Chapter 3 of this thesis presents an in-depth, longitudinal study on several aspects of mobile phone call annotation. It starts with presenting static aspects, such as annotation needs of mobile users, frequently used annotation tools, and factors that influence note-taking; to how annotation needs and behaviours change across time. Later, it also includes an study on the comparative evaluation of annotation methods, that helps complementing the topic and suggests future research and development options for the deign of phone call annotation tools.

1.3 Contributions

Next we summarize the contributions of this thesis to the areas mentioned before. While most of the contributions are related to personal information valuation in the two scenarios described before, some contributions are more specific, so here they are presented framed in their specific contexts. First, in the context of web browsing:

- We found the monetary valuation that Internet users assign to different types of their Personally Identifiable Information (PII) while browsing the web. Notably, we observed that information related to web-browsing history was valued at around 7€. More importantly, we found how the value of different types of PII—including offline and offline—compare. Offline PII information—such as age, salary and address—was valued higher than online PII. PII associated to social networking or finance websites was valued significantly higher than PII related to search and shopping websites. As opposed to this, we did not observe a significant difference in the valuation of different amounts of PII pieces (e.g., 1 email contact or search keyword were considered to have a similar values as 10 items of the same type).

- We also found that, unsurprisingly, web users do not like the idea of their PII being monetized by Internet-based companies. However, they are willing to accept certain goods in exchange of their PII. These goods include, money, improvement of current services, additional free services and recommendations, amongst others.

- We used a refined Experience Sampling Method for extracting PII valuation. While this methodology has been used in other contexts [15, 24], its use for the extraction of PII valuation in a web-browsing context is novel. Before starting to collect the valuation of PII from participants, we measured each participant's baseline browsing behaviour. When comparing it with their browsing behaviour before and during the study we found little difference, meaning that the probing did not affect their normal web browsing activity. Furthermore, extreme valuations (highest and lowest) were justified by participants as a result of the importance of the PII or because of fairness of the provided value. These observations provide evidence of the effectiveness of the chosen methodology in the web browsing context, and suggests new opportunities for similar studies related to information valuation.

Second, in the context of mobile communications:

- The findings presented in this thesis empirically confirm that there is an actual need for the development of solutions that can support note-taking from mobile phone calls in an automated way. We found that users frequently have difficulties when trying to remember information from mobile phone calls. Also they often have the need to take notes to remember information discussed in phone conversations. We also observed that the smartphone was the most frequently used tool for taking notes after a call, followed by pen and paper and sound recording devices. Additionally, we found a number of factors—both call-related and contextual—that influence note-taking during mobile phone calls. The main influencing factor was the presence of certain patterns such as proper names, numbers and question adverbs, as well as the length of the call. We also observed some influence of the objective of the call. The complete discussion provided by our research suggests a path for the implementation of automatic detection of noteworthy information from mobile phone calls [6].
- We provided a comprehensive description of the temporal dynamics of phone call annotation needs. We found that callers seem to underestimate the value of information related to the objective of the call close to the time when the call was made, while finding new important information as time passes. This newly found information is regarded as more important, and is considered to have value for archiving purposes and for refining the information initially considered to be important. Intelligent annotation systems should preserve as much information as possible, and be able to leverage our findings to automatically detect the relevant pieces of information at the right time, even if the user has not acknowledged their initial importance. User's feedback should also be taken into account after notes are detected and presented to the user. Given the importance of the interaction between an automatic annotation system and the user's feedback, the whole process should be done semi-automatically.

In a wider context, our contributions provide important insights for understanding the perceptions of the users as key players in one of the most important business models that support the Internet economy, one that is based on the economic exploitation of personal information. The role of the web users in the information economy is clear for Internet-based companies. Our research suggests the need for these companies to be more transparent in explaining their need to monetise their users' PII. Additionally, there is a need for incentives for users to adopt privacy solutions, as previous solutions seem to have largely fail in the past.

On the other hand, while note-taking is a topic that has been studied in contexts such as work meetings[28, 7, 8, 23] or personal information management [31, 37], it has not been explored in the context of daily mobile phone calls. The findings herein presented help to inform the future creation of solutions for the automatic extraction of annotations in this context. We also believe that beyond the specific application of phone call annotation, our findings help to better understand important aspects of the rationale behind the information management behaviour of users of mobile communication.

1.4 Organization of this dissertation

Three papers contribute to the body of this dissertation. One of them is a journal paper that is accepted for publication in the Transactions of Computer-Human Interaction (ToCHI) journal; a conference paper published in 2013 in the Proceedings of the 22Nd International Conference on World Wide Web; and a third paper that was just submitted to the ACM SigCHI conference.

The first part of this dissertation focuses in the monetary valuation of personal information of web users. Chapter 2, presents this paper:

J. P. Carrascal, C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira. "Your Browsing Behavior for a Big Mac: Economics of Personal Information Online". *Proceedings of the 22Nd International Conference on World Wide Web*. WWW '13. Rio de Janeiro, Brazil, 2013, pp. 189–200

It describes a user study based on a refined Experience Sampling Method on the monetary valuation of Personally Identifiable Information (PII) when browsing the web. It addresses the problem of understanding monetary valuation of personal information during web use, as stated in Research Question 1.

The second part of this thesis includes two papers the were produced after a multi-phase longitudinal study on different aspects of mobile phone call annotation:

J. P. Carrascal, R. de Oliveira, and M. Cherubini. "To Call or to Recall? That's the Research Question". *Transactions of Computer-Human Interaction (ToCHI) (Accepted for publication)* (2015)

J. P. Carrascal, F. Bonin, S. P. Jose, R. De Oliveira, and N. Olivier. "Who can we trust? A Comparative Evaluation of Phone Call Annotation Techniques". *ACM SigCHI Conference (Submitted)*. 2015

The first paper, included in Chapter 3, focuses on both static and dynamic aspects of phone call annotation with the aim of answering Research Questions 2 and 3. The second one deals with Research Question 4 and is included in Chapter 4. It presents the last stage of the study, in which participants performed a comparative evaluation of three phone call annotation techniques: the notes made by the original caller, notes made by a third party, i.e. a context-free annotator, and notes obtained by an automatic algorithm developed. This algorithm was developed by leveraging the findings of the first two phases of the study, as presented in Chapter 3, although the design of the algorithm itself is not part of this dissertation. Besides these two papers, the study also produced another paper [11] and a patent [18] in which there was a significant involvement of the author of this dissertation. For the sake of completion, both documents are included in Annex 1.

Finally, Annex 2 describes some additional research activity that the author conducted while working on this dissertation. This annex also includes one paper [43] that presents a critical view on the ethical aspects of current persuasive technologies. It also proposes three methods for implementing behaviour design technologies with higher compliance to ethical guidelines.

References

[1] A. Acquisti. "Privacy and Security of Personal Information: Economic Incentives and Technological Solutions". *The Economics of Information Security, J. Camp and R. Lewis (eds)* (2004), pp. 165–178.

[2] A. Acquisti and J. Grossklags. "Losses, Gains, and Hyperbolic Discounting: An Experimental Approach to Information Security Attitudes and Behavior". *The Economics of Information Security* (2004), pp. 179–186.

[3] A. Acquisti and J. Grossklags. "Privacy and Rationality in Individual Decision Making". *Security & Privacy, IEEE* 3 (2005).

[4] B. Berendt, O. Gunther, and S. Spiekermann. "PRIVACY IN E-COMMERCE: Stated Preferences vs. Actual Behavior". *Communications of the ACM* (2005).

[5] M. Bernstein, M. V. Kleek, D. Karger, and M. C. Schraefel. "Information scraps: How and why information eludes our personal information management tools". *ACM Trans Inf Syst* 26.4 (2008), pp. 1–46.

[6] F. Bonin, J. San Pedro, and N. Oliver. "A Context-Aware NLP Approach For Noteworthiness Detection in Cellphone Conversations". *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.* Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 25–36.

[7] A. Bothin and P. Clough. "Participants ' personal note-taking in meetings and its value for automatic meeting summarisation". *Information Technology and Management* 13.1 (2012), pp. 39–57.

[8] A. Bothin, P. Clough, and G. Bradley. "Quantitative analysis of individual differences in Note-Taking and talking behaviour in meetings". *Proceedings of IADIS Multi Conference on Computer Science and Information Systems.* Freiburg, Germany: IADIS Multi Conference on Computer Science and Information Systems, 2010.

[9] A. Braunstein, L. Granka, and J. Staddon. "Indirect Content Privacy Surveys: Measuring Privacy Without Asking About It". *Symposium on Usable Privacy and Security (SOUPS).* 2011.

[10] J. P. Carrascal, F. Bonin, S. P. Jose, R. De Oliveira, and N. Olivier. "Who can we trust? A Comparative Evaluation of Phone Call Annotation Techniques". *ACM SigCHI Conference (Submitted).* 2015.

[11] J. P. Carrascal, R. de Oliveira, and M. Cherubini. "A Note Paper on Note-taking: Understanding Annotations of Mobile Phone Calls". *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services*. MobileHCI '12. San Francisco, California, USA: ACM, 2012, pp. 21–24.

[12] J. P. Carrascal, R. de Oliveira, and M. Cherubini. "To Call or to Recall? That's the Research Question". *Transactions of Computer-Human Interaction (ToCHI) (Accepted for publication)* (2015).

[13] J. P. Carrascal, C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira. "Your Browsing Behavior for a Big Mac: Economics of Personal Information Online". *Proceedings of the 22Nd International Conference on World Wide Web*. WWW '13. Rio de Janeiro, Brazil, 2013, pp. 189–200.

[14] M. Cherubini and N. Oliver. "A Refined Experience Sampling Method to Capture Mobile User Experience". *Workshop of Mobile User Experience Research part of CHI'2009*. 2009.

[15] S. Consolvo and M. Walker. "Using the experience sampling method to evaluate ubicomp applications". *IEEE Pervasive Computing* 2.2 (Apr. 2003), pp. 24–31.

[16] J. Cox. *Many Android apps leak user privacy data*. http://bit.ly/b1modi. 2010.

[17] M. Csikszentmihalyi and R. Larson. "Validity and reliability of the experience-sampling method". *The Journal of nervous and mental disease* 175.9 (1987), pp. 526–536.

[18] R. De Oliveira and J. Carrascal. *Method for collecting and storing annotations associated to a voice audio data*. US Patent 8,792,863. July 2014.

[19] M. J. Ferreira, A. K. Caraban, and E. Karapanos. "Breakout: predicting and breaking sedentary behaviour at work". *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. ACM. 2014, pp. 2407–2412.

[20] J. E. Fischer. "Experience-Sampling Tools: a Critical Review". *Mobile Living Labs 09: Methods and Tools for Evaluation in the Wild, part of MobileHCI 2009*. ACM, 2009.

[21] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay. "MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones". *ACM MobiSys '07*. 2007.

[22] J. Gerken. *Longitudinal Research in Human-Computer Interaction*. 2011.

[23] W. Geyer, H. Richter, and G. D. Abowd. "Towards a Smarter Meeting Record–Capture and Access of Meetings Revisited". *Multimedia Tools Appl.* 27 (3 Dec. 2005), pp. 393–410.

[24] G. Iachello, K. N. Truong, G. D. Abowd, G. R. Hayes, and M. Stevens. "Prototyping and sampling experience to evaluate ubiquitous computing privacy in the real world". *ACM SIGCHI CHI '06* (2006).

[25] C. J. *Announcing the AMI meeting corpus*. ELRA Newsletter 11(1):3–5. 2006.

[26] C. Jensen and C. Potts. "Privacy policies as decision-making tools: an evaluation of online privacy notices". *ACM SIGCHI CHI*. 2004.

[27] C. Jensen, C. Potts, and C. Jensen. "Privacy practices of Internet users: Self-reports versus observed behavior". *International Journal of Human-Computer Studies* 63.1-2 (2005), pp. 203–227.

[28] V. Kalnikaitė, P. Ehlen, and S. Whittaker. "Markup as you talk: establishing effective memory cues while still contributing to a meeting". *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. CSCW '12. Seattle, Washington, USA: ACM, 2012, pp. 349–358.

[29] V. Kalnikaité and S. Whittaker. "Software or wetware?: discovering when and why people use digital prosthetic memory". *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. San Jose, California, USA: ACM, 2007, pp. 71–80.

[30] E. Karapanos, J. Jain, and M. Hassenzahl. "Theories, methods and case studies of longitudinal HCI research". *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM. 2012, pp. 2727–2730.

[31] M. V. Kleek, W. Styke, and D. Karger. "Finders / Keepers : A Longitudinal Study of People Managing Information Scraps in a Micro-note Tool". *Proceedings of the 2011 annual conference on Human factors in computing systems*. Vancouver, BC: ACM, 2011, pp. 2907–2916.

[32] B. Krishnamurthy and C. E. Wills. "On the leakage of personally identifiable information via online social networks". *Proceedings of the 2nd ACM workshop on Online social networks*. ACM. 2009, pp. 7–12.

[33] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. "A survey of mobile phone sensing". *Communications Magazine, IEEE* 48.9 (2010), pp. 140–150.

[34] M. W. Lansdale. "The psychology of personal information management". *Applied ergonomics* 19.1 (1988), pp. 55–66.

[35] R. Larson and M. Csikszentmihalyi. "The experience sampling method." *New Directions for Methodology of Social & Behavioral Science* (1983).

[36] D. Lin and M. C. Loui. *Taking the byte out of cookies: privacy, consent, and the Web*. Vol. 28. 2. ACM, 1998.

[37] M. Lin, W. G. Lutters, and T. S. Kim. "Understanding the Micronote Lifecycle : Improving Mobile Support for Informal Note Taking". *Proceedings of the ACM Conference on Human Factors in Computing Systems*. Vienna, Austria: ACM, 2004, pp. 687–694.

[38] D. MacMillan and E. Dwoskin. *Smile! Marketing Firms Are Mining Your Selfies. The Wall Street Journal*. http://on.wsj.com/ZSsro5.

[39] S. Menard. *Longitudinal research*. Vol. 76. Sage, 2002.

[40] F. Mueller and A. Lockerd. "Cheese: Tracking Mouse Movement Activity on Websites, a Tool for User Modeling". *CHI '01 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '01. Seattle, Washington: ACM, 2001, pp. 279–280.

[41] A. Narayanan and V. Shmatikov. "Myths and fallacies of personally identifiable information". *Communications of the ACM* 53.6 (2010), pp. 24–26.

[42] K. O'Hara, M. Perry, A. Sellen, and B. Brown. "Wireless world". Ed. by B. Brown and N. Green. New York, NY, USA: Springer-Verlag New York, Inc., 2002. Chap. Exploring the relationship between mobile phone and document activity during business travel, pp. 180–194.

[43] R. de Oliveira and J. P. Carrascal. "Towards Effective Ethical Behavior Design". *CHI '14 Extended Abstracts on Human Factors in Computing Systems.* CHI EA '14. Toronto, Ontario, Canada: ACM, 2014, pp. 2149–2154.

[44] A. Piolat, T. Olive, and R. T. Kellogg. "Cognitive effort during note taking". *Applied Cognitive Psychology* 19.3 (2005), pp. 291–312.

[45] G. Press. *A Very Short History Of Big Data. Forbes.com.* http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/.

[46] S. Richmond. *Eric Schmidt: Google gets close to 'the creepy line'.* http://tgr.ph/dp8BaH. Oct. 2010.

[47] C. Riederer, V. Erramilli, A. Chaintreau, P. Rodriguez, and B. Krishnamurty. "For sale: Your data, By: you". *ACM Hotnets*. 2011.

[48] D. Sarno. *Apple collecting, sharing iPhone users precise locations.* http://latimesblogs.latimes.com/technology/2010/06/apple-location-privacy-iphone-ipad.html. June 2010.

[49] J. Staiano, N. Oliver, B. Lepri, R. de Oliveira, M. Caraviello, and N. Sebe. "Money Walks: A Human-centric Study on the Economics of Personal Mobile Data". *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* UbiComp '14. Seattle, Washington: ACM, 2014, pp. 583–594.

[50] E. Steel. *A web pioneer profiles users by name.* http://on.wsj.com/riy5Qo. Oct. 2010.

[51] E. Steel and G. A. Fowler. *Facebook in Privacy Breach.* http://on.wsj.com/buDn6d. Oct. 2010.

[52] J. Valentino-Devries and E. Steel. *White House to Push Privacy Bill.* http://on.wsj.com/hlSR7E. Mar. 2011.

[53] S. Whittaker, P. Hyland, and M. Wiley. "FILOCHAT: handwritten notes provide access to recorded conversations". *Proc. of CHI '94.* Boston, Massachusetts, United States: ACM, 1994, pp. 271–277.

[54] S. Whittaker, S. Tucker, K. Swampillai, and R. Laban. "Design and evaluation of systems to support interaction capture and retrieval". *Personal Ubiquitous Comput.* 12 (3 Jan. 2008), pp. 197–221.

[55] T. N. Yorker. *The Virtual Interview: Edward Snowden.* http://www.newyorker.com/new-yorker-festival/live-stream-edward-snowden.

[56] M. Zawadziński. *Six Alternatives to Google and Facebook in the Battle for Online Privacy. International Business Journal.* http://www.ibtimes.co.uk/six-alternatives-google-facebook-battle-online-privacy-1472156.

Part I

Valuation of personal information in Web browsing

The next chapter describes a large scale study based on the monetary valuation of Personally Identifiable Information (PII) of Web users. In it we employed a refined Experience Sampling Method and a reverse second-price auction were used in order to overcome a series of difficulties that commonly appear when probing study participants about privacy-related topics. The chapter consists of this paper:

J. P. Carrascal, C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira. "Your Browsing Behavior for a Big Mac: Economics of Personal Information Online". *Proceedings of the 22Nd International Conference on World Wide Web*. WWW '13. Rio de Janeiro, Brazil, 2013, pp. 189–200

Chapter **2**

Your Browsing Behavior for a Big Mac: Economics of Personal Information Online

Juan Pablo Carrascal, Universitat Pompeu Fabra
Christopher Riederer, Columbia University
Vijay Erramilli, Telefónica Research
Mauro Cherubini, Telefónica Research
Rodrigo de Oliveira, Telefónica Research

Abstract

Most online service providers offer free services to users and in part, these services collect and monetize personally identifiable information (PII), primarily via targeted advertisements. Against this backdrop of economic exploitation of PII, it is vital to understand the value that users put to their own PII. Although studies have tried to discover how users value their privacy, little is known about how users value their PII while *browsing*, or the exploitation of their PII. Extracting valuations of PII from users is non-trivial – surveys cannot be relied on as they do not gather information of the context where PII is being released, thus reducing validity of answers. In this work, we rely on refined Experience Sampling – a data collection method that probes users to valuate their PII at the time and place where it was generated in order to minimize retrospective recall and hence increase measurement validity. For obtaining an honest valuation of PII, we use a reverse second price auction. We developed a web browser plugin and had 168 users – living in Spain – install and use this plugin for 1 month in order to extract valuations of PII in different contexts.

We found that users value items of their online browsing history for about 7 € ($\sim$ 10 USD), and they give higher valuations to their offline PII, such as age and address (about 25 € or $\sim$ 36 USD). When it comes to PII shared in specific online services, users value information pertaining to financial transactions and social network interactions more than activities like search and shopping. No significant distinction was found between valuations of different quantities of PII (*e.g.* one vs. 10 search keywords), but deviation was found between types of PII (*e.g.* photos

vs. keywords). Finally, the users' preferred goods for exchanging their PII included money and improvements in service, followed by getting more free services and targeted advertisements.

2.1 Introduction

A large part of the Internet economy operates by being reliant on online advertisements. In recent years, targeted advertising has become an attractive offering where targeting is facilitated by the collection of large amounts of personally identifiable information (PII) of end-users. However, this collection comes at the cost of erosion of privacy of end-users. Web service providers are collecting more PII about the end-users, often *outside* the scope of their application (*e.g.*, search engines collecting browsing information via third party aggregators like Doubleclick *etc.* [35]). At the same time, users are becoming more aware of various privacy breaches [4, 37, 41], attracting the attention of regulatory bodies [40].

The economics of the online ecosystem can be summed up by the pithy adage 'if you are not the consumer, then you are the product', more specifically, the product being end-users' PII. In such an arrangement, it is easy for service providers to attach a value on each users' PII, based on the revenues they can extract. However, for users to perform a cost-benefit analysis, where the cost is loss of privacy, and the benefit is the service they obtain in return, it is important that they first know the value of *their* PII they are trading away.

There has been a lot of work on users valuating their information [**cvrcek2006**, 9, 17], and in general users' perceptions about privacy [2, 3, 13]. However, there has been surprisingly little to no work on valuating *web-browsing* information, even though it is known that privacy leakages can occur while web-browsing [25, 35]. In this paper, we focus on understanding the value that users attach to their own PII,[1] while web-browsing.

It is challenging to extract the value that users' put on their own PII. First of all, the valuation could change based on *context*. For instance, the value that a user puts on the fact that she is searching for a restaurant can be different than when she is searching for cancer drugs. Even using the same keywords while searching, but in a different context, could lead to different valuations of the same PII (*e.g.* searching for leisure while at home or at work). Past work done in this domain has included valuating personal information (*e.g.*, weight, age [17]) as well as location information [9]. However they all rely on surveys that do not leverage contextual factors when the PII was generated and/or released.

In order to leverage these contextual factors, we rely on the refined Experience Sampling methodology (rESM) [7] (Sec. 2.3). This data collection approach probes users at appropriate times to obtain more reliable answers, as questions are presented to users in-context and hence minimizes retrospective recall and possible errors that come with such recall. We implemented rESM by means of a browser plugin (Secs. (2.3, 2.3)). Users get asked specific questions when they access different types of content/services (social networks, search engines, finance sites, etc.). We recruited 168 participants living in Spain with a diverse range of demographics (Sec. 2.3), and had them participate in our study for 1 month. We used a reverse second price auction to obtain an honest valuation for different types of PII (Sec. 2.3). We also use

[1]We focus on monetary value assigned by the user to their information, although one can imagine other notions of value and utility like satisfaction etc. We consider money as we are interested in the overall ecosystem of online services that partly hinges on monetizing PII. Secondly, money is a tangible concept and easier to arrive at as opposed to user happiness. We will consider other notions of value in future work.

our methodology to obtain users' perceptions and awareness of the economic usage of their PII by online service providers (Sec. 2.4).

The major findings of this work are:

- Users value PII related to their *offline* identity (age, gender, address, economic status) at about 25 € ($\sim$ 36 USD), and this value does not change when the user is probed in different contexts (*e.g.* browsing search sites, webmail, etc.).
- Moreover, users value PII related to offline identity higher than PII related to browsing activity, which is about 7 € ($\sim$ 10 USD)[2].
- In terms of valuating service specific PII (*e.g.* photos uploaded to social networks, search keywords, online purchases, etc.), users gave higher valuations to interactions in online social networks (12 € or $\sim$ 17 USD) and finance websites (15.5 € or $\sim$ 22 USD), when compared to activities like search (2 € or $\sim$ 3 USD) and shopping (5 € or $\sim$ 7 USD).
- The majority of participants in our study were aware that their PII is being collected when web-browsing, and while they were positive about their PII being used to improve services, they were also negative that it could be monetized by service providers.
- Our results reveal that users prefer to trade their PII for monetary rewards or improved services more than trading it for additional free services or targeted advertisements.

2.2 Research challenges

The work presented herein aims at answering the following two research questions:

- **RQ1:** What monetary value do users assign to different types of PII[3] while being online?
- **RQ2:** What are the perceptions of users vis-a-vis their PII being monetized, improving existing services and for personalized advertisements?

In order to answer these questions, it is of great importance to consider a user-centric approach. Previous work addressed related questions and using techniques such as post-study surveys and diaries [6, 29]. These traditional methods could have relevant drawbacks when trying to gather answers for the questions posed above. For instance, consider a web user Alice who browses the web on a daily basis. On a given day, Alice searches for symptoms pertaining to an illness she suspects she has. Alice then sends an email to her friend Bob about this illness. Some time later, she takes part in a traditional survey and/or diary study that aims at answering the aforementioned research questions. These techniques would most likely not collect accurate responses due to a number of reasons, including:

[2]Equivalent to a Big Mac meal in Spain, circa 2011. Hence the title of this paper.

[3]A strict definition of PII does not include browsing behavior. However it has been shown that information leaked on the web via browsing can be combined to form PII [26], so we use PII to refer to all the information that a user can leave online, knowingly or unknowingly.

1. Retrospective recall: Self-report recall surveys and diaries suffer from recall and selective reporting biases [19, 8]. Alice may not be able to remember what she searched for some time ago, or what emails she exchanged. The more time has passed since these actions occurred, the harder it is for Alice to remember and report them accurately in a survey/diary study.

2. Validity: Alice's valuation of the illness related keywords also depends on the context when she shares her PII (*e.g.* how, where and when she came up with the keywords in the first place). So even if Alice is later given keywords related to illness and asked to valuate, she might not remember or recreate the conditions she had when she came up with the keywords and may end up assiging an incorrect value to them.

3. Burden: Alice can be asked to note down her activities and assign values to different PII in a diary. This is however burdensome for Alice, who might even consider dropping out of the study.

4. Honesty: If Alice needs to valuate several different PII in a long survey or in a daily diary study, she will most likely get bored and provide random values just for the sake of getting the job done.

5. Engagement: In order to address response fatigue and have Alice valuate information under a diverse set of conditions as accurately as possible, we need her to be motivated. Answering multiple survey questions could lead to a significant number of drop-outs if the study does not include an element of engagement.

Next, we present the methodology of our study describing *how* we tackled these challenges towards providing trustful results for our research questions.

2.3 Methodology

Tackling Challenges

Users' Need for Recall

In order to address the challenge of users' *retrospective recall* for PII valuation, we used a refined version of the Experience Sampling Method (rESM). Experience Sampling involves asking participants to report on their experiences at specific points throughout the day. The method was originally developed in the psychology domain [5] and recently adapted successfully in many studies of Human-Computer Interaction [8, 18, 19, 28]. As Cherubini *et al.* highlighted [7], the main advantage of ESM is its ability to preserve the ecological validity of the measurements, defined by Hormuth *et al.* as "the occurrence and distribution of stimulus variables in the natural or customary habitat of an individual" [16]. This method is better than recall-based self-reporting techniques by "probing" the participant in close temporal proximity to when a relevant event was produced. One of the drawbacks of the method is that participants often are sampled at random times and therefore the probing might be invasive for many participants. This is why in recent years some researchers have proposed to refine the method by modeling the participants' context [7, 11], and this is what we use.

Validity of PII Valuations

As a means to perform rESM and further address the challenge of *validity* of valuations, we instrumented the web browser of participants with a plugin that was able to log the website they were browsing and probe them at the exact time a certain PII was being shared online. At a high level, the study operated as follows. First, participants installed the plugin and browsed as usual. Then the plugin would categorize every website the user would browse into one of the eight categories: Email, Entertainment, Finance, News, Search, Shopping, Social, and Health. These categories closely correspond to the eight popular categories that online ad-networks like Doubleclick[4] use, as we are interested in the monetary aspect of PII. In addition, the plugin was able to sense when the user was changing context and use this information to trigger a popup, which would have two goals: (i) collect the user's valuation of specific PII related to the category of the site the user is browsing, via an auction and (ii) inquire the user about perceptions of PII usage. Finally, the popup would send this data to a remote server for posterior data analysis.

Engagement and User's Burden

With respect to preventing the user's *burden*, we adjusted the frequency of the popups triggered by the browser plugin and also allowed users to skip them if they wanted to. In order to provide users with an element of *engagement* to participate actively in the study, we created a real setting where participants could trade their PII for money based on their own valuations. More specifically, participants received the winning monetary value of every auction they won.

Honest PII Valuations

In order to persuade participants to provide an honest valuation of their PII, we relied on a reverse second price auction: given a set of k bids, pick the lowest bidder as the winner, and pay that person the amount equivalent to the second lowest bid. We chose this auction mechanism for the following reasons: (i) this mechanism has the strong property of being truth telling; the best strategy for participants in the auction is to be honest about their valuation [23], (ii) it has been used before for valuating location information [9], and (iii) it is relatively easy to explain.

We allowed positive amounts, including 0, in increments of one cent. We also gave the user a choice to not participate in the auctions at all. This was necessary to cover cases where users felt overwhelmed with participation and cases where users did not even want to disclose the fact that their PII was worth a very high amount. In order to reinforce the notion that the user would indeed part with their PII if they won, we had the user verify that they understood their data would be sold in a second popup. We ran an auction whenever we had 20 bids per category from 20 different users. Multiple auctions were run during the study.

All winners of the auctions were notified by email with information including their winning bid, contextual information of the bid (date and time of bid, PII, website they were on). We reinforced the message that as they won, we would use their PII (showing the exact PII they bid on), for a period of 6 months. Likewise, we sent a similar email to the remainder participants, conveying that as they lost the bid, their PII would *not* be used. Only after the end of the study we informed participants that

[4]Doubleclick has more than eight major categories and more than 600 subcategories. We chose eight as a good trade-off between obtaining detailed information without annoying the user, given that the rESM probing would increase linearly with the number of categories.

their PII was actually not going to be used for any commercial purposes. For all our communication with users, we used neutral language with regards to privacy, so as to not prime them one way or another, following the findings in Braunstein *et al.* [6].

Participants

Participants were recruited using a survey published via a major Web portal in Spain. A total of 168 participants (93 male, 55%) installed the Firefox browser plugin and completed all requirements of the study. All participants were users of the Firefox browser and hence had it installed on their computer. Participants' age ranged between 18 and 58 years old ($\bar{x} = 31.83$, $s = 8.15$). With respect to their educational level, 1% did not finish primary school, 8% finished primary school, 14% did secondary school, 75% had a university graduate degree, and 2% a post-graduate degree. Socioeconomic status was also diverse: 28% of the sample said their annual gross salary to be lower than 10K €, 25% said it was between 10K € and 20K, for 22% it was in the range of 20K € and 30K, 11% between 30K € and 40K, and 10% reported earning more than 40K € per year (4% preferred not answering this question). All participants lived in Spain and the vast majority were of Spanish nationality (94%). Each participant was given a gift card voucher worth 10 € (∼ 14 USD) for taking part in the whole study. Our ethical board and legal department approved the experiment. Participants were debriefed about what was being logged and instructed on how to temporarily disable or remove the plugin. Participants were free to leave the experiment at any time.

Apparatus: Browser plugin

In order to capture the browsing context of the users we developed a system consisting of two parts: a browser plugin – to be installed in participants' browsers – and a web server that communicated with the plugin, sending configuration information and receiving data from it.

Firefox Plugin: The plugin had three main tasks. First, it captured and stored all browsing activity of the user. This consisted of the url, time of page access, and a unique ID we assigned to each browser. This data was stored on the local machine and sent to the server at regular intervals.

The second main task of the plugin was to categorize visited websites into one of the eight categories mentioned in Sec. 2.3. In order to do this, we relied on a hard-coded list of 1184 popular sites from different categories for Spain, gleaned from alexa.com. Although some popular sites like Facebook can host content pertaining to health or entertainment, we hard-coded it to Social.[5] For sites that were not on Alexa, we resolved them into categories in real-time by relying on an approach implemented in another browser plugin called Adnostic [39]. The basic idea is to perform a cosine similarity between the set of keywords present on the site the user visits and a massive corpus of words that are associated with specific categories. The category with the highest similarity score is used and the appropriate text is presented in the pop-up. Testing on individual unclassified and Alexa-classified websites gave a high level of accuracy (approx. 98% correct classification).

Third, the plugin presented the participants with two independent popups, as described earlier. The first popup displayed auction questions and the other displayed

[5] Such a monolithic categorization does have limitations; large service providers like Facebook or blogspot host content belonging to multiple categories. However, we consistenly pick the first category as put out by Alexa. This ensures that we do not have any false positives – Facebook will always be categorized as Social. In addition, the questions we pose users (Table 2.1) for a certain category are always consistent; questions on Facebook are always related to Social. We leave a detailed categorization mechanism to future work.

questions related to exploitation of PII. These were configured to be switched on or off from the server. From a UI perspective, the popup displayed the text of relevant auction question, with the type of PII in the auction in bold text, to highlight what is actually being traded in the auction. There was a box below the text where the user could enter an amount, and there was a radio button below the box where the user could select to not participate in the auction. Fig. 2.1 shows an example of a popup for category SOCIAL.

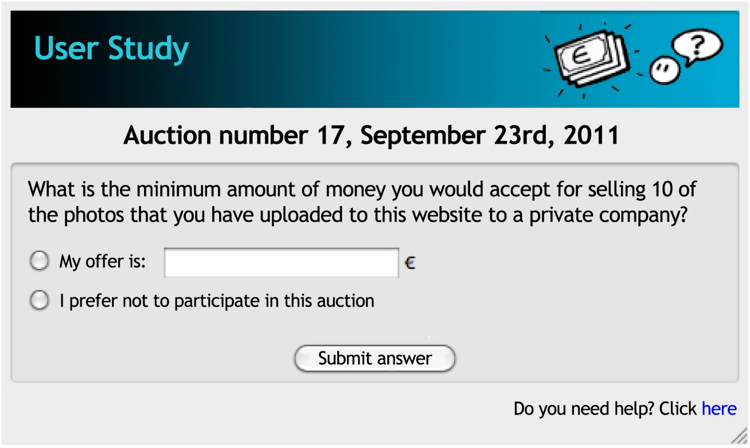


Figure 2.1: The auction popup. Each auction game was identified by a sequential number and a date. The participant had the option to either enter a bid or to not take part in the auction.

Server: We developed a simple, highly responsive webserver that the browser plugin would sync with at regular intervals. The server accepted data (bids, responses to questions) from the plugin and stored it in a database. The main function of the server was to run auctions. For each category of website, and for each type of PII (there were 4 types per category, as explained in Section 2.3, questions a1–a4), we set an auction to run once 20 bids were in. We pooled all these auctions, ran them once every morning, and sent out results to participants via emails. The entire process was automated.

Procedure

The study ran for a period of two months from mid-July to mid-September, 2011. After following our study advertisement on a famous Web portal and signing up for the study, participants were selected based on our unique filtering criteria – users of the Firefox web browser – and invited via email to participate in our study. We asked participants to fill a recruitment questionnaire which focused on demographics as well as their general Internet privacy knowledge and perception. We explained to participants that the study consisted of three phases: (1) an initial week where the popups were inactive, and their browsing behavior would be collected, (2) the actual study that lasted four weeks where popups were active, and (3) the final questionnaire.

Phase 1. During the first week, the plugin silently recorded the browsing behavior of participants (with their consent). The information captured during this phase was used to record a user's baseline browsing behavior. We used this information to make sure that our popups were not interfering with the normal browsing behavior

Table 2.1: Questions asked during the different phases of the study

Code †	Question	Remarks	Type
r1	**Are you concerned about protection of your private data on the Internet?** *[5- A lot, 4- Much, 3- Somewhat, 2- Little, 1-Never]*		5-point
r2	**Do you distrust the way the websites you visit use your data?** [5- I distrust all, 4- Only some, 3- I do not care. 2- Only few, 1- I do not distrust]		5-point
r3	**Do you read the privacy policies of the web sites that you visit?** [5- Always, 4- Often, 3- Sometimes, 2- Rarely, 1- Never]		5-point
r4	**How much do you know about current legislation about data protection?** [5- A lot, 4- Much, 3- Something, 2- A little, 1-Nothing]		5-point
a1	**What is the minimum amount of money you would accept for selling to a private company information about your age, gender, salary and address?**	Context Independent	Numeric
a2	**What is the minimum amount of money you will accept for selling to a private company details about your presence on this webpage?**	Context Dependent	Numeric
a3	**What is the minimum amount of money you would accept for selling * to a private company?**	Cotext Dependent (1 category-customized PI item*)	Numeric
a4	**What is the minimum amount of money you would accept for selling 10 * to a private company?**	Context Dependent (10 category-customized PI items*)	Numeric
p1	**Are you aware that the web site you are currently visiting might generate revenues from the information ** ?** [5- I was fully aware, 4- I did know, 3- I was not fully aware, 2- I figured but I was unsure, 1- I did not know]	Context dependent. Category-customized**	5-point
p2	**How comfortable do you feel knowing that the web site you are visiting might generate revenues with the information you share?** [5- Very comfortable, 4- Comfortable, 3- I do not care, 2- Uncomfortable, 1- Very uncomfortable]	Context independent	5-point
p3	**If the company that uses this information does it in order to offer you a better service, how would you feel?** [5- Much better, 4- Better, 3- The same, 2-Worse, 1- Much worse]	Context independent	5-point
p4	**If the company that uses this information does it in order to present you with customized advertisements, how would you feel?** [5- Much better, 4- Better, 3- The same, 2-Worse, 1- Much worse]	Context independent	5-point
f1	**On day "X" you bid "Y" Euros for sharing "Z", and that was your lowest bid. What was your main motivation for bidding that low?** [answers manually categorized as: 1- to win the bid, 2- not important information, 3- fair value, 4- other] **On day "X" you bid "Y" Euros for sharing "Z", and that was your highest bid. What was your main motivation for bidding that high?** [answers manually categorized as: 1- to win the bid, 2- to prevent selling, 3- fair value, 4- other]	Letters X, Y and Z were replaced by the actual bid date, bid value, and traded PI respectively	Text
f2	**I do not see a problem if *** generate revenues from my personal information as long as they:** [1. pay me some money, 2. improve their existing service, 3. provide more free services, 4. recommend me things that I like, 5. no need to change what they are currently doing, 6. other]	Category-customized***	Nominal
f3	**With respect to the services offered to you by ***, do you believe they have significant operational costs?** [1. Yes, 2. No, 3. I do not know] **With respect to the services offered to you by ***, do you believe they have significant revenues?** [1. Yes, 2. No, 3. I do not know]	Category-customized***	Nominal
f4	**In the case of having a market that you could sell your personal information (*e.g.,* clicks on a website, history of pages visited, contact details, %bank account details, *etc.*), who would you trust to handle that information?** [options to rank: Telecommunication company, Government, Bank, Insurance company, Yourself only, Other]		Ranking

† Codes refer to different phases of the study: 'r' stands for recruitment questionnaire; 'f' stands for final questionnaire, 'a' stands for auction game, and 'p' stands for perception of PI monetization.

* Customized per category: Email: "data about one of the contacts that you email more often"; Entertainment: "that you have visited this web Site"; Finance: "details about your last financial transaction"; News: "the last news or articles that you read"; Search: "the words that you used in your last search"; Shopping: "details about the last product or service that you bought online"; Social: "one of the photos that you have uploaded to this web site"; and Health: "details about the last time you were sick".

** Customized per category: Social: "you share with your friends"; Entertainment: "you share when you fill its forms"; Health: "you are looking for here"; Search:
"your search history"; Finance: "about your finance might be shared with other companies"; Email: "the content of your email messages"; Shopping: "your shopping behavior"; News: "your news reading history".

*** Customized for 3 categories: Social: "online social networks"; Email: "e-mail providers"; Search: "search engines (e.g. Google, Yahoo, Bing, etc.)".

of the participants. We extracted the frequency distribution across the visited sites for every user. We will refer to this as the user's *fingerprint*.

Phase 2. During the experiment, the plugin displayed popups when the participants were browsing the internet. The popups contained two kinds of questions: questions about valuing PII – used as the basis for the auctions – and questions on participants' perceptions and knowledge of PII. To avoid the popups being too invasive the plugin displayed at most one popup per category per day. Additionally, there was a *minimum* delay of 10 minutes between any two popups.

Phase 3. At the end of the experiment, we asked participants to fill in a post-study questionnaire that aimed to clarify the main results obtained during the study.

Measures

Table 2.1 summarizes the questions that we presented to the user during the entire study, along with their associated measures. We customized popup questions (a1–a4 and p1–p4, described below), according to the *context* they were asked in. Questions about PII unrelated to the website currently being visited are *context independent*. For instance, a question about a user's age while at a news website could be considered context independent. Conversely, questions about PII that *are* related to the current website are called *context dependent*. For instance, financial transactions on a banking website. Additionally, the content of some questions was customized according to the category of the website they refer to, as explained in Table 2.1.

Recruitment (r1–r4). Questions in the recruitment questionnaire aimed to gauge participants' knowledge of privacy related issues.

Privacy Valuation (a1–a4). These questions were presented to participants with a plugin popup during the auctions, and asked them to bid the *minimum* value they would accept to sell their PII. We were deliberately vague about *how* we were going to use their PII for two reasons: (i) to realistically reflect the conditions that exist today, as there is little knowledge of how one's PII is being used (targeted advertisements, price discrimination [30]), and (ii) not to bias the user by providing a specific use case of their PII; for instance using PII for behavioral targeting can be construed positively or negatively.

Question a1 is context independent. Its purpose was to assess the validity of our measures by contrasting with results from a2. Indeed, a2 and a3/a4 were context dependent, but while the former asks about the same PII item across categories, the latter is customized for each category of websites. Our goal was not to produce generalized estimates of context valuation but rather to understand whether online context had an influence on the valuation that people attach to certain types of PII. Furthermore, we chose to ask a2 as this is the information that most entities engaged in large scale tracking across the web (like DoubleClick) have access to, and hence can monetize. These are often referred to as 'third' parties. Questions a3-a4 are category specific and in most cases, this PII is available only to the service provider actually providing that service (*e.g.,* photos on social networks, financial transactions, purchase history on e-commerce sites, *etc.*). These are referred to as publishers or 'first' parties.

Privacy Perception (p1–p4). These questions were also presented with a plugin popup, and were designed to understand if users are aware of monetization of their PII by online entities.

Exit (f1–f4). These questions were asked in the final questionnaire in order to clarify results obtained during the study.

Statistical Analysis

Nonparametric analysis was applied to the data considering the ordinal nature of some observed variables and that continuous variables did not follow the normal distribution. Given that participants browsed web pages in their natural environment without being forced to visit sites from all categories mapped in our study – thus preserving ecological validity, our sample had several missing values across categories. Removing subjects that did not provide information for all categories – as they did not browse all types of web pages – would significantly reduce the generalization power of our results and yield unrealistic findings based on the assumption that everybody browses web pages from all categories considered in this study. Therefore we opted for *not* using related sample analysis. Hence differences between median bid values (or Likert scale measures) across categories were tested using the Kruskal-Wallis test and the Mann-Whitney test whenever appropriate. Associations between ordinal/interval variables were assessed using the Spearman's Rho test. Comparisons between related sample distributions of dichotomous variables were performed using both the Cochran's Q test and the McNemar test. The level of significance was taken as $p < .05$.

2.4 Auction and Survey results

We summarize the main results obtained towards addressing our two research questions. Our results are mainly reported in Euros, and at the time the conversion was approx. 1 € gives 1.42 USD.

Results for RQ1: Monetary value of PII

Effect of pop-ups on browsing behavior. We used the $L2$ distance between participants' first week's "baseline" fingerprints and their fingerprints for the second week of the study after pop-ups were turned on and found little differences (165 users had less than 5% difference). Specifically, only three users (2% of the sample) had high browsing behavior deviation and reported being on vacation during the second week, thus explaining why they used their browser sparsely. These findings indicate that users *did not* deviate from their 'normal' browsing behavior when participating in the study.

Representativeness of categories. We look into the bidding behavior of the whole sample ($N = 168$) while browsing websites as they map to each of the 8 categories and also in relation to the nature of the information being sold (see questions a1-a4 in Table 2.1). Overall, participants visited websites from all of the eight categories, Health being the least visited category (Search: 82%, Entertainment: 82%, Social: 78%, News: 76%, Finance: 75%, Shopping: 75%, Email: 64%, Health: 2%). Given the lack of representativeness for the number of subjects visiting health related web pages, we therefore consider only seven categories when comparing participants' bids and other relevant measures across categories.

Bids on context independent PII. With respect to selling their PII that is related to their offline identity (*i.e.*, age, gender, address and salary; see question a1 in Table 2.1), we found no significant difference among participants' median bid values across categories ($p = .702$). Note that this result was somewhat expected as question a1 was context independent – no mention was made to selling the participants' PI

Table 2.2: Median bid values per category calculated from participants' median bids in each category (1st and 3rd quartiles shown between brackets)

Questions	Email	Entertainment	Finance	News	Search	Shop	Social	All Categories	*p*-value
a1	24.5 [1.6, 97.4]	26.5 [3, 115]	20.2 [3.4, 100]	25 [4, 150]	20 [2.5, 150]	10 [2, 100.2]	15 [3.5, 60]	25 [5.5, 151]	.702
a2	5 [1, 25]	5 [0.9, 20]	3 [1, 20]	5 [1, 43.5]	4 [0.7, 20]	5.2 [1, 30]	7.1 [1, 25]	7 [1, 38]	.569
avg(a3, a4)	6 [2, 89]	2 [1, 14.3]	15.5 [3.8, 229.5]	2 [0, 13.5]	2 [1, 12.8]	5 [1, 20.5]	12 [2, 81.5]	5.5 [1, 39.3]	$< .001$

to an entity related to the website they were browsing. The overall median bid value across categories was 25 €.

Bids on context dependent PII. When probed about selling clicks they performed on a given web page (see question a2 in Table 2.1), which represents their browsing behavior, participants' median bids were again not significantly different across categories ($p = .569$). In this case, the overall median bid value was 7 €. Median bid values for highly category specific PII – as captured by questions a3 and a4 in Table 2.1 – revealed significant differences across categories ($p < .001$). The highest median bid values (in euros) were from categories Finance ($\tilde{x} = 15.5$), Social ($\tilde{x} = 12$), and Email ($\tilde{x} = 6$), with Finance similar to the latter two categories ($p = .31$ and $p = .09$ respectively) and significantly different from the remaining categories (Shopping=5, News=2, Entertainment=2, Search=2; $p < .001$). Table 2.2 summarizes the most relevant descriptive statistics of median bid values per category.

Effectiveness of the auction: Categorization of the participants' free text responses to why they bid so low/high in their lowest/highest bids (in euros) was categorized manually with an acceptable inter-rater reliability (lowest bid: $K = .77, p < .001$; highest bid: $K = .78, p < .001$). Even considering the extreme case of each participant's lowest bid, only 15% explained that they bid that low in an attempt to win the auction. The majority said it was because the information was not important ($50 - 51\%$) or they thought it was a fair value ($8\% - 10\%$), or due to some other reason ($25\% - 26\%$). On the other hand, highest bids were mainly due to prevent selling important information ($53\% - 58\%$), although also being explained as a fair value ($16\% - 22\%$) or due to some other reason (22%). Note that there are subtle differences between a 'fair' value assigned to information, and very high/low values assigned because information was very important or not important at all. Fair value indicates a more reasoned approach while bidding very high values indicates focus on the outcome (no selling under any circumstance). Bidding very low values indicates nonchalance; value of information is so low that it isn't worth reasoning about. Very few participants ($3\% - 4\%$) explained their highest bid as a strategy to win the auction. These results could indicate that the rules of the reverse second price auction were understood. Overall, the results indicate that the auction scheme is indeed effective for truth telling, given that the majority gave reasons of fair value or worth of information for bids instead of trying to game the system.

Bulk PII effect. We verified no significant difference between the median bid value for all categories in question a3 ($\tilde{x}_{a3} = 5$) and in question a4 ($\tilde{x}_{a4} = 5$, $p = .59$). This finding indicates that the amount of information being sold was not a factor for participants when placing their bids, as they valued one piece of information (question a3) and 10 pieces of information of the same type (question a4), to be used for the same period of time if sold, in a similar way.

Winning bids and pay-outs. Considering the 40 subjects that won at least one

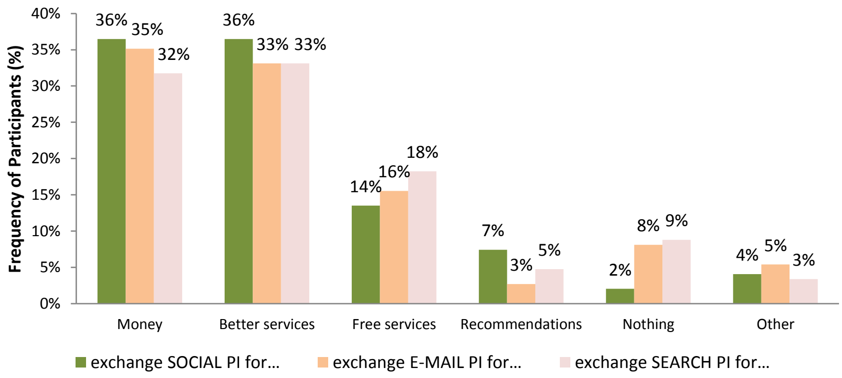


Figure 2.2: Participants' preferred goods in exchange of PI to online social networks, e-mail providers, and search engines.

auction, their median winning bid was of 5 cents of Euro ($min = 0, \bar{x} = 0.19, max = 2.29$). Even though we allowed a bid of 0 as a valid bid, only seven winners bid 0 on 11 occasions, out of 5000+ bids. The other winners' bids were strictly positive. Finally, as we used the reverse second price auction, the median payout was actually 45 cents of Euro ($min = 0.01, \bar{x} = 0.65, max = 5.69$). We describe this result for completeness.

Trading PII for alternative goods. At the end of the study, we wanted to understand if there were preferred goods participants would be willing to trade their PII for, and if the preferred goods would be different across the most popular categories, *i.e.* SOCIAL, EMAIL and SEARCH (question f2 from Table 2.1). According to our results, participants' first choice was to either exchange their PII for money ($32\%-37\%$ of participants) or have improvements in services they are currently using ($33\% - 37\%$). The second choice was to receive more free services ($14\% - 18\%$). No significant differences were found between distributions of each strategy across the three categories (money: $Q = 2.000, p = .37$, better services: $Q = 1.042, p = .59$, free services: $Q = 1.805, p = .41$). Interestingly, receiving PII-based recommendations was the third option for social networks (7%), but rather the fourth for EMAIL (2%) and SEARCH (4%), with a significant difference between them ($Q = 6.167, p = .046$). Fig. 2.2 shows a graph comparing the participants' preferred monetization strategies across the three categories.

Relationship between bids, demographics, and privacy. We next looked into significant associations between variables captured in the recruitment questionnaire and the participants' bids. Our findings reveal a medium negative correlation between participants' age and their median bid values for question SOCIAL-a3 ($n = 64$, $\rho = -.276$, $p = .03$). Similarly, age is negatively correlated to the combination of questions SOCIAL-a3 and SOCIAL-a4 ($n = 69$, $\rho = -.287$, $p = .02$), thus providing evidence that the older people are, the lower they tend to bid on photos they share online. Furthermore, we found a medium positive association between gender and median bids for question EMAIL-a3 ($n = 45$, $\rho = .333$, $p = .03$). This result indicates that men might value information related to their email contacts more than women. Correlations between income levels and bid values were not significant. Finally, we found medium negative correlations between participants' education level and their median bid values for question a2 in most categories (ENTERTAINMENT: $\rho = -.277$, Finance: $\rho = -.282$, Search: $\rho = -.235$, Shopping: $\rho = -.32$).

We also correlated bid values with responses provided to privacy-relevant questions in the recruitment questionnaire. Positive correlations were found between being worried about online data protection and higher bids on context independent PII (question a1, Entertainment: $\rho = .252$, Finance: $\rho = .278$, Search: $\rho = .23$).

Results for RQ2: Perceptions around usage of PII

Results presented in this subsection contribute to the understanding of how users' perceive the economic usage of their PII by online service providers. Note that we considered only the first answers that participants gave to questions p1–p4 per category. This decision guaranteed that their initial opinion would be taken into account instead of a potentially biased opinion due to the effect of long exposure to the study.

Knowledge of PII-based monetization. Participants were aware that PII shared on a particular web site could be used to generate revenue (question p1, $\tilde{x} = 4$, $q1 = 2$, $q3 = 4$). Moreover, no significant difference was found between median ratings across categories ($p = .107$). This finding suggests that knowledge of PI-based monetization is related to Internet services in general and not to a particular set of services.

Comfort with PII-based monetization. In question p2, participants revealed how comfortable they were with web sites extracting revenue out of their PII. With a median rating of 2 ($q1 = 2$, $q3 = 3$), they reported being uncomfortable with it, and this feeling was shared across categories as no significant difference between participants' median ratings per category could be found ($p = .429$). From this finding, we conclude that the act of monetizing from users' PII is what generally makes people uncomfortable, and not the type of online service providers the revenue will go to (*e.g.*, finance, search, *etc.*).

Improving services with PII. Although not comfortable with their PII being monetized, participants pointed out that they would like online companies to improve their web services using their PII (question p3, $\tilde{x} = 4$, $q1 = 3$, $q3 = 4$). No significant difference was found between participants' median ratings across categories ($p = .869$). This finding is consistent with results presented in Fig. 2.2 about money and improved services being the participants' preferred PII monetization strategies.

PII-based publicity/ads. Finally, subjects were indifferent with regards to online service providers making personalized publicity/ads by using their PII (question p4, $\tilde{x} = 3$, $q1 = 3$, $q3 = 4$). Once again no significant difference could be identified between participants' median ratings across categories ($p = .686$). This finding suggests that leveraging users' PII to provide them with personalized ads generally have neither a negative nor a positive impact on people.

Perception of costs and revenues. Participants were more confident about revenues than costs of providing social network, email, and search services (answered "do not know": 3% vs. 29%, 10% vs. 24%, 6% vs. 21% respectively). In general, most participants agree that these service providers have high revenues (93%, 69%, and 89% respectively) and high costs (43%, 45%, and 53% respectively), but the perception of revenues is significantly higher than costs ($p < .001$ for each of the three categories). Finally, more participants perceived search services to have significant costs compared to email (68% $n = 117$ vs. 58% $n = 113$, $p = .02$), while more participants perceived social network and search services to have significant earnings compared to email (97% $n = 143$ vs. 77% $n = 133$, $p < 001$; 94% $n = 139$ vs. 77% $n = 133$, $p < .001$ respectively). These results reveal that users might consider social network services to be more profitable than search or email services.

2.5 Discussion

The conclusions that can be derived from our results (Sec. 2.4) are:

Users value offline PII more and online PII less: If we consider the results for a1, the question on valuating offline PII (Sec. 2.4) users consistently bid high values for their offline PII like age, gender, address and financial status; pieces of PII that form their offline identity, to trade with online entities. Likewise, users attach lower value to a2, a3 and a4, PII that mostly has to do with their online behavior (a2 is exclusively about browsing history, the other two are about online transactions). Digging deeper, we also note that users tend to value category-specific PII (a3 and a4) on Finance and Social, categories that are more explicitly intertwined with one's offline identity, more than Search and News.

We conjecture that the difference in valuation exists because of lack of awareness. Offline PII is easier to valuate as it is more explicit. It is harder to understand the implications of having your PII continuously tracked, data-mined, and linked to an offline identity [12, 36]. As a consequence, users value such PII less.

Higher valuations than previous studies: Previous studies on valuation of privacy or personal information have reported lower values for various PII than what we encounter in our results [22]. This could be for two reasons. First, we use experience sampling that puts emphasis on valuating PII during web-browsing at the appropriate time, and second, specific properties of the demographics (Spanish citizens) could play a role. We note here that the regulatory framework surrounding privacy in EU is much stricter than in other parts of the world and this could affect the norms related to privacy and personal information of users. We also note that cultural norms can play a role. Addressing these concerns is beyond the scope of this work.

Users do not distinguish between quantity of PII, but type: We compared the median bid values for a3 and a4 across categories and found no significant difference. These two auction questions differ only in quantity of information being traded, with the type of PII and the context remaining the same. As reported earlier, there are significant differences between type (Finance and Social being higher than Search, Shopping etc.)

We correlated the values with demographic information as well as the responses to the privacy related questions (r1-r4). We found weak to no correlation. A possible conjecture can be on the lines of what is reported in [**cvrcek2006**], that users factor in diminishing returns of more information in their valuation – although we have no evidence to support or refute this conjecture.

Older users less concerned about online PII: When we correlated bid values against demographics, a high (negative) correlation occurred between age and category specific PII on Social, Entertainment and News, and more so while valuating bulk information (a4). For Social, this can be linked to the fact most older users do not use online social networks, let alone upload photos to online social networks.[6] This result is in contrast to previous work that stated that older users are generally more concerned about their privacy, while being online [32].

Users do not like monetization of their PII: Users are negative when it comes to their PII being used for monetization by entities (question p2), despite knowing that online entities collect and use their PII for monetization (p1). In addition, they prefer their PII to be used for improving the services they are offered (ap3), across all categories. On the one hand, these results are expected – the former deals with monetization of a good (PII) that users probably perceive as theirs, while users view the latter as a positive outcome of their PII being exploited.

In order to understand why users are negative about their PII being monetized, one can posit that most users are not aware of the functioning of the online ecosystem in place – they do not perceive that the services they get for 'free' (storage in Gmail, Bing search, Facebook etc.) actually are expensive (large datacenters, equipment and bandwidth costs) and while users are aware of their PII being monetized, they are possibly not aware that large parts of that monetization goes towards providing them with a 'free' service.

However, when we look at the results from the post-study questionnaire (question f3), we find that users indeed seem to be aware of the costs and revenues of different services with most users assigning *higher* revenues than costs for services. Taken together, users' negative attitude about monetization of their PII by services can be due to a feeling of unfairness.

Users are indifferent when it comes to the use of the PII to send them personalized ads (p4), again across categories. This is somewhat in contrast to results in [29] where the authors report that 64% of the survey respondents (all Americans) find behavioral targeting invasive. The differences between our results and theirs can be due to cultural differences (our sample consists mainly of people from Spain) and/or methodological differences – we used experience sampling to capture the context, while the results reported in [29] were gathered via traditional surveys.

2.6 Implications for Design and Future Research

Our study has direct implications on the monetization of personal information online. As the focus of the study has been towards understanding the *economic* aspects of PII, we believe the findings can help in the following future research topics and new offerings. We propose three major implications.

Incentives for adoption of privacy solutions

A prominent reason for the failure of adoption of most privacy solutions are the lack of proper incentives (economic or otherwise) for various parties to support the adoption [31]. Consider online privacy; on one side there are online service providers who have stated that they want to move up to the 'creepy' line [33] on accessing and using PII, while on the other side users are resorting to unilateral measures like anti-tracking plugins etc. to prevent data collection, hence deterring service providers from supporting such privacy preserving measures.

Recent privacy preserving solutions have been designed to preserve privacy of the users as well as provide means for online service providers to access and monetize

[6] http://www.comscoredatamine.com/2010/09/visitor-demographics-to-facebook-com/

PII via targeted ads, thereby preserving the business models of these providers [14, 39]. Based on our findings (Fig. 2.2) these solutions can have a better chance of adoption if they incorporated some form of economic incentives, by way of monetary compensation to the end-user. Such economic incentives based solutions have been proposed as well [27, 34], with some start-ups going for such a model[7].

The results in this paper provide the first empirical foundation for economic incentives by demonstrating how users value different types of PII for a variety of actions performed while online. The prices can be taken to be the reserve prices[8] that users will be willing to accept to part with their PII. Likewise, we have seen that different types of PII have different valuations (e.g. photos in social networks *vs.* online purchase history). These differences can be used by service providers to strategically target different types of PII. The findings in our paper can be used as inputs to drive models to better understand the ecosystem. For instance, a recent proposal to address privacy breaches using insurance can benefit from our analysis to set premiums [15].

In addition, other types of incentives can also help drive adoption of privacy preserving solutions. If we consider Fig. 2.2, users also prefer improved services that use their PII. If service providers can convince users that there have been improvements to the respective services and which PII bits went into the improvements, users may be less concerned about their privacy.

We asked participants of our study about who they would trust to handle their PII in the case that an entity enables economic transactions around their PII, in the post-questionnaire. Users trusted themselves more than any other entity ($\bar{x} = 5.2$, $\tilde{x} = 6$, $q1 = 6$, $q3 = 6$, 6-point scale). Government was the second most trusted entity ($\bar{x} = 3.8$, $\tilde{x} = 4$, $q1 = 2$, $q3 = 5$), followed by banks ($\bar{x} = 3.5$, $\tilde{x} = 4$, $q1 = 3$, $q3 = 4$) and telecommunication companies ($\bar{x} = 3.4$, $\tilde{x} = 3$, $q1 = 2$, $q3 = 4$) tied in the third place ($Z = -.299, p = .77$). Finally, insurance companies were considered the least trustful entities for handling people's PII ($\bar{x} = 3.1$, $\tilde{x} = 3$, $q1 = 2$, $q3 = 4$). Trusting oneself with one's PII could point to a totally decentralized architecture for a privacy solution. However, more work needs to be done to verify if users can undertake the burden of dealing with all the transactions around their PII themselves.

Transparency on monetization of PII

One of the findings reported in Sec. 2.4, is that while users have knowledge of their PII being collected, they are not comfortable about their PII being monetized. This lack of awareness also plays out in valuations – while offline PII and certain types of online PII like photos and financial transactions have high valuations, presence of the user on different sites were valued very low. This is interesting as a behavioral profile can be constructed just by tracking users across sites (via cookies etc.) and this profile can be used to identify users and be monetized [10]. We believe that most privacy concerns that arise are due to lack of awareness of precisely this fact – that PII is being monetized (participants knew their PII could be monetized by entertainment and search related websites, but not for the other categories).

The findings reported in this paper indicate that if online service providers are explicit and up front about the fact that they provide a service (email, video streaming, a social network, etc.) for free and in return collect and monetize PII, along with

[7] www.personal.com

[8] http://en.wikipedia.org/wiki/Reservation_price

details on the specific types of PII they collect, the privacy concerns of most users will be tempered. Long privacy policies written in complicated legalese that are seldom effective [20], can be dispensed with.

For example, we can think about agreements that could expose the amount of money required to run the service the user is signing up for and how the revenues generated by exploiting PII help cover those costs. This implication is further strengthened when we factor in that majority of the users perceive that service providers have higher revenues than costs (Sec. 2.4), hence being transparent about costs can help educate users. Additionally, we can have alternative business models where the user has the *option* to pay for the service that s/he is signing up for either with his/her PII or with real money.

Bulk data mechanism

A final implication for design is related to the indifference in valuation for bulk quantity of data. Specifically, participants assigned a similar value to a certain piece of PII as to 10 pieces of the same information. This has a direct consequence for the design of trading PII. In fact, it does not make sense to implement mechanisms for the trade of a single piece of information. Rather, it makes more sense –according to these results– to design solutions that would allow interested users to trade a bulk amount of PII. For instance, such a mechanism could be presented during registration to a new service and extended for bulk amounts of PII that the user will be sharing throughout the use of the service. The effect of such a design could be two fold: on one hand it would minimize the user's effort and mental load, while on the other hand it would maximize the effectiveness of the service provider's budget expenditure.

2.7 Related Work

Previous research has shown that valuation can depend on the type of information release. For instance, Huberman [17] reported that valuation of certain bits of PII like weight and age depends on the desirability of those bits of information in a social context. Likewise, valuation of location information has been found to depend on factors like the distance traveled by the user and other factors [**cvrcek2006**, 9]. Our work differs in multiple regards. First, we focus on *web browsing information* of users that is of economic interest to online services (*e.g.*, search providers, social networks) and such information raises privacy concerns [24, 26]. Second, we study the effects of demographic information like age, gender, education levels and socio-economic factors on valuation of one's PII. While the aforementioned works used mostly surveys to figure out different valuations, we use a methodology based on experience sampling to capture PII context and obtain valuations in-situ. Finally, whereas previous works used hypothetical payments to determine PII valuation [9], we use actual payments, hoping to obtain a more accurate value and have user engagement.

Another body of work that is related has to do with studying the dichotomy that exists between willingness to pay (WTP) to buy privacy protection and willingness to accept (WTA) to reveal PII. A difference between WTP and WTA can be indicative of an *endowment* effect [38]: people can place a higher value on an object that they own, in this case PII. In our paper we do not deal with WTP vs WTA explicitly, instead we focus on extracting WTA for web-browsing, while leveraging contextual factors when PII is generated and/or released.

A majority of the work done on understanding the awareness levels of users in terms of how their PII is exploited and related privacy concerns has focused on how the actual behavior of people deviates from what they state. This deviation has been noted by Jensen & Potts [21] who also found that there is a difference between reported knowledge and reality; in general people do not seem to know as much about privacy protection measures as they state. They also report that surveys as a method should not be taken as indicative of users' actual behavior. Acquisti studies the reasons that affect people's behavior vis-a-vis privacy and reports bounded rationality as well as the practice of hyperbolic discounting [1]; assigning a higher value to actions involving immediate gratification than those actions leading to long-term protection. In this work, we focus on understanding people's knowledge and perception of how their PII is exploited from an economic view-point, and use experience sampling to capture the behavior and context and as a result, do not suffer from the limitations seen in survey-based studies.

2.8 Conclusions

Our paper deals with the economic value that users assign to PII. Previous literature has focused on different types of PII, but not web-browsing behavior, which is the focus of this work. Previous work has also shown that privacy valuation is a difficult problem, as is affected by a number of technical, legal, social and psychological factors that lead to inconsistencies between what people say and what they actually do. We attempt to overcome these issues through the use of the refined Experience Sampling Method and a truth-telling auction mechanism that incentivizes users to participate honestly.

We found that users give more importance to PII related to their offline identities than to PII that is related to their online behavior. They mostly do not care about the quantity of PII released but they do care about its type. Users tolerate the use of their personal information for improving service, they do not like their information to be used to generate revenues. Users also preferred trading in their PII for money or improved services, and targeted advertising, in this order. We hope the results in this paper can guide future privacy research and solutions.

References

[1] A. Acquisti. "Privacy and Security of Personal Information: Economic Incentives and Technological Solutions". *The Economics of Information Security, J. Camp and R. Lewis (eds)* (2004), pp. 165–178.

[2] A. Acquisti and J. Grossklags. "Privacy and Rationality in Individual Decision Making". *Security & Privacy, IEEE* 3 (2005).

[3] A. Acquisti and L. John. *What is privacy worth.* Tech. rep. Heinz College, Carnegie Mellon University, 2009.

[4] J. Angwin and S. Stecklow. *Scrapers dig deep for data on Web.* http://on.wsj.com/P17bnx. Oct. 2010.

[5] L. F. Barrett and D. J. Barrett. “An Introduction to Computerized Experience Sampling in Psychology”. *Social Science Computer Review* 19.2 (Summer 2001), pp. 175–185.

[6] A. Braunstein, L. Granka, and J. Staddon. “Indirect Content Privacy Surveys: Measuring Privacy Without Asking About It”. *Symposium on Usable Privacy and Security (SOUPS)*. 2011.

[7] M. Cherubini and N. Oliver. “A Refined Experience Sampling Method to Capture Mobile User Experience”. *Workshop of Mobile User Experience Research part of CHI'2009*. 2009.

[8] S. Consolvo and M. Walker. “Using the experience sampling method to evaluate ubicomp applications”. *IEEE Pervasive Computing* 2.2 (Apr. 2003), pp. 24–31.

[9] G. Danezis, S. Lewis, and R. Anderson. “How Much is Location Privacy Worth?” *Online Proceedings of the Workshop on the Economics of Information Security Series (WEIS) 2005*. 2005.

[10] C. Dwyer. “Behavioral Targeting: A Case Study of Consumer Tracking on Levis.com”. *Risk Analysis* (2009).

[11] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay. “MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones”. *ACM MobiSys '07*. 2007.

[12] S. Goel, A. Broder, E. Gabrilovich, and B. Pang. “Anatomy of the Long Tail: Ordinary People with Extraordinary Tastes”. *in WSDM*. 2010.

[13] J. Grossklags and A. Acquisti. “When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information”. *Workshop on Economics of Information Security*. Citeseer, 2007.

[14] S. Guha, B. Cheng, and P. Francis. “Privad: Practical Privacy in Online Advertising”. *NSDI*. 2011.

[15] S. Guha and S. Kandula. “Act for affordable data care”. *HotNets*. 2012, pp. 103–108.

[16] S. E. Hormuth. “The sampling of experiences in situ”. *Journal of Personality* 54.1 (1986), pp. 262–293.

[17] B. Huberman, E. Adar, and L. Fine. “Valuating privacy”. *Security & Privacy, IEEE* 3.5 (2005), pp. 22–25.

[18] G. Iachello, K. N. Truong, G. D. Abowd, G. R. Hayes, and M. Stevens. “Prototyping and sampling experience to evaluate ubiquitous computing privacy in the real world”. *ACM SIGCHI CHI '06* (2006).

[19] S. S. Intille, J. Rondoni, C. Kukla, I. Ancona, and L. Bao. “A context-aware experience sampling tool”. *CHI '03 extended abstracts on Human factors in computing systems*. CHI EA '03. Ft. Lauderdale, Florida, USA: ACM, 2003, pp. 972–973.

[20] C. Jensen and C. Potts. “Privacy policies as decision-making tools: an evaluation of online privacy notices”. *ACM SIGCHI CHI*. 2004.

[21] C. Jensen, C. Potts, and C. Jensen. “Privacy practices of Internet users: Self-reports versus observed behavior”. *International Journal of Human-Computer Studies* 63.1-2 (2005), pp. 203–227.

[22] N. Jentsch, S. Preibusch, and A. Harasser. “Study on monetising privacy: An economic model for pricing personal information”. *ENISA* (Feb. 2012).

[23] “Studies in resource allocation processes”. Ed. by J. Kenneth and L. Hurwicz. Cambridge University Press, 1978. Chap. On informationally decentralized systems.

[24] B. Krishnamurthy, D. Malandrino, and C. E. Wills. “Measuring privacy loss and the impact of privacy protection in web browsing”. *SOUPS*. 2007.

[25] B. Krishnamurthy, K. Naryshkin, and W. Craig. “Privacy leakage vs. Protection measures: the growing disconnect”. *WISP* (May,2011).

[26] B. Krishnamurthy and C. Wills. “Privacy diffusion on the web: a longitudinal perspective”. *WWW 2009*. 2009.

[27] K. C. Laudon. “Markets and Privacy”. *ICIS*. 1993.

[28] C. Mancini, K. Thomas, Y. Rogers, B. Price, L. Jedrzejczyk, A. Bandara, A. Joinson, and B. Nuseibeh. “From spaces to places: emerging contexts in mobile privacy”. *Proceedings of the 11th international conference on Ubiquitous computing*. ACM. 2009, pp. 1–10.

[29] A. M. McDonald and L. F. Cranor. “Americans’ attitudes about internet behavioral advertising practices”. *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society*. 2010.

[30] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. “Detecting price and search discrimination on the internet”. *HotNets*. 2012.

[31] A. Narayanan, V. Toubiana, S. Barocas, H. Nissenbaum, and D. Boneh. “A Critical Look at Decentralized Personal Data Architectures”. *CoRR* abs/ 1202.4503 (2012).

[32] C. Paine, U.-D. Reips, S. Stieger, A. Joinson, and T. Buchanan. “Internet users’ perceptions of ’privacy concerns’ and ’privacy actions’”. *Int. J. Hum.-Comput. Stud.* 65 (6 June 2007), pp. 526–536.

[33] S. Richmond. *Eric Schmidt: Google gets close to ‘the creepy line’*. http://tgr.ph/dp8BaH. Oct. 2010.

[34] C. Riederer, V. Erramilli, A. Chaintreau, P. Rodriguez, and B. Krishnamurty. “For sale: Your data, By: you”. *ACM Hotnets*. 2011.

[35] F. Roesner, T. Kohno, and D. Wetherall. “Detecting and defending against third-party tracking on the web”. *NSDI’12: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. 2012, pp. 12–12.

[36] E. Steel. *A web pioneer profiles users by name*. http://on.wsj.com/riy5Qo. Oct. 2010.

[37] E. Steel and G. A. Fowler. *Facebook in Privacy Breach*. http://on.wsj.com/buDn6d. Oct. 2010.

[38] R. Thaler. “Toward a positive theory of consumer choice”. *Journal of Economic Behavior & Organization* 1.1 (1980), pp. 39–60.

[39] V. Toubiana, H. Nissenbaum, A. Narayanan, S. Barocas, and D. Boneh. “Adnostic: Privacy Preserving Targeted Advertising”. *NDSS*. 2009.

[40] J. Valentino-Devries and E. Steel. *White House to Push Privacy Bill.* http://on.wsj.com/hlSR7E. Mar. 2011.

[41] WSJ. *WSJ: What they know series.* http://online.wsj.com/public/page/what-they-know-digital-privacy.html. Apr. 2012.

Part II

Valuation of personal information in mobile communication: a longitudinal study on mobile phone call annotation

The next two chapters present the "CallNotes" research project, a longitudinal, multiple-phase study on several aspects of phone call annotation that lasted almost two years. The CallNotes Project comprised three phases:

1. A study of phone call annotation behaviours, including phone recall needs; common artefacts used by mobile phone users to take notes or to remember information from their calls; and factors that influence note-taking in mobile calling circumstances. This first phase of the study resulted in a short paper [3] that was published in the Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI2012), and is included in this dissertation in the Annex 1.

2. The second phase of the project allowed us to obtain insights on the dynamic aspects of mobile phone users' call annotation needs and behaviour. A paper was accepted for publication in the Transactions of Computer-Human Interaction (ToCHI) journal:

   J. P. Carrascal, R. de Oliveira, and M. Cherubini. "To Call or to Recall? That's the Research Question". *Transactions of Computer-Human Interaction (ToCHI) (Accepted for publication)* (2015)

   This paper also expanded of the first phase of the CallNotes project, thanks to a larger dataset that allowed to improve statistical significance of the results. Additionally, the results of this paper were compiled and leveraged to file as a patent that was granted in July of 2014. It is also included in the Annex 1.

3. The third phase of the CallNotes project conducted a user-centered comparative evaluation of three annotation techniques: original notes as determined by the participants of the study; notes selected by external or context-free annotators; and notes obtained by means of a machine-learning algorithm, that was created on the basis of the findings of previous phases of the project. The results of this phase were written down in a paper that was submitted to the ACM SigCHI2015 conference:

   J. P. Carrascal, F. Bonin, S. P. Jose, R. De Oliveira, and N. Olivier. "Who can we trust? A Comparative Evaluation of Phone Call Annotation Techniques". *ACM SigCHI Conference (Submitted).* 2015

The aforementioned publications are presented next, as Chapters 3 and 4.

Chapter **3**

To call or to recall? That's the research question

Juan Pablo Carrascal, Universitat Pompeu Fabra
Rodrigo de Oliveira, Telefónica Research
Mauro Cherubini, Telefónica Research

Abstract

We present findings of a study with 62 subjects who had 796 of their outgoing mobile phone calls recorded and transcribed for their later annotation—by highlighting important information shared during calls. We found that patterns in these calls (numbers, names, interrogative adverbs) as well as some contextual parameters are better indicators of annotation needs than the callers' profile or call quality. Callers highlight information in both parties' turns (caller and callee) more often than highlighting solely information provided by the callee, which is mostly due to annotating questions with contextual information for the highlights in the callee's turns. We discuss how this behavior changes according to call purpose. Finally, we found that annotation needs change over time: while some annotations might not be considered relevant after weeks, others originally considered irrelevant might become important archival notes. We present implications of these findings for the design of mobile phone annotation tools.

3.1 Introduction

Mobile phones are one of the most pervasive personal electronic devices ever made. With a number of mobile subscriptions that approaches the 7 billion, [2], these devices are used for various purposes, such as keeping in touch with friends, doing business, or even for emergency situations. Moreover, making mobile phone calls has been part of people's day-to-day life in the past couple of decades. A lot of information is exchanged through phone calls. While a consistent part of this information could be ephemeral as supporting our social needs, another part might be worth remembering as being functional for our lives. For example, we might make a phone call to confirm whether our partner is feeling better, to inquiry which groceries to buy on our way back home, to discuss a specific topic with a work colleague, or to double check the exact time and place to meet a friend during the weekend. While the first example deals with information that is consumed immediately and does not necessarily require notes to be taken, the other ones suggest the need to archive information for future use.

To date however, there is little support of applications for annotating these information appropriately. Previous work has focused mostly on investigating how people take notes during work-related meetings [4, 34, 12, 35] and in other general settings [20, 3, 24]. Indeed mobile phone calls could be seen as two-person meetings, but they are quite different from formal meeting contexts: calls are usually less planned, less structured, shorter in time, and performed in mobile contexts. Aside from these differences, it is often experienced by many the cumbersome situation of having hands busy and needing to take notes while on the phone, a restriction typically not experienced during work meetings. Perhaps the scarce number of note-taking solutions for mobile phone calls is due to the lack of studies that have tackled this issue. In consequence, we have little evidence of what is important to remember during mobile conversations, as well as which factors play a role in this process. Furthermore, it is unknown how phone call annotation needs change over time and which actions should be taken in order to support such dynamic needs.

In a previous work, we disclosed initial findings related to some of these topics, which pointed to the importance of studying more adapted forms of phone call annotation [6]. In this paper we expand our previous one tackling all topics mentioned before and using a larger dataset with increased statistical power (CL: 95%, MOE: $\pm 3.4\%$). More specifically, we discuss results obtained with a 2-month user study conducted with 62 subjects who had their outgoing phone calls transcribed for annotation by means of highlighting. The analysis of the participants' calls ($N = 796$) and highlighting behavior confirms that there are specific call patterns that are frequently annotated, and that some contextual variables might influence the note-taking behavior. We observed that callers surprisingly highlight only information on their turns as often as highlighting only information on the callee's turns. We also observed that people's annotation needs and strategies do change across time, and these dynamics should be taken into account when designing mobile phone call annotation tools.

The main contributions of this work include our findings on:

1. People's needs to annotate mobile phone conversations, as supported by a survey conducted with 62 subjects;

2. Factors that seem to influence the note-taking activity, as supported by the analysis of 2-month user study data in which participants (i) highlighted parts of the transcribed mobile phone calls that they considered to be important, and (ii) provided contextual details for each call by answering a post-call questionnaire.

3. The frequency and circumstances that each party in a phone call—caller and callee—provide information considered to be relevant for annotation;

4. Influence of time on annotation needs, which was investigated by a follow-up study phase enabling participants to (re)annotate—if needed—a subset of their calls;

5. Derived recommendations for the design of note-taking support tools for daily mobile phone conversations.

We expect our work to provide new insights in these topics and to stimulate the design of memory prosthetic support for mobile phone calls.

3.2 Related Work

The act of taking notes is an activity people frequently perform to record information from other sources for later use. Geyer and colleagues [12] state that "personal notes primarily serve as a memory aid for individuals to remember important facts, actions, ideas, and decisions but are hardly useful for persons other than the author". Annotations may be as simple as highlighting passages or producing a new written product. In the first case, certain fragments of information from the original piece are marked as they are considered important. In the second, new written pieces are composed on the basis of the original information. These freetext annotations might provide additional information beyond the original piece that can enrich the information, e.g. for providing context or indicating on what to do with the information in the future. To provide similar functionality, text highlighting can be enriched by certain behaviors, such as highlighting contextual information beyond the specific information to be remembered (e.g. instead of highlighting only the address of a meeting point, the annotator can also highlight the question related to the address).

In both note-taking strategies, there is the need to preserve certain pieces of information for later access and use.

It is often the case one has to take notes in cooperative situations, such as in meetings, lectures or phone calls. Related research in latest years has focused primarily on studying *needs* for annotation in work-related meetings. In this setting, notes have a structured form and usually include action items. Often these notes are recorded to create a shared group memory and to make the meeting more efficient [12]. Alternatively, notes can be used as memory cues for participants to recall events of a meeting rather than being full recordings of the activity [34]. Moreover, these notes can serve as markers to add structure to meeting recordings [33, 12, 35, 4]. In both cases, attention and active participation is required, and taking notes at the same time may become an additional cognitive load [26] that reduces the person's ability to participate [34].

These needs have been lately supported by various *artifacts*, including electronic annotation tools that leverage desktop computers [20, 3] or mobile devices [14, 19, 10, 13], as well as the common paper and pen approach, still frequently used in the form of post-it notes, miscellaneous text files, or the corner of other printed documents [3, 34].

The work presented herein focuses on annotations of daily mobile phone calls, that is, parts of a phone conversation that participants of a phone call would consider worth preserving for later use. Phone calls are different from work-related gatherings in a number of ways. For example, typical mobile phone calls tend to be relatively shorter, they are frequently not planned beforehand and they lack the structure of a meeting, relaying instead on a series of salutations and informal dialogs. In consequence, the pieces of information that the participant of a phone conversation may need to remember—to annotate—and his/her motivations to take notes, differ from those of a participant of a work meeting. Also, it has been observed that during phone calls, participants often have their hands busy, either by performing another activity (e.g. driving) or by holding the phone, documents or other objects [25]. Despite these differences, annotation of mobile phone calls has received little coverage in current research literature.

These observations support the need to study note-taking habits explicitly in contexts of informal mobile phone calls. As an initial step of this work, we aimed at understanding current needs and artifacts for annotating mobile phone calls as stated by our first research question:

> *RQ1: What needs and artifacts do people have to annotate information exchanged in mobile phone calls?*

Researchers have also dedicated to understand *influencing factors* when annotating information, particularly in work meetings. As Lin and colleagues [24] previously identified, the first step in the lifecycle of a note is the need to annotate something. Studies conducted by [5, 4] revealed that demographics and contextual information might be significant influencing factors of note-taking in work meetings. According to their findings, women tend to take more notes and more often than men. They also observed that older participants take more notes than younger ones. Additionally, the role of the participant in the meeting was found to have a direct influence on the amount of notes taken: meeting leaders talk more and hence have less opportunity to take notes, while project managers have to take more notes because of their responsibility to produce meeting minutes.

Note-taking influencing factors have also been studied in the reverse order, i.e., by looking at annotation occurrence in order to infer other pieces of information. For example, [4] explored the relationship between topics in meetings and the annotating behavior of their participants, and suggested that the presence of annotations could have some predictive power to estimate when something important is about to be discussed.

However, there is little evidence that the aforementioned findings can be generalized to annotations of mobile phone calls. The second goal of our research is therefore stated as:

> *RQ2. Which factors mostly influence the need for creating annotations during mobile phone calls?*

The information we take notes of is usually the information that we consider to be the most important to be preserved for later use. Understanding how often each party—caller and callee—can be the source of such important information is key to simplifying automatic approaches for mobile phone call annotation. In order to investigate this, we leveraged concepts from Conversation Analysis, a research area that has recently focused on studying communication in mobile phone calls [1, 16, 22, 31] besides its frequent studies on landline phone calls [29, 32, 23]. We used the concept of turn [27] when analyzing our dataset to investigate whether callers annotate information shared by the callee more often than they annotate information shared by themselves. We aimed to understand this, both on a general level as well as in relation to contextual variables—such as relationship with the callee, location and companion at the time of the call, and objective of the call—. Hence, our third research question is:

> *RQ3. How often and in which circumstances does each mobile phone party—caller and callee—provide information that the caller considers to be worth remembering?*

For simplicity, we split this broad research question into the following ones:

> *RQ3.1 Do callers annotate—by means of highlighting mobile phone call transcripts—information shared by the callee significantly more often than information shared by themselves?*
> *RQ3.2 Which factors mostly influence callers to highlight information in the turns of the two parties involved in a call (i.e. caller's turns, callee's turns, or both parties' turns)?*

Finally, scholars have looked at temporal effects on people's annotation needs in contexts of meetings and lectures. As people's memory recall abilities diminish with time, a number of tools for aiding their memory have been applied, thus composing their prosthetic—rather than the natural organic—memory [18]. In the lab study conducted by [18], participants' prosthetic memory supported by paper and pen notes were shown to have high recall value and retrieval efficiency in the short term. However, accuracy decreases rapidly to the point of becoming useless after one month when compared to organic memory. This tradeoff between recall accuracy and retrieval efficiency has been also previously observed by [33].

As time passes, actions taken over notes vary. After initial information consumption is completed, notes are either discarded or archived [24]. Strategies used to perform these actions vary [20]. [34] found that most annotators access their personal records of meetings afterwards, and 75% access them frequently. The same work found that meeting participants try to keep notes accurate, and half of them even rewrite their notes. In terms of archiving, 75% of the participants keep their notes for a year on average.

These findings are mostly related to the context of work meetings and hence might not be generalized for annotation of informal mobile phone calls. We therefore state our fourth research question as:

> *RQ4. Does the annotation behavior—by means of highlighting mobile phone call transcripts—change over time?*

In case we find the answer to this question to be affirmative, additional questions should also be investigated in order to better understand such change:

RQ4.1 Does the topic and amount of information in highlights change across time?
RQ4.2 How does this information change—if any—unfolds across time?
RQ4.3 Does the importance of the highlighted information change across time?
RQ4.4 What are the reasons presented by participants for the change in their highlighting behavior?
RQ4.5 How can annotation behavior change—if any—be explained in terms of the previously investigated factors?

The rest of the paper is organized as follows: in Section 3.3 we explain the design and methodology of the study. In Section 3.4 we address each research question separately, presenting relevant results and discussing them in detail. Finally, in Section 3.5 we present a set of implications for the design of a mobile phone application to highlight important information shared during mobile phone conversations.

3.3 Methodology

We deployed a user study consisting of two phases. The first phase ($P1$), which spanned over 64 days, allowed us to collect a large sample of outgoing[1] mobile phone calls, the highlights of important information inside them—if any—, and contextual parameters at the time of the calls. The second phase ($P2$) was based on questionnaires that produced a second round of highlighting on a subset of the same calls, allowing us to understand the effects of time on the highlighting behavior of the participants.

Participants

A total of 62 subjects (20 female), with a mean age of 31.5 years ($s = 7.52$, $min = 20$, $max = 51$) participated actively in the user study by answering the pre-study questionnaire and contributing at least one mobile phone call. All of them were residents of Spain and reported being fluent in the Spanish language—a requirement of the study. The sample was geographically well distributed (38 unique cities), and included only subjects that had received basic education at least (primary school: 3.2%), followed by 3.2% who finished secondary school, 79% that concluded technical school or obtained a bachelor degree, and 14.5% who had either a masters or a doctorate degree. The reported annual income suggests that all social classes were represented in the sample (27.4%, 19.4%, 25.8%, 19.4%, and 8.1% earned up to €10K, €20K, €30K, €40K, and more than €40K a year respectively).

Procedure

Participants were recruited amongst people who voluntarily registered after following advertisements in popular Web portals in Spain. We opted for asking participants to install a specific VoIP application in their smartphones to enable recording and transcription of their calls for later analysis. The application was available for Android and iPhone platforms only. Candidates who owned a mobile phone with either

[1] We considered only outgoing calls given that these are intentional with a clear user need, thus reducing the number of random and/or undesired calls in our sample.

of these platforms were invited via email to be part of the study, and asked to answer an online pre-study questionnaire. Besides collecting general demographics, the questionnaire also inquired participants about their calling habits and general note-taking habits during phone calls. We then conducted a user study divided in two phases, as explained in the following section.

Study Phase One ($P1$)

The goal of this phase was to gather a large sample of mobile phone calls, including their audio, full transcription, contextual information, and textual fragments of each call transcription that the respective caller would consider important or worth to be remembered in the future. This data was used to address research questions $RQ1$ and $RQ2$.

Phone Call Setup: During $P1$, we offered participants free calls to mobile or fixed phone lines inside the Spanish territory. To be able to make free calls, they had to install a VoIP application on their mobile phones and configure it to connect through our servers. Once the application was installed, making mobile phone calls through the application was transparent for users: whenever a call was dialed using the native phone keypad, it was automatically redirected through the VoIP system.

Privacy and Information Security: We explained participants how to deactivate the application, so as to prevent us from having access to more private phone conversations. To further comply with privacy protection laws, whenever participants placed a call through our system, a short message was played to both the caller and the callee informing them that the call was going to be recorded and transcribed.

Call Transcription: Since it was out of the scope of our study to work on or improve the state of art in speech-to-text, we hired an external service provider to transcribe the phone conversation recordings into text. Transcriptions were first generated by the provider automatically—by means of a Hidden Markov Models-based method [17]—and later manually inspected and corrected by a human expert before being presented to participants. The resulting transcription for each call was an annotated file that contained the text of the whole call and the adequate labels for both caller's and callee's turns. Besides helping to identify each party's turns, we used the labels as a separator for parsing the transcriptions. This way, by using a simple regular expression, we could divide the call into individual turns for quantitative analyses. Example 1 shows a fragment of a transcription as they were provided by the transcription service[2]:

Example 1:

```
         (...)
CALLER:  Hey...  where are you?
CALLEE:  Still here in Fnac.
CALLER:  Ok, we are down here, by the door.  Near the tree,
         in Plaza de Callao.
CALLEE:  The tree?  Is there a tree there?  Well, I'm here
         still undecided, but I'm going down, but we probably
         have to pay a book that Irene bought, so...
CALLER:  Ok, then come to Callao when you are out, to the
         Christmas House.
CALLEE:  Ok, good, we'll go as soon as possible.  See you
         later.
CALLER:  See you.
```

[2]Call transcription examples were translated from Spanish (their original language) to English by the authors.

Annotations in the form of highlights: Annotations are usually studied in the form of spontaneous, free-text *note-taking*, which involves a high cognitive load and effort for synthesizing the original information and composing new pieces of derived text [26]. Another form of annotation is that of *highlighting* text fragments, which is consistent with the *offline/explicit* indexing strategy explained by [12] for marking meaningful points in meeting records. Both cases have the same goal: to preserve pieces of important information for later use. On this basis, we opted for studying annotations in the form of highlights rather than free-text generated by the note-taker due to a couple of reasons. First, generating free-text form annotations for each phone call would demand higher effort from the participants, which could lead to negligent behavior when deciding whether a call should be annotated or not. And second, collecting useful fragments from calls would allow us to later investigate whether annotations can be generated automatically at an acceptable error rate.

Call Questionnaire: We developed a web application which allowed participants to interact with call data. It displayed a list of their calls, and for every call, its transcription, date, duration, status and callee number, as well as a related *call questionnaire*. Participants were free to complete this questionnaire whenever they wanted for the duration of $P1$. They were also allowed to delete[3] a call within a 24-hour period if they considered it to have sensitive content. In the call questionnaire, we gave participants the following task for all of their transcribed calls: *"Highlight the parts that you consider to be important and that you would like to remember for later*[4]". We therefore asked participants to regard information as *important* when it was worth of being preserved to be used at a later time. In cases when they did not find any important text to highlight, we asked them to explicitly declare so. Participants received a monetary incentive for answering call questionnaires regardless of highlighting anything in the call transcript or not. We expect that this approach motivated participants to provide us with more detailed information about their calls without biasing their highlighting behavior. Finally participants were presented a series of contextual questions related to each phone call: 1) Relationship with callee; 2) Who was with the caller at the time of the call; 3) Location of the caller at the time of the call; 4) Objective of the call; 5) Level of importance of the call; 6) Level of importance of the notes; and 7) General questions about sound and transcription quality. Given that participants reported objectives of calls in free text, we needed to develop a coding scheme to further analyze this variable. In order to minimize the investigator bias, we asked a colleague researcher who was not aware of this study to inspect the reported objectives and to suggest a coding scheme. He then met with one of the authors and they agreed on a final coding scheme which included the following categories: "discuss a topic", "ask and/or receive specific information", "set an appointment", "social" and "other". They both then proceeded to classify the answers separately. Inter-rater reliability was tested using Cohen's Kappa, and it was found to be highly acceptable ($K = .86, p < .001$).

Analysis of highlighted turns. We leveraged some concepts from the Conversation Analysis area to study the call transcriptions, in order to obtain insights on how frequently important information is highlighted among the contributions of the caller and the callee. A phone call conversation is composed by several alternating interactions contributed by the two involved parties. Each of these interactions is called a *turn* [27]. They can be considered as the basic building blocks of any conversation. Consecutive turns constitute units called sequences. For instance, Example 2 com-

[3]Participants deleted a total of 65 calls recorded during the study, which represents 7.5% of the final working dataset ($N = 796$ calls).

[4]This is the exact phrasing that we used, as translated from the original Spanish version: "*resalta las partes que consideres importantes y que quisieras recordar para después*".

prises a sequence of four turns extracted from one of the conversations of our dataset (text in bold indicates the highlights made by the caller):

Example 2:

4		(...)
5	CALLER:	**Have you bought something for mom already?**
6	CALLEE:	No, I was going to do it, but I don't know... after the next week or the other one.
7	CALLER:	Did you have something in mind already?
8	CALLEE:	As I told you **I thought about buying her a coat** because she once said she would like to have one.
9		(...)

Therefore we first selected the calls that contained at least one highlight ($N = 299$) and parsed their corresponding transcriptions in order to extract the turns. Afterwards, for each call, we counted the number of *highlighted turns*, i.e., the turns containing at least one highlighted piece of information. It is important to note that we asked participants to highlight the pieces of text from the transcription that they considered to be important or worth remembering. This is our definition of *important information.* With this in mind, we consider that if a turn contains a piece of highlighted information, the turn itself contains important information, i.e. it is a *highlighted turn.* Having counted highlighted turns for both parties in every call, we determined whether: 1) the caller highlighted information only in his own turns; 2) the caller highlighted information only in the callee's turns; or 3) the caller highlighted information in both parties' turns, i.e. the caller highlighted some information that s/he said during the call, *as well as* some other information—not necessarily the same information—that the callee said during the call. Every call was therefore classified into one of these three groups. To further explain these groups, let us reconsider Example 2 that has turns number 5 and number 8 highlighted. Given that the former is a caller's turn (turn number 5) and the latter is a callee's turn, we conclude that the caller highlighted information in *both parties' turns* for that specific call. In Example 3 though, all of the highlighted turns[5] are caller's turns, and hence we conclude that the caller highlighted information *only in his/her turns* for the given call.

Example 3:

7		(...)
8	**CALLER:**	**Listen, I am calling to tell you that I am going to Burgos tomorrow in the morning.**
9	CALLEE:	Really? Why?
10	**CALLER:**	**Well because Diego will leave no sooner than six and I'll leave tomorrow at twelve so I have bought the one o'clock ticket.**
11	CALLEE:	Good, Ok. So will you be home for lunch?
12	CALLER:	I don't know for lunch, because one plus three is four... You eat, if you can leave something for me, then great.
13		(...)

After manually inspecting the highlighted turns, we observed that many of them contained questions, and that they mainly provide context to the actual information sought by the person who asked. As stated by [28], interrogative sentences *"signal the*

[5] Although Examples 3 shows only a fragment of the call, no other turns, besides those highlighted in bold, were highlighted in the entire call.

desire of the speaker to gain information from the addressee". Given this main role of questions, they were highlighted by participants most likely in order to provide contextual information for framing a forthcoming answer. For instance, in Example 2, the highlighted question in Turn 5 (*"Have you bought something for mom already?"*), provides context to the information sought by the caller and hence highlighted in Turn 8 (*"I thought about buying her a coat"*). Although both Turn 5 and Turn 8 were highlighted, we consider that the callee (Turn 8) provided relevant information while the caller's turn (Turn 5) was highlighted mostly to provide context. Therefore, in a similar way as we studied the frequency of highlights in call parties' turns, we also studied how frequently *questions* were highlighted among the caller's, callee's, and both parties' turns[6]. This helped us to understand who is the party (either the caller or the callee) who provides more questions—the main information seeker—and thus, in those cases, who provides more important contextual information. To round up this analysis, we also analyzed the highlighted turns excluding questions, in order to determine how frequently highlighted information, excluding contextual information contained in questions, appears in each parties' turns. Finally, we searched for correlations between the presence of highlighted in-formation—in each and in both parties' turns—, and the contextual variables obtained during the study. All the aforementioned measures were used for addressing $RQ3$.

Study Phase Two ($P2$)

From $P1$, we obtained an initial data set to understand the highlighting needs of participants in a time close to their calls. The aim of $P2$ was to extend this data with a second round of highlights that participants could take further in time from calls, which would allow us to study how annotation needs change over time. This data was used to address $RQ4$.

In order to collect such data, we invited participants to fill a questionnaire—similar to the call questionnaire in $P1$—about the calls they made during the first phase of the study. Phase two was not mandatory and participants of $P1$ received a monetary incentive to also participate in $P2$. Probing participants on every call they made in $P1$ would reduce reliability of our results due to survey response fatigue, given that each subject made 13 calls in average during $P1$. We hence decided to restrict participants' questionnaires to a subset of their calls. In a pilot conducted before $P2$, test subjects—not related to the study sample—answered the questionnaire for up to six calls without any fatigue-related complaint. Therefore, we opted for selecting a maximum of six calls per participant: three calls that s/he previously highlighted in $P1$ (if any), and three calls that s/he did not highlight in $P1$ (if any). The selection of the six calls had to be representative about each user's participation in the entire Study Phase 1. In order to achieve this goal, we ordered each participant's *highlighted* calls according to their annotation date and time, and then selected his/her first call, mid-point call, and last call. The same process was applied for *non-highlighted* calls towards selecting a maximum of six calls per participant. By following this procedure, we obtained a more representative set of calls per participant covering their entire experience in $P1$ as evenly as possible. Furthermore, this procedure balanced the amount of data points across time, thus maximizing statistical power for the study of research question $RQ4$.

The $P2$ questionnaire look and feel was similar to the one used in $P1$. For each of the selected calls, we again offered participants the opportunity to highlight the

[6] In order to detect questions, we first manually reviewed the dataset to verify the correct presence of question marks (in spanish ¿ and ?) in turns representing questions, and then parsed the transcriptions.

pieces of information from the transcription that they considered relevant. It is worth remarking that we were not measuring recall, as participants could highlight whatever information in $P2$ they considered to be important at that particular time. Up to this point, we did not present initial highlights they made in $P1$, if any. If participants considered that call transcriptions in $P2$ had no important information, they were required to inform that. Next, participants were asked about the level of importance of the call, reasons for highlighting it (if applicable), and level of importance of the notes.

Once they finished this step, participants were presented with a dynamic section, which could branch into one of the following:

1. *Call was highlighted both in $P1$ and $P2$:* Both sets of highlights were presented. Participants were asked if the highlights made during $P2$ were less important, equally important, or more important than those made in $P1$, and to explain why.

2. *Call was highlighted in $P1$ but not in $P2$:* Original highlights were presented and participants were asked about the reason for not highlighting anything in $P2$.

3. *Call was not highlighted in $P1$ but it was highlighted in $P2$:* Recent highlights made in $P2$ were presented and participants were asked about the reason for not having highlighted anything in $P1$.

4. *Call was not highlighted in either $P1$ or $P2$:* No further questions were asked.

Participants' reasons for highlighting different information in $P1$ and $P2$ were coded using the same procedure described for coding the objectives of calls. The categories used were "temporal effect", "information refinement", "archiving", "same content" and "no clear answer". Inter-rater reliability for this coding scheme was tested and evaluated as highly acceptable ($K = .85, p < .001$).

Finally, we aimed to quantify thematic differences between notes taken in both phases. In order to achieve this, two coders inspected all the highlights in $P1$ and $P2$ and agreed on classifying differences in four categories:

- *No change of information*: Both notes dealt with the same topic and conveyed the same amount of information.

- *Information increase*: Both notes dealt with the same topic. Notes from $P2$ conveyed more information than notes from $P1$, *e.g.* *"Where are you", "I am in the garage"* vs. *"I am in the garage because I had to get some oil and it took some time"*.

- *Information decrease*: Both notes dealt with the same topic. Notes from $P2$ conveyed less information than notes from $P1$, *e.g.* *"I'm going to take it back / the T-shirt / Don Jaime street / I will wait for you at <supermarket name>"* vs. *"at <supermarket name>"*.

- *Different topic*: The topic of the notes changed from $P1$ to $P2$, thus the information they contain could not be compared. *e.g. "Let's meet tomorrow at five / at Sevilla on February the 19th"* vs. *"I work from home on Friday / call you later"*.

The two coders applied this classification scheme, with a highly acceptable inter-rater reliability ($K = .81, p < .001$).

Data Preprocessing

Before analyzing data collected in both phases, we first reviewed all of the participants' comments and objectives of calls reported in the $P1$ call questionnaires. From this initial analysis, we observed that most of the participants' first calls were justified as being test calls. These calls were removed from our data. Furthermore, calls that could not be properly transcribed were also filtered out. The sample reported herein (796 calls and 62 participants) is therefore our final working dataset after applying these filtering heuristics.

3.4 Results and Discussion

During the study deployment, participants made 796 calls with a total duration time of $141,741$ seconds ($\bar{x} = 178.07$; $s = 364.55$). Quality of the calls was considered acceptable ($\tilde{x} = 3$: acceptable, $Q1 = 3$: acceptable, $Q3 = 4$: good). Transcriptions of the calls and highlights yielded a total of $1,241,956$ characters ($\bar{x} = 1,560.25$; $s = 3,094$) and $49,382$ characters ($\bar{x} = 62.04$; $s = 198.01$) respectively. The average number of calls per participant was 12.84 ($s = 11.57, min = 1, max = 49$), and they highlighted an average of 4.92 of their calls ($s = 5.78, min = 0, max = 30$). Hence, 37.6% of all phone calls were annotated by highlighting, which is consistent with the participants' self-reported annotation habits captured by the pre-study questionnaire (34% and 45% indicated taking notes frequently using paper/pencil and their mobile phone respectively). Likewise, participants called family members more often than friends, and called friends more often than work colleagues, which reveals the same order reported in the pre-study questionnaire. These findings support consistency between the participants' behavior in the study and how they perceive their behavior in real life. Next we present and discuss results related to each of the research questions described in the Related Work section.

Addressing RQ1: Phone recall needs and artifacts

According to what participants reported in the pre-study questionnaire, recalling information from call conversations is a frequent need and not necessarily an easy task. Almost half of the sample agreed that this need occurs sometimes (47%), and over one third indicated it happens frequently (37%), while only 16% of the participants reported that this need rarely occurs. No one reported the absence of this need. When evaluating the easiness to recall information obtained in phone calls, 39% said it is either easy or very easy, 35% reported it is neither easy nor difficult, and the remaining 26% agreed the recall task is at least difficult. These results suggest the importance of supporting recall of phone conversations.

In addition, we observed that only 4% of the characters in the transcribed calls were highlighted by the participants. This indicates that full transcripts of conversations is not an optimal solution for phone recall, as they would most likely overload users switching their problem from recall to information retrieval. As [21] points out, this is one of the major problems in personal information management, which could be targeted by mobile applications. An optimal approach should include the recognition of annotation patterns together with contextual information of the call so that the user does not have to go through the entire transcript to retrieve the important pieces of information. Then the stored contextual information—i.e. metadata—can be used in order to optimize retrieval. This is in agreement with [34] on their suggestions for work meeting capturing tools.

With respect to the main artifacts used for recall, mobile phones and regular paper and pencil were reported as the most important ones–45% and 34% respectively use them frequently for this task. More specifically, participants reported taking offline notes of phone conversations using text-based notepads and audio-based memo applications (24% record audio notes for phone calls at least once a week).

It is worth noting that our sample is composed of smartphone users only, hence the popularity of mobile phones as the primary annotation source. Nevertheless, our results indicate that: (1) the majority of this population segment often have needs to recall information exchanged in daily phone calls; and (2) smartphones are becoming relevant tools for annotating these conversations. In fact, smartphones seem to be the most convenient devices for that purpose as most commonly reported annotation needs can be handled by a specific mobile application: 33.9% reported usually taking notes of phone calls to remember calendar-related information (dates and appointments), 33.9% said call notes are useful to remember contact information (phone numbers, names and emails), and 22.3% wanted to remember to-do items (e.g. shopping lists).

Addressing RQ2: Variables that influence call annotation

We looked into a number of variables that could potentially reveal the likelihood for one to highlight information in any given phone call. We collected and analyzed four types of variables:

- *Patterns and call related variables:* In the preliminary questionnaire, participants reported the kinds of information they find themselves trying to remember after a phone call. About 81% mentioned pieces of information that necessarily include numbers (e.g., phone numbers, dates, prices, addresses) and 47% mentioned information related to names (e.g., addresses, contacts). On this basis, we implemented two parsers to count numbers and names in call transcriptions aiming to evaluate whether this information could be a relevant driver for highlighting information in the call. Example of text fragments that were considered to have a number include: “three hundred euro”, “half kilo”, “fourth street”, among others. We also implemented a third parser to count interrogative adverbs (i.e. why, where, how, when). Our reasoning is that the occurrence of an interrogative adverb implies the presence of a question, which is, by definition, an explicit request for information. Such request would therefore be followed by important information that might not necessarily appear in the form of numbers or names. Additional call-related variables include call length both in characters and seconds.

- *Participant's profile:* These variables were collected in the preliminary questionnaire, and they helped us to understand whether highlighting behavior can change based on demographic features.
- *Quality of service:* We asked participants about the quality of calls and transcriptions in order to investigate how they might influence highlighting.
- *Contextual variables:* We supposed that the context of a call might influence the need for highlighting information. Furthermore, given that we asked participants to highlight *important* pieces of information, we also looked for an association between the stated importance of the highlights and the stated importance of the call.

Figure 3.1 shows the tested variables, their correlation with the highlighting parameters (*Highlighted* i.e. whether the call was highlighted or not, and *Highlight length* in chars), and their meaning. We report both variables that displayed a significant correlation and those who did not. This helps us to provide a more complete discussion of which of them would be more useful for predicting annotation of mobile phone calls as well as those that might not. Next we discuss these results. We refer to the names of studied variables in italics.

Patterns in mobile phone call annotations are better indicators of note-taking than most of the variables observed in the study. According to Figure 3.1, these pattern variables (*Total numbers/call, Total names/call, Total adverbs/call*) have some of the highest correlation coefficients with the length of the highlights ($\rho = .29, \rho = .28, \rho = .25$ respectively). We further normalized values of these pattern variables by call length, and correlations were also among the highest ones observed in our study. ($\rho = .24, \rho = .19, \rho = .25$ respectively). In other words, the more numbers, names, and interrogative adverbs are mentioned in a call, or the higher the amount of them per unit of time, the higher the probability to highlight information—take notes—and also the longer the highlights might be. In the specific case of interrogative adverbs, the ones with higher correlation with highlight length were "when" ($\rho = 0.24$), "how much/how many" ($\rho = 0.18$), "who" ($\rho = 0.15$), and "where"($\rho = 0.10$). This suggests that the questions that would lead most frequently to highlight information are respectively those related to planning events (e.g. meetings, special events, holidays), discussing numeric values (e.g. prices), people and places, respectively. Lastly, *Call length* is the only variable that surpasses correlation results for patterns (in chars: $\rho = .35$; in seconds: $\rho = .29$).

Additionally, we found some weak yet significant correlations between contextual variables and the occurrence of patterns in calls. They are however worth being reported as they might indirectly influence highlighting behavior. We found that calls made to the caller's mate tended to have less proper names ($\rho = -0.1$) while calls made to friends seemed to have more proper names ($\rho = 0.13$). This suggests that couples tend to discuss topics less related to things that need to be referred by a proper name (such as places, people, or brands) maybe focusing more in their common daily activities. On the other hand, friends seem to mention more information related to third parties, news, recently discovered places, etc. (*e.g.*[7] *CALLER: "(...) Who's coming? Aitor, Raul...?"—CALLEE: "Aitor, Gorka, Raul and I don't know who else. (...)"*). In terms of the objective of the call, calls made with the goal to

[7] Actual examples from the study dataset.

set up a meeting/appointment were positively correlated with the number of names mentioned during the call ($\rho = 0.17$). This is most likely due to the mention of addresses or meeting points during the call, as these items frequently include streets or specific place names (*e.g. CALLEE: "(...) Should we meet at eight in... some bar of the San Juan Street? Is it good for you?"*). The opposite happened to social calls ($\rho = -0.13$). While this kind of conversations might also exchange names casually, social calls are more frequently made without a specific goal, instead with the need to "say hello" or to see how the callee is doing. In consequence, they less frequently contain proper names.

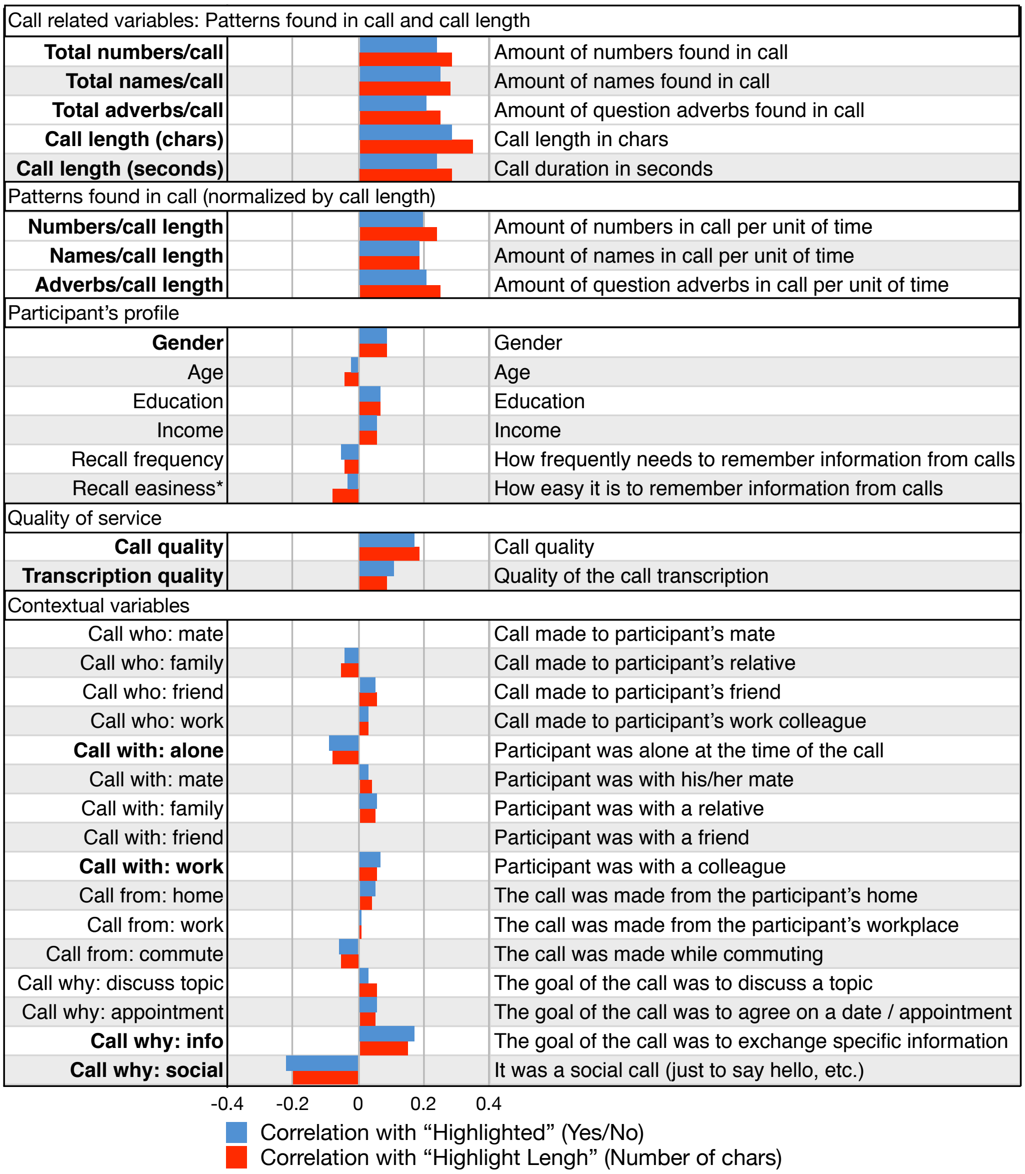


Figure 3.1: Correlations/Associations between highlight-related variables (*Highlighted*: Yes/No; *Highlight Length* in number of characters) and other variables related to the call task. Correlations between ordinal and non-normal interval variables were assessed using Spearman's Rho (ρ). Associations between dichotomous variables were assessed using the χ^2 derived Phi coefficient (ϕ). Variables in bold have significant coefficients at $p < .05$. * Recall easiness was found to be significantly correlated with Highlight Length only.

Profile and demographic information of the caller do not seem to be related to phone call annotation. According to Figure 3.1, most of the callers' demographic variables (i.e. age, education and income) did not reveal significant correlations with the highlighting variables. *Gender* (coded as 1: male, 2: female) seems to be an exception, implying that women might highlight more information than men. [5] obtained a similar finding for the context of work meetings. However, they reported a medium effect size for women to take notes more frequently and for longer than men, whereas we observed an effect below the standard weak correlation threshold ($\rho = .09 < .10$) [9]. Similarly, our participants' self-reported *Recall easiness* also revealed a below-weak correlation with length of higlights, implying that the easier one considers the recall task, the shorter his/her highlights are ($\rho = -.08 > -.10$). Future work should clarify whether gender and recall easiness are indeed significantly correlated with phone call annotation.

Quality of Service (QoS) parameters are weakly correlated to phone annotation. All QoS variables—as reported by participants—were positively correlated to making phone call highlights. Moreover, *Call quality* was positively correlated to both the highlighting activity ($\rho = .17$) and length of highlights (in chars: $\rho = .37$, in seconds: $\rho = .26$). One possible explanation is that the poorer the quality of calls, the less users speak and spend time in a phone conversation, thus reducing the probability of highlighting information.

Contextual variables play distinct roles in the note-taking activity. Information related to the call place, caller's companion and callee information did not reveal any significant relationship with call highlighting, or rather only below-weak correlations (see variables *Call from, Call with* and *Call who* in Figure 3.1). However, call objective (*Call why*) seems to be more connected with the users' highlihgting needs. For example, whenever participants made social calls (*Call why:social*, i.e. call objectives reported as *"just to chat"* or *"to say hello/goodbye"*), less highlights were made ($\rho = -.22$). On the other side, calls to give and/or ask for information (*Call why:info*) received more highlights ($\rho = .17$).

Importance of calls and importance of their notes are related to call context. By means of studying associations between call/notes importance and contextual variables, we observed some significant—although mostly weak—correlations that are worth mentioning. We found that calls made to the caller's mate had the tendency to have lower importance, as evidenced by a negative correlation ($\rho = -0.12$), while the opposite happened for calls made to friends ($\rho = 0.1$). It is possible that phone conversations between partners tend to have routine information that is more predictable and then frequently regarded as less important, while phone conversations with friends include information considered to be new or unexpected, thus being considered to be more important. Call importance tended to be lower for social calls ($\rho = -0.35$) while calls made to discuss a topic showed a tendency to be more important ($\rho = 0.13$). As mentioned before, social calls are mostly made for very informal or casual purposes, thus these calls tend to be regarded as unimportant. On the other hand, calls made with the goal of discussing a topic have a clear goal, which is probably the reason why they are more frequently considered important. With regards to the importance assigned to notes, we observed a positive association with calls made to a friend or to a service provider, like restaurants, shops, etc. ($\rho = 0.13$, $\rho = 0.14$, respectively). This suggests that highlighted information from these calls are frequently important for conducting further actions, for instance, attending a appointments based on highlights of location details, buying artifacts based on highlights of shopping items, or performing specific tasks based on highlights of instructions. Still regarding note importance, we found a similar positive association with calls made with the objective of setting up a meeting/appointment and for discussing a topic ($\rho = 0.15$ and $\rho = 0.13$ respectively). The first case might be due to the importance of the information in the notes for attending the commitment, and the second suggests that the discussion of topics frequently produces information which is important for later use. Finally, social calls tended to have less important notes ($\rho = -0.23$). This once again confirms that social calls—and notes derived from it—are generally not very useful for information exchange. Their purpose is probably to keep in touch with people we know, to nourish our social circles.

Important notes are taken from important calls. As a last result in this section, our experimental data corroborate what one would expect: importance of calls as evaluated by the callers, strongly correlates to the importance that they attributed to highlighted information in those calls ($\rho = .51, p < .001$). This finding might indicate that the higher one thinks is the importance of a call, the more likely important annotations will be created for the conversation.

These results support and expand our major findings from [6]. and further clarify some early doubts. For example, in our previous work we inquired whether the associations between the *Call why:info* variable and the highlighting-related variables (*Note Taken* and *Note Length*) were significant. With the larger dataset used in the work presented herein, we were able to corroborate that these associations were indeed significant. Similarly, the early apparent significant association between *Gender* and *Notes Taken* was corroborated in this work, which is also in accordance with results from [5].

In summary, our findings indicate that users tend to highlight—or annotate—mostly information containing numbers and names, such as phone numbers, addresses, dates, shopping lists, or contacts. These patterns can be easily identified signaling the importance to annotate calls. Profile and demographic information of the caller are most likely not relevant indicators of annotation, while QoS parameters could potentially inform it—poor quality calls may lead to shorter calls with fewer notes. Finally, contextual information such as the call objective and call length seem also to be good indicators for call annotation: while calls that intended to give or receive specific pieces of information did require annotations to be taken, social calls did not. Future work should investigate whether these variables related to mobile phone call annotation could also influence note-taking in work-related meetings [5, 4].

Addressing RQ3: How often and in which circumstances does each mobile phone party—caller and callee—provide information that the caller considers to be worth remembering?

With this research question we want to understand, in terms of turns, how frequently and in which circumstances the important information exchanged during a mobile phone call appears in each of the parties' turns. The dataset used to address this question—all calls that contained information highlighted by participants in the initial Study Phase ($N = 299$)—contained a total of $11,958$ turns ($\tilde{x} = 25$ turns per call, $min = 2$, $max = 463$). The amount of caller's and callee's turns per call were balanced across the dataset (caller: $\tilde{x} = 12$ turns per call, $min = 1$, $max = 232$, callee:

$\tilde{x} = 12$, $min = 1$, $max = 231$). From all annotated turns, 833 (7%) contained at least some information highlighted by the participants, i.e. were *highlighted turns*. From these highlighted turns, 441 (3.7%) were caller's turns and 392 (3.3%) were callee's. Next we will go deeper into understanding where the highlighted information tends to appear more frequently.

RQ3.1 Do callers annotate—by means of highlighting mobile phone call tran-scripts—information shared by the callee significantly more often than information shared by themselves?

As explained in the methodology section, we defined three groups of calls depending on where the highlighted information was contained: in the caller's turns, in the callee's turns or in both. Figure 3.2 shows the frequencies of these groups (confidence intervals in Figures 3.2 and 3.3 are Exact Binomial confidence intervals, calculated according to [8]).

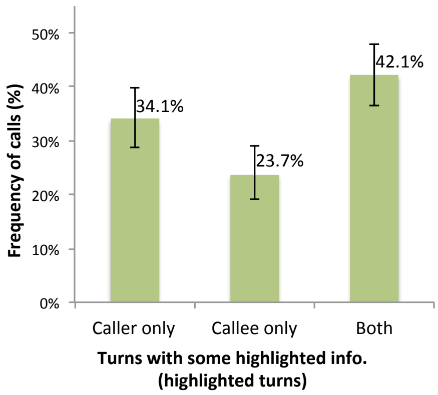


Figure 3.2: Binomial proportions of calls according to what party—caller, callee or both—had some information highlighted in his/her turns. Error bars indicate 95% confidence interval.

An analysis using McNemar's chi-squared test using a significance level of 0.05 indicates that the proportion of the first group (information highlighted in caller's turns only) is significantly higher ($\chi^2 = 5.20$) than the second (information highlighted in callee's turns only), and the proportion of the third group (information highlighted in both parties' turns) is also significantly higher ($\chi^2 = 14.80$) than the second. Therefore, there is a higher proportion of calls with important information provided by the caller (or by both parties) than by the callee alone.

This result may appear counterintuitive, as one might think that given that the caller is taking the initiative when dialing, probably more information should be contributed by the callee, as a response to the caller's inquiries. As stated in the Methodology Section, we looked for a possible explanation by performing a manual inspection of call transcriptions. This revealed that in many cases, the callers' highlighted turns were questions. Furthermore, the information included in those questions is contextual, frequently highlighted for adding meaning to the information exchange. For instance, in one of the calls including a discussion about buying a gift, the participant highlighted both the question *"what color?"* and the answer *"Dark grey, because I wasn't sure if I should buy the green or the blue one [...]"*. The actual information requested is in the second turn, while the first was highlighted to add meaning to it.

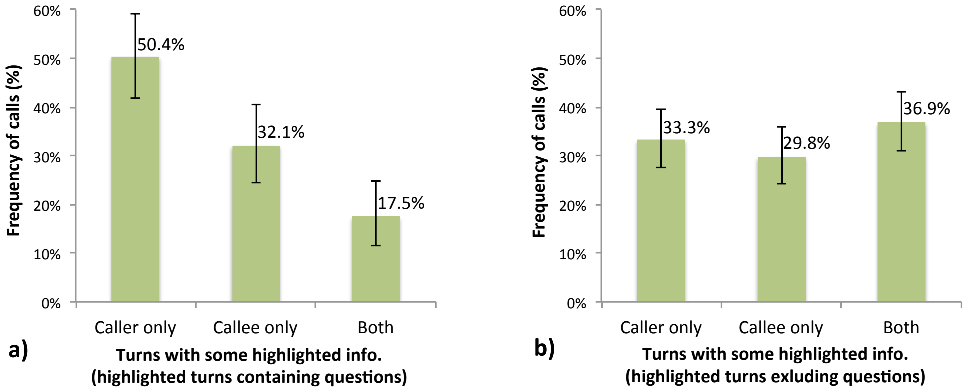


Figure 3.3: a) Binomial proportions of calls according to what party—caller, callee or both—had information highlighted in his/her turns. b) All highlighted turns excluding those with questions. Error bars indicate 95% confidence interval.

We found that about one quarter of the highlighted turns (23%) were questions. Figure 3.3a shows the proportions of highlighted turns containing questions, according to their location on each parties' turns. A McNemar's test confirmed that there are more calls with highlighted questions in the caller's turns than calls with highlighted questions in the callee's turns ($\chi^2 = 5.1$) or in both parties' ($\chi^2 = 20.82$). According to these results, from the caller's perspective, it is him/herself who makes most of the important questions in a given call, while also providing context for information exchanged during the call.

We further looked at the *number* of highlights—rather than the presence of highlights per call—in each party's turns that included questions versus those that did not. From the 441 caller's highlighted turns, 107 (24.3%) were questions—i.e. context establishing turns—and the remaining 334 (75.7%) could be considered as integral to what is worth being remembered. On the other hand, from the 392 callee's turns that were highlighted, 82 (20.9%) were questions, and 310 (79.1%) were not. In this analysis, however, we did not observe a significant difference between the rate of highlighted questions in the caller's turns versus in the callee's turns (as we observed previously when analyzing presence of highlights *per call*). From these findings, we argue that it is more likely for calls to have questions highlighted only in the callers' turns than only in the callee's turns, which suggests that context to important information tend to be provided by the caller. However, whenever callers highlight questions only in the callee's turn, often more questions per call are highlighted.

Figure 3.3b, shows that removing the highlighted questions from the highlighted turns, made the proportions of highlighted turns between the three groups similar. In this case, no significant difference was found using McNemar's tests ($\chi^2 = 0.4, p = 0.53; \chi^2 = 0.3575, p = 0.55; \chi^2 = 1.7, p = 0.19$). In this case, roughly one third of the calls contained highlighted information in the caller's turns exclusively, one third contained highlighted information in the callee's turns exclusively, and one third contained highlighted information in both parties' turns. These results suggest that when looking for important pieces of information exchanged during a mobile phone call (excluding questions that usually just provide context), one might find them in any of the parties turns, or even in both of them.

While the exchange of information excluding highlighted questions might give a clear view of where the important information more frequently appears, the data suggests that questions should be involved in the annotation process—specially those made by the caller. It is worth noticing that this is backed up by the results from section 3.4, where we found that the presence of interrogative adverbs was one of the variables that displayed highest correlation with the note-taking activity. While questions may play a number of roles in a conversation [11], most of the time they are used for gaining information [28]. However, given our findings, we believe that an automatic annotation system should not consider questions as being merely markers for important information that is yet to appear during the conversation. Instead, questions themselves contain relevant contextual data that should be linked with information from their corresponding answers in order to construct meaningful notes.

Regarding our RQ3.1, we conclude that callers do *not* highlight information on the callees' turns more often than on their own turns. In fact, callers more often highlight information in their turns exclusively or in both parties' turns. Analysis based on manual inspection of calls suggests that this is mostly due to annotating caller's own questions to provide context for the information highlighted in the callee's turns, thus composing one meaningful note. Moreover, these findings can indicate collaborative construction of important information by both caller and callee. Next, we look further into this latter case.

Which factors mostly influence callers to highlight information in the turns of the two parties involved in a call (i.e. caller's turns, callee's turns, or both parties' turns)?

We investigated if call context is related to the occurrence of important information in each parties' turns. More specifically, we looked for associations between contextual variables (e.g. calling a family member) and the presence of highlights in the caller's turn exclusively, in the calle's turn exclusively, or in both parties' turns. Associations between binary variables were calculated using the χ^2 derived Phi coefficient (ϕ) and a significance level of 0.05 was used.

Calls to one's significant other are collaborative: highlights are usually made in both parties' turns. For calls made to the caller's partner (variable *Call who: mate*), we found a significant positive association with the presence of highlighted information in both parties' turns ($\phi = 0.12$). We also found a negative association between calling the partner and highlighting information in his/her turns—i.e. callee's turns ($\phi = -0.12$). These results suggest that when having a phone conversation with their significant ones, callers found important information in both what they said and in what their partners said, as opposed to finding it in one party's turns exclusively. Manual inspection of call transcripts suggest that calls between members of a romantic relationship tend to be collaborative, sharing common interest items such as activities to attend together, discussing what groceries to buy, etc. A recurrent example of this is the case of conversations where partners synchronize their schedules (highlighted turns in bold):

Example 4:

```
CALLEE:  So that's it.  What time are you going to the gym?
CALLER:  I think today I'll go at 8:15, so we can meet before.
CALLEE:  Fine.  Yes, I hope so.  In fact, we have work here, but
         I think I'll leave no later than 6.  So at 7:00
         I should be at home.
CALLER:  Fine it's just that...
CALLEE:  Yes, tell me, tell me.
CALLER:  Of the activities they had before, I didn't like
         anything.  Then I'll go to Body Combat, which is
         at 8:15.
```

Social calls are also collaborative: highlights are usually made in both parties' turns. We also found a positive association between making a social call and highlighting information in both parties' turns ($\phi = 0.16$). After inspecting transcripts of these conversations, we observed how they are frequently related to daily and informal events, or to common acquaintances and news. Moreover, these conversations frequently flow without time constraints, and thus new conversation topics arise, bringing up spontaneous bits of information that can be considered relevant. For instance, the conversation between a woman and her mother shown in Example 5:

Example 5:

```
CALLEE:  They have a baby girl, Alba, and now they are waiting
         a kid, in about a month, they told us.
CALLER:  And how they are going to call him?
CALLEE:  What?  I couldn't hear you
CALLER:  I said how they are going to call him?
CALLEE:  Mario, they told us.
CALLER:  Vicky's child is going to be called Mario, too.
         (...)
CALLER:  And next thursday is La Candelera*, isn't it?
CALLEE:  Yes, this thursday, La Candelera[8].  Your father has a
         party, he says he is going to buy some seeds.
```

In informational calls, the caller has important things to say. Conversely, we found a positive association between making phone calls to exchange specific information (variable *Call why: info*) and highlighting the caller's turns ($\phi = 0.16$). For this type of calls, the need to share information with the callee is probably the main trigger for placing the call, thus the important information is more frequently in the caller's turns. This happens frequently in situations when the person calls to inform the callee of past or forthcoming events, for instance: *"Hey, about the scarf I told you somebody forgot in my house, well I think it's yours. [...]"*, or: *"I called to tell you that I am going to Burgos tomorrow in the morning"*. A negative association between making these information-based calls and highlighting information in both parties' turns ($\phi = -0.18$) further suggests that these types of calls tend to be task oriented. Therefore, giving or—less frequently—receiving information is enough without much collaboration between parties to generate content worth to be highlighted in both turns.

Other contextual variables, such as companion during a call (variable *Call with*) and location at the time of the call (variable *Call where*), did not show significant associations with highlighting a certain party's turns (i.e. caller's turn, callee's turn, or both turns). In the next section we study time-based changes in annotation behavior, in order to address $RQ4$.

Addressing RQ4: Temporal effect on highlighting needs

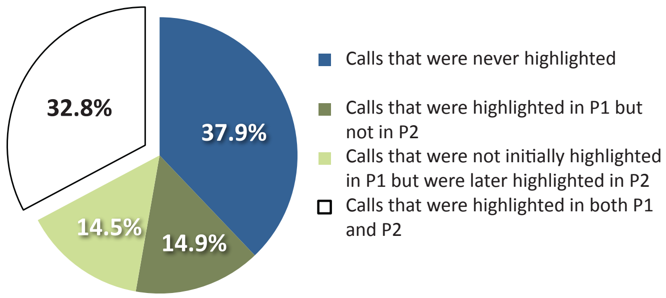


Figure 3.4: Overview of all calls used to study the temporal effect on highlights ($N = 235$). Calls in shades of green (67.2%) suffered a clear change in highlighting behavior. Calls in white need to be further investigated.

As described in the methodology section, the effects of time on highlighting behavior were studied using a subset of calls ($N = 235$) that participants revisited during phase two ($P2$). From this set, 37.9% were neither highlighted in the first phase of the study ($P1$) nor in the second phase ($P2$), meaning that participants did not change their highlighting behavior for these calls. On the other side:

- 14.9% of the calls were highlighted in $P1$ but not in $P2$. This decision was justified by the participants in most cases as due to a temporal effect in the original highlighting needs (e.g. participant 54: *"this event already happened"*).
- 14.5% of the calls were not initially highlighted in $P1$ but were later highlighted in $P2$. These calls were later considered to have useful information, thus indicating that users' first intention about discarding information exchanged in mobile phone calls can also be affected by time.

These results alone corroborate our fourth research question given that almost one third of the calls had their highlights drastically changed as an effect of time (29.4%; see Figure 3.4). Nevertheless, it is worth inspecting calls with highlights in both phases $P1$ and $P2$ (32.8%, $N = 77$) in order to investigate differences as a result of the influence of time.

RQ4.1 Does the topic and amount of information in highlights change across time?

As explained in the methodology section, differences between highlighted information in $P1$ and $P2$ were categorized, and we then used these categories to calculate the proportion of highlights in $P2$ that differed, in terms of their topic, from highlights in $P1$. Our findings reveal that 6.38% of the calls had highlights with different topics in $P1$ and $P2$, 10.64% of calls had highlights with the same topic but more information in $P2$, and 10.21% of calls also had highlights with the same topic but more information in $P1$. On the other hand, only 5.53% of the calls kept the same amount of information in highlights made during both phases of the study (see Figure 3.5).

These results indicate that 56.6% of the calls (133 calls) used for studying temporal effects were approached differently by participants in $P2$, either by highlighting different pieces of information, by highlighting information in calls that were initially considered not to have valuable content, or by not highlighting the calls anymore. Conversely, only 43.4% (102 calls) of the calls received the same highlighting approach in both phases, because they were never highlighted or because the highlighted information was the same.

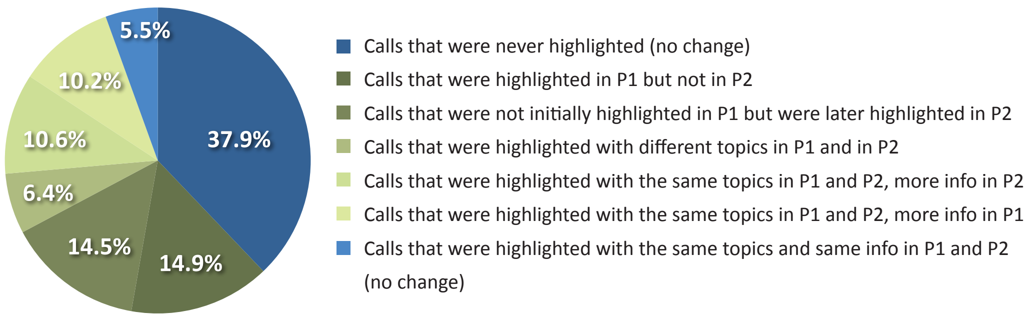


Figure 3.5: Overview of changes on highlighting behavior between $P1$ and $P2$. Calls in shades of green (56.6%) were approached differently by the participants, while those in shades of blue (43.4) suffered no change.

More than a half of the mobile phone calls involved in the study suffered a change in their highlights after time passed. The forthcoming research questions are aimed to further investigate and understand this phenomenon. In order to answer them, we considered the time difference between highlights in $P1$ and highlights in $P2$. This difference spanned from a minimum of 22 days to a maximum of 82 days. We grouped calls into temporal bins according to these time differences. We tried different numbers of bins (2 bins of 31 days each; 3 bins of 20 days each, etc.), and our results were consistently similar for all combinations. Therefore, we opted to use three bins since they provided enough temporal detail to understand highlighting dynamics and increased statistical power for data analyses across bins (3 bins have more data points per bin compared to 4+ bins). Hence, $Bin1$ contained calls with an interval of 22 to 41 days between their highlights in $P1$ and $P2$ ($N = 54$), $Bin2$ contained calls with an interval of 42 to 61 days ($N = 105$), and $Bin3$ contained calls with an interval of 62 to 82 days ($N = 76$). We used this grouping strategy to address the next research questions.

RQ4.2 How does this information change unfolds across time?

In the methodology section we explained how highlights in $P1$ and $P2$ were manually compared so as to identify whether highlights preserved the same topic. Figure 3.6 shows the proportion of calls for which highlighted information changed topics in $P2$ compared to $P1$ for each of the three temporal bins. While none of the highlights in $Bin1$ changed topic between $P1$ and $P2$, 25% of the highlights in the following temporal bins ($Bin2$ and $Bin3$) did change topic. This suggests that time affected participants' highlighting behavior in a way that, after 42 or more days, they found some pieces of important information that were thematically different to the ones they originally highlighted.

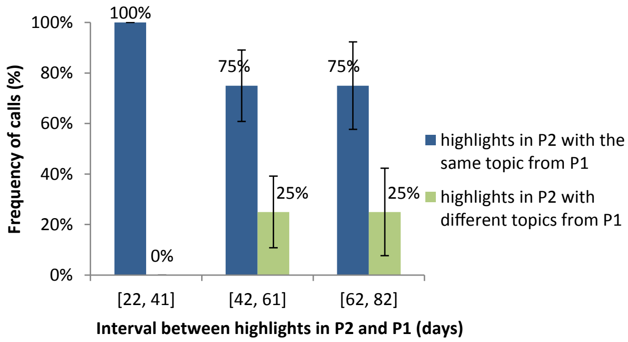


Figure 3.6: Change of topics for highlights made in study phase 2 compared to highlights in study phase 1. Error bars indicate 95% confidence interval.

We further looked into those calls that preserved the same topics in their highlights in both study phases ($N = 62$) to verify if the change in *amount* of highlighted information could also be explained by a temporal effect. Figure 3.7 shows the proportions of calls with the same topic and amount of information in $P1$ and $P2$ *vs.* calls with the same topic and different amount of information in $P1$ and $P2$ ($p < .05$). It shows that the former decreases and the latter increases across bins—i.e.across time. This means that participants had the tendency to change the amount of highlighted information more often for calls that were highlighted later in time.

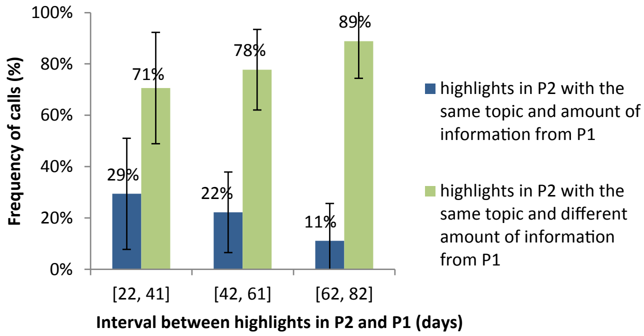


Figure 3.7: Change of amount of information highlighted in study phase 2 (P2) compared to study phase 1 (P1) for call highlights with the same topic. Error bars indicate 95% confidence interval.

To better understand how this change happens for calls that kept the same topic in their highlights during both study phases, we looked at their difference in amount of information between highlights from $P2$ and highlights from $P1$ (see Figure 3.8). During the first two time bins (i.e. when $P2$ and $P1$ are separated by 22 to 61 days), the same amount of information is highlighted in $P2$ compared to $P1$. However, in $Bin3$ apparently less information is highlighted in $P2$ for this set of calls.

So far we have looked into how changes of topic and amount of information for calls highlighted in both study phases evolved across time. Next we investigate if the perceived importance of highlights also changes as time passes, and later the reasons for all these changes. While *RQ4*, *RQ4.1* and *RQ4.2* were based on the analysis of the highlighted information, next two research questions leverage the participant's self-reported perceptions of these changes.

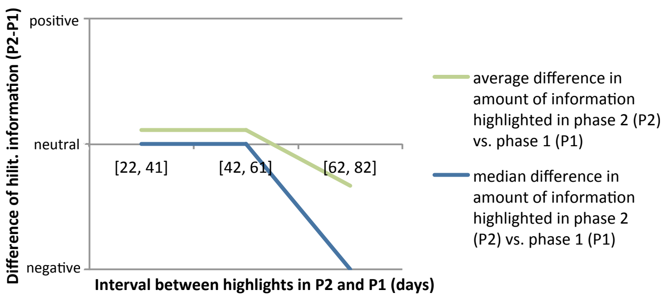


Figure 3.8: Difference between the amount of information highlighted in P2 compared to the amount of information highlighted in P1. Participants made fewer highlights in P2 compared to P1 when these phases were separated by 62 to 82 days (Bin3).

RQ4.3 Does the importance of the highlighted information change across time?

Figure 3.9 shows how importance of highlights made in $P1$ changed for highlighs made in $P2$, as reported by the participants. For $Bin1$, the increase in importance is negative, meaning that highlights made about a month after the first highlighting phase were perceived to be less important. For longer intervals between highlights, they were perceived to have an importance similar to that of the previous bin ($Bin2$ and $Bin3$: $42 - 82$ days). The change of importance observed in $Bin1$ was significantly different than the change of importance observed in $Bin2$ ($Z = -2.698$, $p < .01$) and in $Bin3$ ($Z = -2.823$, $p < .01$). Conversely, the change of importance observed in $Bin2$ was not significantly different than the one observed in $Bin3$ ($Z = -.281$, $p = .78$).

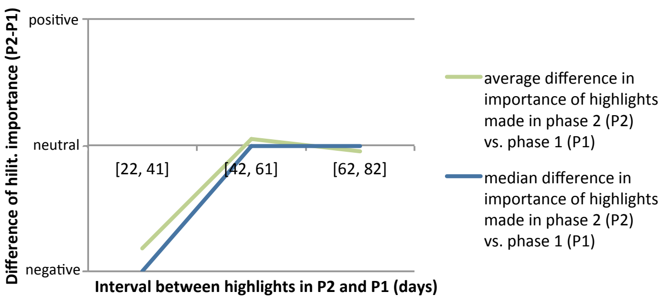


Figure 3.9: Difference between the importance of highlights made in P2 compared to the importance of highlights made in P1. The importance of P2 highlights is lower than the importance of P1 highlights in Bin1 (22 to 41 days) and it increases later on.

RQ4.4 What are the reasons presented by participants for the change in their highlighting behavior?

We studied participants' answers to our question regarding *why* their annotation behavior changed, when it was the case. As explained in the methodology section, participants' explanations for the differences—if any—between recent highlights in $P2$ and older highlights in $P1$ were manually categorized in a few reasons. We found that time between highlights seemed to influence these explanations. Figure 3.10 shows how the reasons evolved across time. While reasons related to *temporal effect* tend to decrease over time, participants cited increasingly more reasons related to *archiving* highlighted information and *refining* highlights to explain the changes of highlighting.

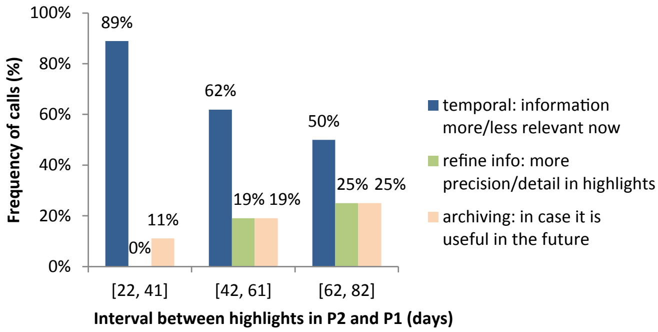


Figure 3.10: Reasons for changing highlighting strategies in study phase 2 (P2) compared to study phase 1 (P1).

RQ4.5 How can annotation behavior change be explained in terms of all previously investigated factors?

From the calls studied during both study phases ($N = 235$), 56.6% revealed a change in participants' annotation behavior, i.e. they highlighted calls that were not highlighted before, they did not annotate previously highlighted calls, or the information they highlighted was different between study phases. Next, we describe and discuss the possible causes for this change over time.

About one month after the first annotation ($Bin1$: from 22 to 41 days), all highlighted calls were related to the same topic (see Figure 3.6). This result might indicate that participants were still highlighting information based on the *original purpose of the call*. By the time they answered call questionnaire in $P2$, probably highlighted information was already consumed, thus explaining why highlights made in $P2$ were considered to be less important than those taken in $P1$ (Figure 3.9). Also for 89% of these calls, differences between highlights in $P1$ and $P2$ were justified to be due to a temporal effect. In other words, while still highlighting based on the original purpose of the call, participants perceived the new highlights to be less important because the old ones were already consumed and, due to a temporal effect, they were not important anymore (e.g. participant 54: *"this event already happened"*). This seems to agree with findings from [18] for the context of work meetings, in which pen and paper notes were found to be missing important information after around 30 days.

Up to two months after the first annotation ($Bin2$: from 42 to 61 days), participants started to change their annotation behavior by *annotating new topics* beyond the original purpose of the call (Figure 3.6). By considering the newfound topics worthwhile, the overall valuation of highlights increased with respect to the previous bin: while in $Bin1$ the valuation of highlights taken in $P2$ was depreciated with respect to the valuation given to highlights made in $P1$, in $Bin2$ both valuations were similar, revealing an appreciation for highlights made in $P2$ in the mid-term (Figure 3.9). A possible explanation for this is found in Figure 3.10, which shows that participants' annotation strategies changed from addressing immediate needs to recording information that might be useful in the long-term (*e.g.* participant 23: *"Because it is important to remember the name to be able to search information about it whenever it is possible"*), or refining previously highlighted information (e.g. participant 18: *"I've highlighted more details"*). Although our results are related to daily mobile phone calls rather than work meetings, they might shed light on the reasons why previous work has found that meeting participants try to keep their notes accurate after meetings, and half of them even rewrite them [18, 34].

Finally, from 62 to 82 days after the first annotation ($Bin3$), participants' behavior was quite similar to that observed in the previous interval $Bin2$: importance of new highlights in $P2$ was once again the same as highights in $P1$ (Figure 3.9), annotations were again taken beyond the original purpose of the call (see Figure 3.6), and more often participants considered archiving information for future use rather than addressing an immediate need (Figure 3.10). Nevertheless, the amount of information highlighted in $P2$ decreased with respect to information highlighted in $P1$, a behavior not observed in any of the previous intervals $Bin1$ or $Bin2$ (Figure 3.8). Therefore, this information is captured in a different way across time (Figure 3.7), particularly with less details due to a temporal effect in the long-term.

From the results discussed herein, we conclude that there is an appreciable effect of time on people's annotation needs of their daily mobile phone calls. Particularly, we found it to be quite interesting that users can depreciate information in the short term and look at them with renewed appreciation in the long term, being the inverse relationship also possible—valuing information only in the short term. These findings suggest the need to identify and archive information that could potentially be useful at some point in time (either in the short or long term). Information overload might be reduced by allowing users to hide annotations that are not relevant at a specific point in time.

3.5 Implications for Design

Our results suggest the need to create tools to support annotation of mobile phone calls. In this section we describe some design recommendations that might aid to that goal, aiming to satisfy the requirements of mobile phone users. The section was organized according to the tasks that such application should perform: recognizing potentially useful notes, annotating them in a way that satisfies users' needs as reported in our study, supporting users in taking advantage of their notes, and finally facilitating the consumption of annotated information.

On What to Annotate

Similarly to recommendations in the context of work meetings [34], we found that saving complete transcriptions of calls is not an adequate solution: in our study, only 4% of characters in the transcribed calls were annotated. A more efficient approach is to either process the call audio or parse its transcribed text towards automatically identifying potential fragments that the user would be interested in annotating. Deciding which parts of a call should be annotated without consulting the caller's opinion is not a trivial task. Nevertheless, our study revealed a number of *patterns* in mobile phone calls that are usually annotated, such as phone numbers, dates, addresses, prices, shop/to-do lists, contact names, activities, among others. In fact, we automatically identified these recurrent patterns in participants' calls using simple text parsers—as described in the results and discussion section—and verified that their presence in calls are indeed significantly correlated to whether calls are annotated or not. Other techniques shall be used to automatically annotate patterns without requiring full transcription (e.g. dynamic time warping as applied by [7]).

Our study reveals that, given the collaborative nature of mobile conversations, information worth being annotated is not always going to appear in a few phrases of single turns. Instead it is going to be spread over a number of turns contributed by both participants. For example, while questions help to find *when* important information might be about to appear (completion of a question-answer adjacency pair [30]), annotations should not be focused on answers only. Instead, it should start by analyzing questions in order to understand call purpose [11], and then extracting and analyzing meaningful relationships with the corresponding answers. A similar approach extended to other types of relations between turns can lead to the construction of useful notes.

While our results show that the caller finds information worth to be annotated in both parties' turns, some deviations from this behavior were also found. For instance, in informative calls it was more frequent for callers to highlight information in his/her own turns. If noteworthy information in this type of calls could be obtained by only recording an analyzing the caller's channel, bandwidth and audio processing might be reduced with the consequent improvement of quality of service by providing faster and/or better results[9].

Therefore, the annotation application should also take into account *call-derived data*, such as relationship with the interlocutor and call time, as well as leverage embedded sensors to gather relevant *contextual information* for the note-taking activity, such as the objective of the call (e.g. to determine if a call fulfills only social purposes or if it was made to get or receive specific pieces of information). By identifying call context, call QoS parameters—via analysis of the microphone signal—and patterns in the calls, the need to annotate a phone call might be detected and potential annotations inferred.

On How to Annotate

The majority of participants in our study reported using their *mobile phones* to annotate information exchanged in daily phone calls. Although our sample was composed of only smartphone users, these devices are becoming predominant worldwide[10] and should be considered one of the most relevant annotation media for daily phone conversations. They are also very convenient given that users always have them nearby during a phone call, whereas other annotation artifacts are not necessarily available at the same time, such as laptops or paper and pen used in the context of work meetings [3, 34].

Although being more convenient, mobile phones also impose a significant restriction for annotating calls. When making a phone call, one's hands are usually busy, thus preventing annotations to be appropriately taken on-the-fly. In fact, in only 8.8% of all calls made during the study, participants had both their hands free for taking notes. Our findings suggest that annotations of phone calls should be better addressed using an offline approach—i.e. performed after the call. In addition, the process seems to have potential to be automated given evidence of general

[9] Additionally, unintentional privacy breaches could be limited by complying with one-party consent laws, such as: http://www.gpo.gov/fdsys/pkg/USCODE-2011-title18/pdf/USCODE-2011-title18-partI-chap119-sec2511.pdf

[10] According to the International Data Corporation (IDC), in the first quarter of 2014, a total of 281.5 million smartphones were shipped worldwide. (press release: http://www.idc.com/getdoc.jsp?containerId=prUS24823414).

non-personalized patterns in phone conversations that are usually considered to be relevant (i.e. numbers and names in the form of addresses, phone numbers, codes, prices, contact names, shopping lists, etc.). Nevertheless, we found mostly moderate associations between presence of patterns in phone conversations and annotation of the calls, which indicates that fully automated methods might not achieve high accuracy rates when based mostly on these factors. Further research in audio processing and sentiment analysis might reveal other important sources of information that shall enable fully automated annotation tools.

Given all of the aforementioned recommendations, we conclude that a *semi-automatic annotation approach* to phone call annotation is required. As mentioned before, important information patterns should be automatically detected, extracted and stored immediately after the call. This would reduce the information overload that browsing through complete call recordings or transcriptions would mean for the users. On the other hand, we observed that only a fraction of the automatically identified patterns were actually annotated by participants of our study (e.g. 46.6% of numbers in calls that had numbers annotated, 42.5% of names in calls that had names annotated). That said, users should also have the option to manually inspect these pre-annotated patterns and approve those of particular interest to them, thus reducing overload in future recall tasks. The semi-automatic approach should also enable user-derived notes.

On What to Do with Notes

Phone call annotations can be related to a number of activities and used in different ways. For example, one might annotate details of a doctor's appointment discussed over the phone and—right after the call—transpose the notes to an electronic calendar tool. This suggests that once notes are automatically detected and presented to users, the annotation tool should allow them to take *actions* on the given notes by associating them to the appropriate application, e.g. creating reminders or appointments in the phone's calendar tool, saving phone numbers and email addresses in the contacts list, etc.

Other possible actions that may be taken on notes are related to sharing. Given the usual two-people setting of most phone calls, we foresee the opportunity to explore collaboration between caller and callee for the note taking activity. Collaboration can also be implemented to connect third-parties mentioned during calls. According to data from our study, 40% of all annotated calls mentioned either the caller's or callee's relatives, friends or colleagues. The possibility to share (e.g. by email, tweets, SMS, etc.) information extracted from mobile phone calls in an effective way might lead to the design of innovative collaborative tools.

On How to Consume Notes

Our findings reveal strong influence of time in people's phone call annotation needs. While users initially annotated information related to the purpose of the call and disregarded the remaining information exchanged with the callee, later on they considered the non-important pieces of information to actually be worth remembering. Moreover, what was once said to be important, later on it was not annotated at all. This means that users' first impression about annotating any given call is commonly related to their short-term needs, which does not exclude the remaining information from being useful in the long-term. This aspect is notably missed by audio-buffer-based solutions such as those depicted by [15] and [13]. While their proposed solutions provide fast access to the most recent segments from a conversation, they do not offer mechanisms for deeper browsing additional pieces of information that may become more relevant in the long term.

In order to support these time-based needs, we suggest that mobile phone annotation tools should: (1) automatically *record every note candidate* to avoid discarding notes that shall become prominently important in the future, and (2) offer to the user the possibility to provide feedback on which of the automatically detected notes are more relevant, so as to highlight them using a *multilayer annotation visualization interface*. One possible implementation of this visualization technique is suggested in the following: notes that users manually select as important should be put in the first layer for privileged retrieval; remaining notes should be stored in the second layer, effectively preserving every annotation for long-term annotation needs (e.g. archiving); finally the whole call transcription—or recorded audio—could be stored in the third layer (in case such information is available). The whole transcription might be specially useful for providing important contextual information not initially contained in the automatically generated notes—for example, inside questions asked during the conversation. When users browse or search notes, first layer information should be ranked higher and thus presented before second layer notes. If users cannot find the information they are looking for in the first two layers, the third layer could be used. Temporal effect on importance of notes can be further addressed by letting users manually downgrade notes from the first layer whenever they become less important. Similarly, users should also be able to upgrade second layer notes to reflect their dynamic annotation needs.

Another interesting result on the temporal effect of notes is the change of annotation strategies over time. We observed that notes taken about one month (22-41 days) after the first annotation still focus on the original purpose of the call, whereas beyond this point people might annotate other pieces of information not previously considered to be important. That said, the proposed multilayer interface should leverage this finding and *provide awareness on annotations "time-to-live" by decreasing the emphasis on the first layer notes and/or increasing the emphasis on the second layer notes as time passes*. A sudden switch of layers should be avoided since the majority of users tend to keep annotating information related to the original purpose of the call. Alternatively—and more appropriately—the interface could keep the relevance ordering of layers while using special cues for the temporal effect. An example of implementation is to attach a thermometer-like indicator to the first layer notes so as to inform how "fresh" they are with respect to notes in the second layer. Other examples include greying first layer notes and highlighting access to second layer notes.

We consider that in order to facilitate the eventual retrieval of notes, the preservation of additional contextual information is of great importance. As pointed out by [21], the interpretation of the context where information arises greatly influences our ability to remember it. Metadata related to the calls—party name, time, location at the time of the call, etc.—, which are easily gathered by modern smart phones, can help to the purpose of retrieval of the important information.

As a summary, we propose that a *semi-automatic phone call annotation method*, implemented by means of a mobile phone app developed after our design implications, should perform these steps: (1) storing the full call transcript, (2) automatically detecting candidate notes and present them to the user, and (3) letting users fine tune these notes if they want to. The app could present notes by means of a *multilayer annotation visualization interface*, as described in the previous subsections. Our ongoing work leverages findings presented herein towards investigating how mobile technology can best support the annotation of daily mobile phone calls.

Acknowledgements.

We thank Diego Urdiales, Isabel Ordas, Miguel Angel Santiago, Sergio Romero, Edgar Olivares, Martin Pielot, José San Pedro, Nuria Oliver, Josep Blat and all participants of the study. This work has been partially funded by the EEE (TIN2011-28308-C03-03) project. Telefonica Research participates in the Torres Quevedo subprogram (MICINN), cofinanced by the European Social Fund, for researchers recruitment.

References

[1] I. Arminen and M. Leinonen. “Mobile phone call openings: tailoring answers to personalized summonses”. *Discourse Studies* 8.3 (2006), pp. 339–368.

[2] T. W. Bank. *ITU releases latest global technology development figures ITU releases latest global technology development figures*. DOI: 10.1596/978-0-8213-8991-1; Available at: http://www.worldbank.org/ict/IC4D2012. License: Creative Commons Attribution CC BY 3.0. 2013.

[3] M. Bernstein, M. V. Kleek, D. Karger, and M. C. Schraefel. “Information scraps: How and why information eludes our personal information management tools”. *ACM Trans Inf Syst* 26.4 (2008), pp. 1–46.

[4] A. Bothin and P. Clough. “Participants ’ personal note-taking in meetings and its value for automatic meeting summarisation”. *Information Technology and Management* 13.1 (2012), pp. 39–57.

[5] A. Bothin, P. Clough, and G. Bradley. “Quantitative analysis of individual differences in Note-Taking and talking behaviour in meetings”. *Proceedings of IADIS Multi Conference on Computer Science and Information Systems*. Freiburg, Germany: IADIS Multi Conference on Computer Science and Information Systems, 2010.

[6] J. P. Carrascal, R. de Oliveira, and M. Cherubini. “A Note Paper on Note-taking: Understanding Annotations of Mobile Phone Calls”. *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services*. MobileHCI ’12. San Francisco, California, USA: ACM, 2012, pp. 21–24.

[7] M. Cherubini, X. Anguera, N. Oliver, and R. de Oliveira. “Text versus speech: a comparison of tagging input modalities for camera phones”. *Proc. of MobileHCI ’09*. Bonn, Germany: ACM, 2009, 1:1–1:10.

[8] C. Clopper and E. S. Pearson. “The use of confidence or fiducial limits illustrated in the case of the binomial”. *Biometrika* 26.4 (1934), pp. 404–413.

[9] J. Cohen. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.

[10] B. J. Cragun. "Telephony annotation services. US Patent. US 7602892". Patent US 7602892 (US). Oct. 13, 2009.

[11] A. F. Freed. "The form and function of questions in informal dyadic conversation". *Journal of Pragmatics* 21.6 (1994), pp. 621–644.

[12] W. Geyer, H. Richter, and G. D. Abowd. "Towards a Smarter Meeting Record–Capture and Access of Meetings Revisited". *Multimedia Tools Appl.* 27 (3 Dec. 2005), pp. 393–410.

[13] G. R. Hayes, S. N. Patel, K. N. Truong, G. Iachello, J. A. Kientz, R. Farmer, and G. D. Abowd. "The personal audio loop: Designing a ubiquitous audio-based memory aid". *Mobile Human-Computer Interaction-MobileHCI 2004*. Glasgow, Scotland: Springer, 2004, pp. 168–179.

[14] K. Hinckley, S. Zhao, R. Sarin, P. Baudisch, E. Cutrell, M. Shilman, and D. Tan. "InkSeine: In Situ search for active note taking". *Proc. of CHI '07*. San Jose, California, USA: ACM, 2007, pp. 251–260.

[15] D. Hindus and C. Schmandt. "Ubiquitous audio: capturing spontaneous collaboration". *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*. CSCW '92. Toronto, Ontario, Canada: ACM, 1992, pp. 210–217.

[16] I. Hutchby and S. Barnett. "Aspects of the sequential organization of mobile phone conversation". *Discourse Studies* 7.2 (2005), pp. 147–171.

[17] B. H. Juang and L. R. Rabiner. "Hidden Markov models for speech recognition". *Technometrics* 33.3 (1991), pp. 251–272.

[18] V. Kalnikaitė and S. Whittaker. "Software or wetware?: discovering when and why people use digital prosthetic memory". *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. San Jose, California, USA: ACM, 2007, pp. 71–80.

[19] M. Kam, J. Wang, A. Iles, E. Tse, J. Chiu, D. Glaser, O. Tarshish, and J. Canny. "Livenotes: a system for cooperative and augmented note-taking in lectures". *Proc. of CHI '05*. Portland, Oregon, USA: ACM, 2005, pp. 531–540.

[20] M. V. Kleek, W. Styke, and D. Karger. "Finders / Keepers : A Longitudinal Study of People Managing Information Scraps in a Micro-note Tool". *Proceedings of the 2011 annual conference on Human factors in computing systems*. Vancouver, BC: ACM, 2011, pp. 2907–2916.

[21] M. W. Lansdale. "The psychology of personal information management". *Applied ergonomics* 19.1 (1988), pp. 55–66.

[22] E. Laurier. "Why people say where they are during mobile phone calls" (2001).

[23] S.-H. Lee. "Extended requesting: Interaction and collaboration in the production and specification of requests". *Journal of Pragmatics* 41.6 (2009), pp. 1248–1271.

[24] M. Lin, W. G. Lutters, and T. S. Kim. "Understanding the Micronote Lifecycle : Improving Mobile Support for Informal Note Taking". *Proceedings of the ACM Conference on Human Factors in Computing Systems*. Vienna, Austria: ACM, 2004, pp. 687–694.

[25] K. O'Hara, M. Perry, A. Sellen, and B. Brown. "Wireless world". Ed. by B. Brown and N. Green. New York, NY, USA: Springer-Verlag New York, Inc., 2002. Chap. Exploring the relationship between mobile phone and document activity during business travel, pp. 180–194.

[26] A. Piolat, T. Olive, and R. T. Kellogg. "Cognitive effort during note taking". *Applied Cognitive Psychology* 19.3 (2005), pp. 291–312.

[27] H. Sacks, E. A. Schegloff, and G. Jefferson. "A simplest systematics for the organization of turn-taking for conversation". *Language* (1974), pp. 696–735.

[28] J. M. Sadock and A. M. Zwicky. "Speech act distinctions in syntax". *Language typology and syntactic description* 1 (1985), pp. 155–196.

[29] E. A. Schegloff. "Beginnings in the telephone". *Perpetual contact: Mobile communication, private talk, public performance* (2002), p. 284.

[30] E. A. Schegloff. *Sequence organization in interaction: Volume 1: A primer in conversation analysis.* Vol. 1. Cambridge University Press, 2007.

[31] A. Weilenmann. ""I can't talk now, I'm in a fitting room": formulating availability and location in mobile-phone conversations". *Environment and Planning A* 35 (2003), pp. 1589–1605.

[32] J. Whalen, D. H. Zimmerman, and R. Whalen Marilyn. "When words fail: A single case analysis". *Soc. Probs.* 35 (1988), p. 335.

[33] S. Whittaker, P. Hyland, and M. Wiley. "FILOCHAT: handwritten notes provide access to recorded conversations". *Proc. of CHI '94.* Boston, Massachusetts, United States: ACM, 1994, pp. 271–277.

[34] S. Whittaker, S. Tucker, K. Swampillai, and R. Laban. "Design and evaluation of systems to support interaction capture and retrieval". *Personal Ubiquitous Comput.* 12 (3 Jan. 2008), pp. 197–221.

[35] L. D. Wilcox, B. N. Schilit, and N. Sawhney. "Dynomite: a dynamically organized ink and audio notebook". *Proc. of CHI '97.* Atlanta, Georgia, United States: ACM, 1997, pp. 186–193.

Chapter **4**

Who can we trust? A comparative evaluation of phone call annotation techniques

Juan Pablo Carrascal, Universitat Pompeu Fabra
Francesca Bonin, Trinity College, University of Dublin
José San Pedro, Telefónica Research
Rodrigo de Oliveira, Google
Nuria Oliver, Telefónica Research

Abstract

While millions of mobile phone calls take place everyday, most of the information exchanged through them is lost due to the lack of proper automated annotation tools. The problem of extracting and storing—annotating—information from a wide spectrum of sources has been addressed by scholars. However, the information exchanged during informal mobile phone calls remains largely unexplored to date. In this paper we evaluate and compare machine and human-based methods for the extraction of relevant information from real mobile conversations. We found that a high level of subjectivity is involved in the process of selection of noteworthy information from phone calls. We discuss some of the implications of our findings for the development of annotation solutions, and provide insights for future research on the topic.

4.1 Introduction

Mobile phones are increasingly pervasive in our lives, and are used for much more than just making phone calls. They have become sophisticated devices that provide access to limitless amounts of information online, and maybe more importantly, about ourselves. Even though there is a broad spectrum of mobile applications to satisfy a wide variety of information needs, voice information, as it is exchanged through every day calls, remains untapped at large. The saying *"words are carried by the wind"* certainly applies to this scenario.

In a recent study [9], 49% of the participants experienced the need to take notes from mobile phone calls at least sometimes, while 36% of them experienced this need often. In contrast, as pointed out by [21], the case of having both hands busy while having a phone conversation, preventing the user to take notes, is quite frequent. General purpose annotation tools [5, 28] may fall short in satisfying annotation needs in mobile situations. This suggests that there is an actual need for the development of automatic information extraction methods—annotation—for mobile phone calls.

In order to start proposing solutions, we have to take into account the characteristics of the data we are dealing with. As any type of conversation, mobile phone calls—and phone calls in general—have a particular structural feature, which is the sequence of *turns* [23], i.e. alternating interactions of the participants. Being the minimum structural unit of conversations, we want to study the process of identifying noteworthy turns i.e., turns that contain information that, because of its importance, should be preserved for later reference. We refer to them as *relevant turns*.

In this context, the work presented in this paper has two goals, both of them from a human-centric perspective:

1. To validate the machine-learning algorithm that was proposed in [6] for the automatic identification of relevant turns from mobile phone conversations.

2. To compare the quality of the relevant turns identified by the person involved in the conversation, by an automated system—the algorithm we just mentioned—and by an external annotator with no knowledge of the call context.

We therefore obtained three sets of relevant turns—one for each of the annotation techniques—, for a number of calls that were contributed by volunteer participants. Later we presented the same participants with the three sets of notes from their calls, and asked them to evaluate and compare those sets. Their feedback provided insights on which strategy to follow in order to provide adequate automatic annotation services.

4.2 Related work

A note may be the selection of fragments of information from the original source (e.g. when highlighting a piece of text from a book) or a new, derived written piece [12], created after mentally processing the original information [22]. In both cases, the objective is to preserve information that is considered to be important for future use. This *noteworthy* information frequently arises in collaborative situations such as lectures, meetings and phone conversations.

Scholars have widely studied the process of note-taking in work meetings. Research on the topic has found that notes are useful in order to create collective memories and optimize meetings [12]. Notes serve as memory cues to recall important events, instead of having to record every piece of information [28]. Notes have also been seen as markers that add structure to complete meeeting recordings. [26, 12, 29, 7]. In any of these cases, an individual's decision to take a note is the result of a number of factors that have been studied from different perspectives. Bothin et al. [7], found evidence that demographic and contextual factors, as well as the role of the participant, influence the needs to take notes from the meeting participants.

They also suggested that the presence of annotations during meetings might have some power to predict when something important is about to be discussed [8, 7].

Different artifacts can be used for taking notes. Simple techniques are still widely used nowadays, in the form of paper notes, corners of printed documents or text files [5, 28]. The process of taking notes, however, limits the annotators' possibilities to actively participate in the meetings [28]. In consequence, a number of computer-based solutions have been proposed, ranging from desktop computers [18, 5] to mobile devices [14, 17, 11, 13].

A few attempts have been done in *automating* the process of note-taking in meetings, in order to relieve the participants from the burden of such task. Banerjee et al. [3] investigated the feasibility of automatically extracting noteworthy pieces of information in meetings. They annotated a corpus of meetings with noteworthy segments, and reported a low inter annotator agreement (IAA), confirming the subjectivity of the task. They also conducted a wizard of oz experiment, reporting a precision of 35% and a recall of 41.5% from the human annotator. In a related work, Banerjee et al. [4] apply extractive meeting summarization techniques to automatically detect noteworthy utterances in meetings. They train a Decision Tree classifier over a collection of 5 meetings, obtaining an F-score of 0.14.

While there has been some previous work on note-taking in meetings, little work has been done to date about note-taking in mobile phone conversations. Unfortunately, many of the findings related to work meetings cannot be directly applied to mobile phone conversations. Mobile phone calls tend to be shorter and more informal than meetings, mostly only involve two people and they are often conducted while being on-the-go. Furthermore, meetings and phone calls are structurally very different, as meetings usually follow an agenda that is set up beforehand.

Conversely, mobile phone calls typically lack a predefined plan and are instead based on sequences of *turns* [23], interactions performed in an alternating fashion by the participants of the conversation. In fact, the alternating nature of a phone call has been widely studied in the field of Conversation Analysis (CA) [1, 16, 20, 25]. But even though CA has frequently used mobile phone calls as a subject of study, the specific topic of *note-taking* in mobile phone calls, that is, extracting noteworthy pieces of information in mobile phone calls is scarce in the academic literature.

An initial approach is presented by Carrascal et al. [9], who found that certain patterns from mobile phone calls and several contextual factors are good indicators of note-taking, at least from the caller's perspective. On the other hand, using a Machine Learning approach, Bonin et al. [6] proposed an algorithm for automatically identifying noteworthy turns in telephone conversations. The work therein presented offers a quantitative metric of accuracy but lacks a human-centric validation of the proposed algorithm. A part of our work tries to complement that work by providing such validation, hence our first research question is:

> *RQ1: How does a state-of-the-art machine learning-based algorithm that automatically identifies relevant turns in mobile phone calls performs in terms of satisfying real users annotation needs?*

It is still not clear which would be the optimal strategy—manual, automatic, hybrid—to follow to take notes from mobile phone calls. Being able to define an optimal strategy for this activity is a key step toward the implementation of a system or service that could provide a solution. Hence, our next research question focuses on comparing, from a human-centric perspective, different methods to generate annotations from mobile phone conversations.

It is apparent that the participants of the call would be the most adequate annotators. However, as O'hara et al. [21] mentioned, this approach might not be a realistic option as it is common to have both hands busy while having a mobile phone conversation and mobile phone calls are frequently conducted while on-the-go, not to mention the cognitive load implicit in the note-taking process [22], that may interfere with the natural flow of the conversation.

A third approach to obtain annotations is *crowdsourcing*, as it is commonly done with translation or transcription [15, 2] services. In this case, a human with no knowledge of the context of the call—an external annotator—would perform the annotation task.

In particular, we compare three approaches to take notes in mobile phone calls (the original user, an automatic algorithm and an external annotator without any contextual information) from the perspective of the original person who was involved in the phone call. This leads to our second research question:

> *RQ2: According to the original caller's opinion, how well do three different annotator's approaches perform the task of note-taking in mobile phone calls?*

Given that these annotators are (1) the original caller; (2) a novel machine learning algorithm developed for the task; and (3) context-free, external annotators. So, who shall we trust?

4.3 Methodology

We designed and deployed a quantitative study to evaluate and compare three techniques for the identification of important information in mobile phone calls. We built on previous work by Carrascal et al. [9] as a basis, but we had access to a larger participant pool and a larger dataset.

Participants

We recruited 62 participants (20 female), with ages ranging from 20 to 51 years. Mean age was 31.5 (SD = 7.52). All participants were *[omitted for blind review]* residents and reported being fluent in the *[omitted for blind review]* language, which was a requisite for participating in the study. The sample was well geographically distributed (38 distinct cities). All subjects had at least basic education, 3.2% finished primary education, 3.2% finished secondary education, 79% had a technical or university degree, and 14.5% had a superior degree (master, doctorate or postdoctorate). Annual income, as reported by participants, suggested that all socioeconomic levels were represented, with 27.4% having an income of less than €10K, 19.4% between €10K and €20K, 25.8% between €20K and €30K, 19.4% between €30K and €40K, 8.1% more than €40K.

Procedure

The study consisted of two phases:

Phase One ($P1$): Initial collection of call data

This phase lasted 64 days and its objective was to collect a set of mobile calls, their transcriptions, and information related to their context. Participants were recruited through popular web portals in *[omitted for bling review]*. They were asked a pre-study questionnaire to obtain demographic information, calling habits and annotation habits. Participants installed a VoIP application on their mobile phones. During this phase, whenever a participant made a call with his/her mobile, it was routed through our servers and recorded. The use of the application was transparent for the user, since it was installed as the default phone application. However, we explained to participants how to temporarily disable the application in case they would not want their calls to be recorded. As a compensation for participating in the study, all calls made through the system to national landlines or mobile lines were free of charge.

An external service was hired to transcribe the calls into text. After the calls were transcribed, they were made available to participants by means of a Web application, where participants could see a list of their recorded calls, and for each call, its audio recording and its transcription. Participants had the possibility to delete any call they considered to have sensitive content in the 24 hours after each call was made available in the Web application. Otherwise, the calls were considered to be contributed to the study by the participants, as stated and accepted by participants in the consent terms of the study. For these calls, we asked participants to highlight the pieces of text from the transcription that they considered to be important for later use, i.e. the information from the call that they would have wanted to take note of. In case that the participants did not find any important information, we asked them to explicitly say so. We also asked them questions related to the context of the call. A total of 658 calls were collected during this phase. Their lengths ranged from 26 to 4912 seconds ($\mu = 212.8, SD = 391.71$).

Phase Two ($P2$): Evaluation of annotations

Around two years after $P1$, we contacted 34 of the original participants and invited them to participate in the second phase of the study. In this phase, we asked them to evaluate and compare three types of annotations extracted from the calls they initially contributed to the study, as described later in this section. The time between $P1$ and $P2$ was important for reducing the ability of the participants to recognize the notes as their own, so the focus instead on the relevance of the information as included in the calls. This provided us with an unbiased evaluation, and thus a higher validity of the experiment. For this phase, we used a randomly selected subset ($N = 61$) from the calls collected during $P1$ in which these 34 participants highlighted at least one piece of information (the dataset obtained in $P1$ included highlighted and non-highlighted calls). Data collection for this phase took 20 days. In exchange, we offered participants a monetary compensation in the form of gift cards.

Relevant turns and annotators

For the purpose of this study, an *annotation* is a piece of text considered to be relevant[1] by an *annotator* who *highlights* it. Furthermore, given the importance of the turn as the minimum structural component of sequences in conversations [23], we defined the minimum unit of relevant information to be a turn and called it a *relevant turn*. In other words, as long as one word in a turn was annotated, the entire turn was considered to be *relevant*. While this scope sacrifices some precision when compared to working at word level, it provides us with less ambiguous units of information. Due to lack of context, isolated words or very short phrases are easily misinterpreted. To exemplify this, here is an excerpt of a call from our dataset[2]:

```
           (...)
1*  CALLEE: Let's talk about our meeting in
            [city omitted] on February 19th.
2   CALLER: Man, will you be there?
3   CALLEE: Yes, dude, I will go.
            (...)
```

The text that the participant highlighted is in bold. Therefore, in this example, the flagged turn (1) is a relevant turn. It can be seen that the whole turn conveys a more complete unit of information than the highlighted words alone. Furthermore, an inspection of the dataset showed that participants frequently highlighted entire turns, as shown in the next excerpt:

```
           (...)
1   CALLEE: How do you format it?
2*  CALLER: When you introduce the pen drive
            already in the partition, I guess
            a message should appear asking
            you to format it before
            installing Windows.
3   CALLEE: Oh, Alright.
            (...)
```

This suggests an intention to capture contextual information as opposed to highlighting more specific pieces of text. With this criteria, we obtained three sets of relevant turns for every call:

1. *Original relevant turns*: These are the relevant turns as highlighted by the original participants during $P1$. Whenever a participant highlighted a piece of text, we consider the whole turn where that piece is contained as a relevant turn. Though this is the ground truth obtained during $P1$, we wanted to verify if the relevance of the original highlighted information is consistent for participants across time.
2. *Automatically-extracted relevant turns*: We used an automatic machine-learning algorithm which we trained using the dataset obtained in $P1$. This algorithm analyzed the entire dataset and determined whether a turn was relevant or not. Using the results of this algorithm in $P2$ allowed us to measure its performance by asking the original caller about its quality.
3. *Manually-extracted relevant turns*: We asked two external annotators to determine the important information from the calls, in a similar way as the original participants did. These annotators, however, did not know the participants

[1] *"Information that is important or worth been preserved for later use",* as we consistently referred to it through the study.

[2] All examples are translated from the original *[ommited for blind review]* to English by the authors.

or had access to any contextual information of the call, except for the recordings and the transcriptions only. Our goal with the external annotators was to investigate the potential for third-parties to provide annotation services for informal calls.

4.4 Automatic extraction of relevant turns

One of the objectives of our study is the assessment of automatic noteworthy information detection models to fulfill the note-taking needs of users. To this end, we consider a state-of-the-art method for automatically extracting noteworthy information from phone calls [6]. This work poses the task as a binary classification problem. Turns in the conversations are represented as a collection of features, and are used along with a collection of annotations from the original callers to train a supervised classification model. This model can later be used to predict noteworthiness of new conversation turns.

In order to assess the subjective performance of this automatic method, and compare it to our human-supervised baselines, we fully replicated it. We therefore use a feature representation scheme comprising to two main classes of information: content features, which model the content information of the turns, and contextual features which model contextual information about the conversation and its particpants. This section provides a brief description of the most salient characteristics of the algorithm used. For further details and comprehensive description of the method we refer the reader to the original publication [6].

Content features

Content features refer to features extracted by analyzing the content of the conversations, that is, the actual words used by the speakers. This textual content can be extracted directly from the transcriptions of the calls. A subsequent pre-processing step is then required to parse the transcription and extract Named Entities (NE), such as names of persons, locations, organizations, other proper names, together with temporal expressions, numbers and dates. There are several software packages that provide this functionality; in our implementation we used freeling[3].

With this dataset, we built a feature set using the three main feature groups defined in [6]: *Turn-based*, *Dynamics*, *Conversational*. *Turn-based* content features allow to model the lexical and semantic aspects of the turns. They include information such as: presence of NE in the turn, presence of dates and other temporal expressions, Tf-Idf descriptive statistics or frequency of the different parts of speech (i.e. number of nouns, pronouns, verbs, etc.)

Dynamics content features model the semantic relations between a turn and its neighbors. In particular, this class of features considers the repetitions among consecutive turns, grouped by their part of speech function, as well as the identification of question-answer pairs.

Finally the *Conversational* content features are designed to capture the conversational flow and they include information about: the position of a turn in the conversation, the duration of the turn, the speaker who is uttering the turn, and the relative dominance of the speaker in the overall call.

[3] http://nlp.lsi.upc.edu/freeling/

The method also considers the generation of a bag of words representation for each turn, aimed at capturing specific terms that could act as triggers for noteworthiness (e.g. address, number, appointment). This bag of words is merged with the rest of the features to generate the Content-based side of the turn feature representation.

Context features

As previously found by Carrascal et al. [9], certain contextual factors influence the need to take notes during mobile phone calls. This is supported by the results of [6], in which the joint use of context and content features significantly outperform the classification performance of the content-only based representation. To leverage contextual features, we used the information obtained during $P1$. Again, we follow the approach described in [6] and consider the following groups of contextual features: *Call-based* features and *User-based* features.

Call-based features aim to capture contextual information at the call level, namely information about the location where the call happened (work place, etc), the time of the day and the duration of the call.

We also consider the objective of the call, which was gathered from participants in P1 by means of the Web interface. Note that, although this information is not directly accessible from the mobile data collected during the call, previous literature on conversation classification support the feasibility of inferring this information from the conversation content [19].

User-based features capture personalized information about the user, such as gender or age group. We also include the educational level, income and marital status.

Method Validation

We conducted a test for validating our implementation of the automatic noteworthiness detection method proposed in [6]. To this end, we used the same modeling algorithm (Support Vector Machines with Radial Basis Function kernel). We used a 10-fold cross-validation approach to predict the noteworthiness of all turns in our database, and computed the mean and the 95% confidence interval for the 10 values generated of precision, recall and F-score. Given that our objective is to measure the subjective preference of users for manually vs automatically generated notes, we only consider the 295 calls with annotations from our collection. The results obtained, reported in Table 4.1, are in consonance with the accuracy metrics reported in [6].

Precision	Recall	F-score
0.25 (0.22-0.27)	0.48 (0.44-0.51)	0.33 (0.30-0.35)

Table 4.1: Results of the automatic extraction evaluation: mean precision, recall, F-score and 95% confidence intervals in brackets.

4.5 External annotators

Context-free annotations

Our goal with this step was to obtain important information from the calls from the point of view of a person who does not know the context of the original call. We used a subset of 61 calls randomly sampled from the original dataset obtained in $P1$. We recruited two external annotators, one male and one female. We presented them with the transcriptions of these calls and gave them the same instructions we gave to the original participants: we asked them to highlight the information that they considered to be important for later use i.e., the information they thought it was worth taking a note of. This yielded two sets of relevant turns according to each external annotator.

Consequently, we measured the agreement between the external annotators at the turn level, i.e. if the two annotators highlighted words in the same turn, we considered that they agreed on that particular turn. Inter-coder agreement, tested using the Cohen's Kappa coefficient [10], was low (0.41, $p \ll 0.001$). Inter-coder agreement between the original participants and the external annotators was also low ($0.3, p \ll 0.001$; $0.27, p \ll 0.001$), suggesting that a high subjectivity is involved in the task.

In order to evaluate these relevant turns, we created two sets: the union and the intersection of the relevant turns provided by the two annotators. We then randomly divided the participants in the second phase of the study into two groups with 17 individuals each, and presented the union to the first group and the intersection to the second one. The intersection most likely would lead to less ambiguous but scarcer information, while the union might lead to more profuse but less precise information. This way we could find out if there is a preference for either option.

4.6 Measures

During $P1$, besides highlights from the transcriptions, we asked questions related to the call context including: 1) Relationship with callee; 2) Who was with the caller at the time of the call; 3) Location of the caller at the time of the call; 4) Objective of the call; 5) Level of importance of the call; 6) Reasons for highlighting text (if applicable); 7) Whether the caller could take written notes a the time of the call; 8) How important was to take notes during the call; and 9) General questions about sound and transcription quality. The participants reported objectives of calls in free text. For this reason, the objectives were classified by two coders (unrelated to the external annotators) as: "discuss topic", "appointment", "give/receive information", "ask favor", or "social". Inter-rater reliability was highly acceptable ($K = .81$, $p < .001$).

During $P2$, we asked participants to review their calls and transcriptions, and to evaluate each of the three sets of relevant turns. They were presented with the transcription of the calls including the highlighted pieces of text corresponding to the relevant turns, and the audio recordings as a reference. A maximum of three calls were presented to each participant, in order to avoid fatigue effects. We asked participants to suppose that a system had automatically analyzed their calls and produced a result containing a set of notes. Even though we presented complete turns as notes, we avoided using the term "turn" to avoid confusion. We presented the three sets of relevant notes (original, automatic, and context-free), one by one. The three sets were rotated for the different calls of every participant, so as not to biased her towards a particular set. They could read the notes, and then they had to select the level of agreement in a 5-point Likert scale, ranging from "Highly disagree" to "Highly agree", to each one of the following sentences:

- *These notes contain all the important information from this call.*
- *These notes contain unnecessary information.*
- *This result perfectly approaches the notes that I would have liked to take for this call.*

The first two questions were used to invite the participant to reflect upon the fact that the presented relevant turns might or might not miss important information—i.e. participants' perception of recall—and also that they might contain more than the necessary information—participants' perception of precision. The third question was the one we actually expected to express the participants' general satisfaction with respect to the set of relevant turns. We asked participants to explain the reasons for their answers. After the three sets of notes were individually rated, we showed the participants all the three sets, now side by side, and we asked them which set was the best, which was the worst, and why.

4.7 Results

The 61 calls used for the study contained 2038 turns ($\bar{x}$ = 33.4 per call, SD = 30.22). During $P1$, participants selected 223 turns (11%) to be relevant whereas the automatic algorithm found 532 turns (26.1%) to be relevant. From the total number of turns in the dataset, 380 (18.7%) were considered to be relevant by *at least one* of the external annotators—the union— and 120 (5.9%) by *both* external annotators—the intersection.

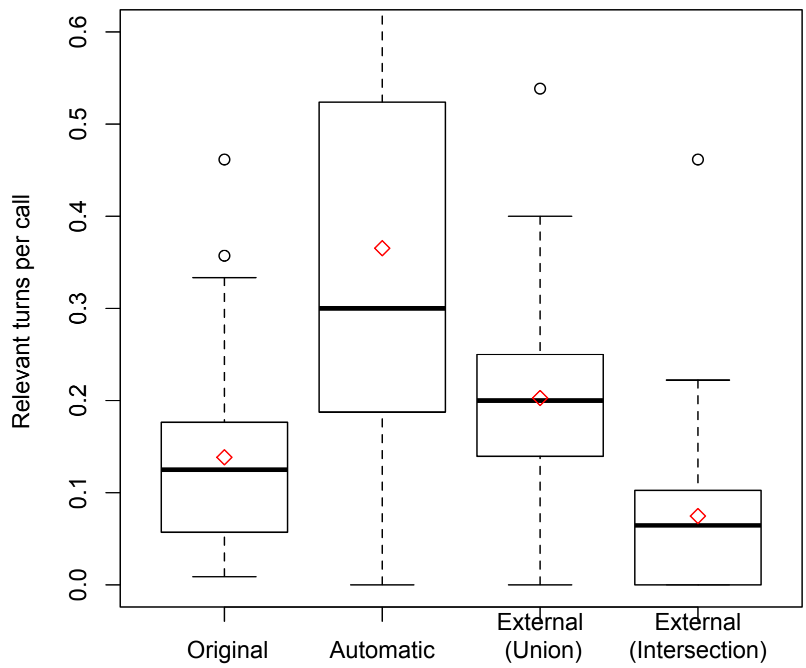


Figure 4.1: Relevant turns per call, according to the different annotation techniques. Upper whisker of the automatic algorithm's results is truncated for better visualization, but it extends up to 1. Red diamonds represent the mean.

Figure 4.1 shows the ratio of turns that were considered to be relevant per call, according to the different annotation techniques. It gives a comparative view of how selective the techniques were when determining if a turn is relevant or not. Compared to the Original relevant turns, It can be seen that the Automatic annotation technique produced the highest amount of relevant turns, while the Intersection of the External annotators produced the lowest. Both the External-Union and the Automatic techniques produced more relevant turns than the Original, while the Intersection provided less. Mann-Whitney U Tests between the distribution of the Original relevant turns per call and the Automatic, External-Union and External-Intersection confirmed these differences ($U = 775, Z = -5.6; U = 1009, Z = -3.89; U = 2337.7, Z = 2.5$; respectively, all significant at $p < 0.05$).

Evaluation of the annotations (relevant turns)

As stated in the *External annotators* section, during $P2$ we presented the union of the relevant turns as selected by the External annotators to half of the participants and the intersection to the other half. The union dataset included 32 calls containing 1031 turns, 170 (16.5%) of which where considered to be relevant by *at least one* of the external annotators. The intersection dataset included 29 calls and 1006 turns, 62 (6.16%) of which where found to be relevant by *both* external annotators.

Hence, we report results and carry out data analysis separately for the Union and the Intersection conditions. Figure 4.2 shows the distributions of ratings assigned by participants to the relevant turns for each of the 3 annotation techniques. In our analysis we use nonparametric methods due to the ordinality of the measures obtained in the questionnaires and the non-normality of the studied variables.

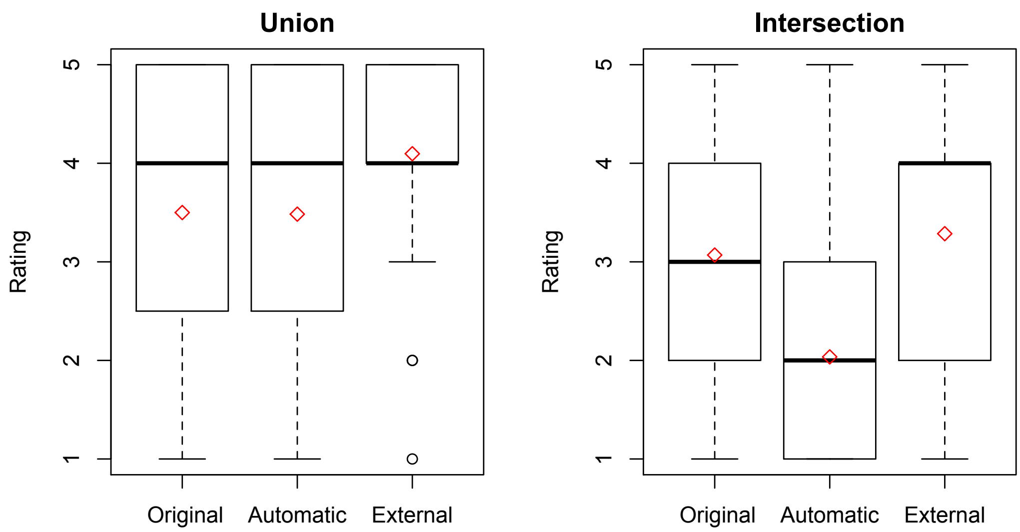


Figure 4.2: Ratings for the relevant turns for each annotation technique, given the Union and Intersection conditions. Red diamonds represent the mean.

In order to understand if there was a significant difference among the ratings, we applied a Kruskal-Wallis test to the Union and Intersection groups. For the union group, the test did not reveal any significant difference between ratings ($\chi^2 = 3.52, p = 0.17$). Post-hoc analyses using Mann-Whitney U tests to the ratings of the different techniques confirmed this ($U = 415, Z = -1.36, p = 0.18; U = 537, Z = 0.35, p = 0.74; U = 643, Z = 1.83, p = 0.07$).

Interestingly, the Kruskal-Wallis test revealed a significant difference among ratings in the Intersection group ($\chi^2 = 10.55, p < 0.01$). Post-Hoc analysis showed a significant difference at $p < 0.01$ between ratings of the Original and Automatic techniques ($U = 596.5, Z = 2.81$) and between ratings of the External and Automatic Techniques ($U = 596.5, Z = 2.82$). However, we found no significant difference between the Original and External techniques ($U = 407, Z = 0.21, p = 0.83$). From this analysis we conclude that participants in the Intersection group considered that the relevant turns provided by the Automatic technique were worse than those provided by the other two techniques.

Further observation of Figure 4.2 shows an unexpected difference in the distributions of the ratings assigned to the Automatic technique between the Union and Intersection conditions. This difference is significant ($U = 729.5, Z = 3.93$, at $p < 0.01$), and given that the different conditions should only affect the ratings of the External annotator, we further investigated this difference. We found that even though the calls in $P2$ were chosen at random, there was a difference in the performance of the Automatic algorithm in the calls chosen for the Union condition (F-score=0.44) and those chosen for the Intersection condition (F-score=0.34). (The F-score was measured with respect to the Original annotations provided by the callers, the Ground Truth.) Although this difference is marginally nonsignificant ($U = 442.5, Z = 1.93, p = 0.06$) we hypothesize it could have played a role in the observed decrease of average rating.

Note that the automatic algorithm extracted many more relevant turns when compared to the Original and the Intersection techniques (see Figure 4.1). Hence, it seems plausible that participants who rated the relevant turns in the Intersection condition found the Automatic turns to be too verbose when compared to the other two techniques (Original and Intersection). This effect did not happen in the Union condition, as the annotations provided by the Union were closer in number to the annotations provided by the Automatic algorithm. Though not conclusive, this is supported by a weak yet significant correlation between the F-score and the rating across the entire dataset ($\rho = 0.22, p < 0.01$).

Role of the order of presentation

The order in which the results of the different techniques were presented to the participants appears to have influenced participants' opinions. Figure 4.3 is an overview of how the relevant turns were rated depending on which order they were presented to participants.

The figure suggests that participants from the Union group were consistent in rating the different annotation approaches independently of the order they were shown. On the other hand, the Intersection group appears to be less consistent. The ratings provided for the second and third annotation strategies presented to participants for a given call were, in average, lower than for the one presented in first place ($U = 548.5, Z = 2.03; U = 540, Z = 1.9$, both at $p < 0.05$). Even though we adequately shuffled the order in which we presented the results of the different annotation approaches to the participants, one third of the times they received contrasting sets in terms of annotation size, i.e. the Automatic with a high amount of turns followed by the Intersection or Original with a much lower amount of turns. Such a contrast

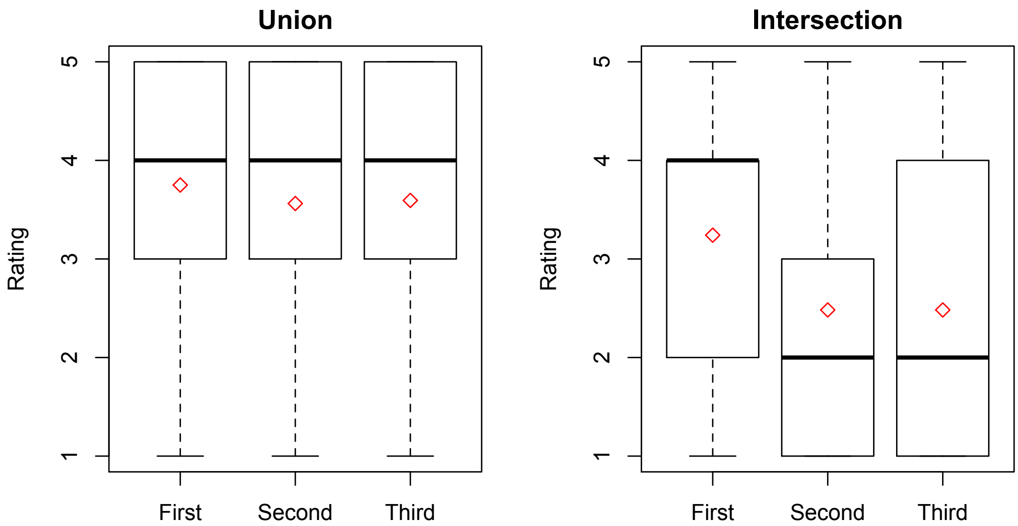


Figure 4.3: Ratings for the relevant turns depending on the order they were presented to participants, given the union and intersection conditions. Red diamonds represent the mean. The difference between the First position between Union and Intersection groups is nonsignificant ($U = 564.5, Z = 1.5, p = 0.14$)

might have biased participants towards giving lower ratings to the second technique (comparative effect). For the Third position, participants did kept similarly low ratings; no significant difference was then found between the Second and Third places ($U = 433, Z = 0.2, p = 0.86$).

In general, the Automatic technique seems to have performed rather poorly, at least when its results were presented inside the Union group. The next section we further explains these findings with the help of the rankings provided by the participants.

Comparison of annotation techniques

As explained in the Measures section, in addition to individual ratings of the annotations, we were interested in directly comparing the 3 annotation techniques to answer our second research question. Therefore, for each call we showed participants the three sets of relevant turns, side by side, and asked them to chose the best and the worst ones.

In the Union group, the External technique was considered to be the best, and the Automatic technique the worst. For the Intersection group, the best tehcnique was the Original and the worst was the Automatic again. We performed a Chi-square "goodness of fit" test to check if the results were significantly different from an equal probability of all the possible outcomes. We found a significant difference for both the Union (best: $\chi^2 = 36.75, p < 0.05$; worst: $\chi^2 = 9.75, p < 0.05$, $df = 2$ for all) and the Intersection groups (best: $\chi^2 = 8.9, p < 0.05$; worst: $\chi^2 = 10.76, p < 0.05$, $df = 2$ for all).

In order to have a clearer view of the results, we added all the times each technique was "voted" to be the best, and subtracted the times it was voted to be the worst. This overall view of the participants' preferences is shown in Figure 4.4.

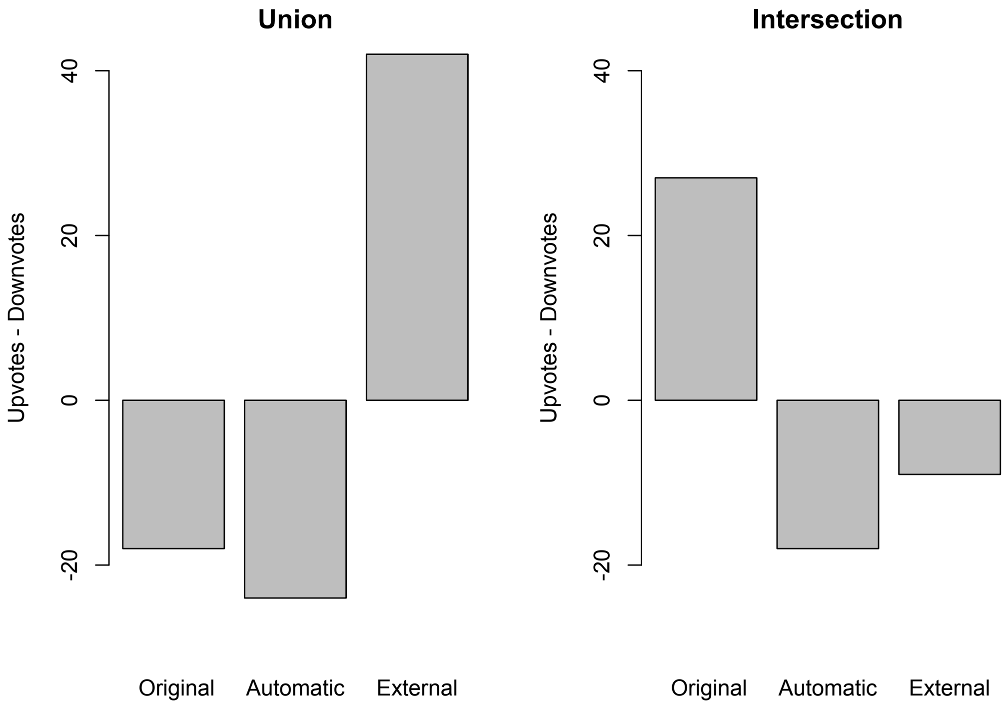


Figure 4.4: Voting results for the three different techniques, given the Union and Intersection conditions.

4.8 Discussion

As mentioned previously, inter-coder agreement between the three annotators, including the original participants and the two external annotators was low. This aligns with similar findings in studies on note-taking during meetings [27] and text summarization [24]. Low inter-coder agreement suggests that the task of determining what is the important information inside an informal phone call, as it happens in other collaborative contexts, is highly subjective.

Subjectivity is also supported by the influence that the *presentation order* had in the ratings assigned to the annotations (see Figure 4.3). This result is an indication of the presence of a contrast effect.

Interestingly, for an important proportion of calls (67.2%), the information that the participants highlighted during the first phase of the study did not satisfy the participants' annotation needs. A temporal effect is a possible explanation for this. When a person has recently made a call, she might be focused on certain pieces of information—for instance those closely related to the purpose of the call—while ignoring other pieces that might be important as well. Given his/her lack of knowledge of the call context, a third party could offer an unbiased judgement that could be useful for spotting noteworthy information from a neutral perspective. This was supported by answers the members of the Union group of our study, as the majority chose the External annotation method as the best. In other words, the external annotators might have been more objective when identifying the important information.

But the results also show that the process of selecting what is important must not be too strict. When comparing the original relevant turns with the intersection of the external annotators' turns, participants preferred their own notes. The intersection of the external annotators' turns probably produced an excessively distilled set of relevant turns, which ended up being too few to satisfy the participants' needs. For instance, a user criticized: *"It lacks information"*, while another complained of a missing piece of important information: *"Some data is missing... like the position in the train"*.

Finally, the automatic algorithm for the extraction of relevant information—in the shape of turns—did not produce satisfactory results in general. It provided excessive information in some cases, (*"I think there are more notes than necessary (...)"*), and it was not as precise as desired (*"It shows the trivial part of the conversation. That is not important information."*) resulting in being ranked as the overall worst of the three methods. Given the challenging nature of the automatic note-taking problem, we think the poorer performance when compared to the manual information extraction approaches was to be expected. The continuous advances in natural language processing and machine learning technologies lead us to think that we will be able to build automatic noteworthiness prediction models able to provide acceptable results.

To sum up, one could imagine a function expressing the user satisfaction level with an annotation service against the amount of notes provided. It would probably have a bell shape. Too selective notes (as those provided by the Intersection of the External annotators) would miss some information. As the amount of notes increase, satisfaction increases. But after a central point of maximum satisfaction, as the amount of annotated information continues to increase, it begins to overwhelm the user, decreasing the satisfaction level again. As the content of the notes approximates the total information contained in the call, the notes get useless. The central, highest point of the curve is where any annotation technique should aim for.

4.9 Implications for design

Annotations are not Personal after all. Surprisingly, the External annotations obtained the highest mean rating and were chosen as the best annotation technique by most of the participants in the Union group. Hence, it seems plausible that a crowd-sourced system could be used to detect noteworthy information in mobile phone calls. It is of key importance to notice the privacy concerns involved if such solution is considered. Opt-in policies on a per-call basis and proper anonymization would be of key importance when sharing with third parties data as intimate as phone call content might be.

It remains as part of future work to determine which refinements of the technique could be made in order to optimize the task, as the intersection of the information selected by the annotators seemed to miss important pieces. Fine tuning of the number of annotators and the combinations of unions an intersections might be the key for improving the performance.

The Ground Truth might not be so true. The annotation behavior of the Original Annotator (caller) should neither be disregarded nor be taken as the ground truth. Factors such as time and context might influence a person information needs after a call has been made, leading him/her to chose poorly in certain situations. Our findings suggest that an optimal annotation solution would take into account contextual information of the call (as stated in [9]), would learn from the user's manually taken notes, and would also offer to the user annotations obtained by crowdsourcing.

The Challenge of the Automatic System. The Automatic note-taking algorithm obtained good results in off-line evaluations with an F-score similar to that of a human in a related task. However, when judged directly by human assessors in terms of satisfying note-taking needs, the Automatic system has done significantly worse than the annotations generated by humans. An interesting possibility that our study suggests is to use the context-free external annotators' relevant turns as an input for the learning phase of the algorithm. We are considering this for future work.

But in the search for a system that can automatically detect noteworthy material from phone calls an important challenge remains. It lies in the observation that the annotations made by the original caller cannot necessarily be trusted as the most appropriate ground truth. Who can we trust then? Only future work will help answering this question.

References

[1] I. Arminen and M. Leinonen. "Mobile phone call openings: tailoring answers to personalized summonses". *Discourse Studies* 8.3 (2006), pp. 339–368.

[2] K. Audhkhasi, P. G. Georgiou, and S. S. Narayanan. "Analyzing quality of crowd-sourced speech transcriptions of noisy audio for acoustic model adaptation". *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.* IEEE. 2012, pp. 4137–4140.

[3] S. Banerjee and A. Rudnicky. "Detecting the Noteworthiness of Utterances in Human Meetings". *Proceedings of the SIGDIAL 2009 Conference.* London, UK: Association for Computational Linguistics, Sept. 2009, pp. 71–78.

[4] S. Banerjee and A. I. Rudnicky. "An extractive-summarization baseline for the automatic detection of noteworthy utterances in multi-party human-human dialog". *SLT.* 2008, pp. 177–180.

[5] M. Bernstein, M. V. Kleek, D. Karger, and M. C. Schraefel. "Information scraps: How and why information eludes our personal information management tools". *ACM Trans Inf Syst* 26.4 (2008), pp. 1–46.

[6] F. Bonin, J. San Pedro, and N. Oliver. "A Context-Aware NLP Approach For Noteworthiness Detection in Cellphone Conversations". *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.* Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 25–36.

[7] A. Bothin and P. Clough. "Participants ' personal note-taking in meetings and its value for automatic meeting summarisation". *Information Technology and Management* 13.1 (2012), pp. 39–57.

[8] A. Bothin, P. Clough, and G. Bradley. "Quantitative analysis of individual differences in Note-Taking and talking behaviour in meetings". *Proceedings of IADIS Multi Conference on Computer Science and Information Systems.* Freiburg, Germany: IADIS Multi Conference on Computer Science and Information Systems, 2010.

[9] J. P. Carrascal, R. de Oliveira, and M. Cherubini. "A Note Paper on Note-taking: Understanding Annotations of Mobile Phone Calls". *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services.* MobileHCI '12. San Francisco, California, USA: ACM, 2012, pp. 21–24.

[10] J. Cohen et al. "A coefficient of agreement for nominal scales". *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[11] B. J. Cragun. "Telephony annotation services. US Patent. US 7602892". Patent US 7602892 (US). Oct. 13, 2009.

[12] W. Geyer, H. Richter, and G. D. Abowd. "Towards a Smarter Meeting Record–Capture and Access of Meetings Revisited". *Multimedia Tools Appl.* 27 (3 Dec. 2005), pp. 393–410.

[13] G. R. Hayes, S. N. Patel, K. N. Truong, G. Iachello, J. A. Kientz, R. Farmer, and G. D. Abowd. "The personal audio loop: Designing a ubiquitous audio-based memory aid". *Mobile Human-Computer Interaction-MobileHCI 2004.* Glasgow, Scotland: Springer, 2004, pp. 168–179.

[14] K. Hinckley, S. Zhao, R. Sarin, P. Baudisch, E. Cutrell, M. Shilman, and D. Tan. "InkSeine: In Situ search for active note taking". *Proc. of CHI '07.* San Jose, California, USA: ACM, 2007, pp. 251–260.

[15] C. Hu, B. B. Bederson, and P. Resnik. "Translation by iterative collaboration between monolingual users". *Proceedings of Graphics Interface 2010.* Canadian Information Processing Society. 2010, pp. 39–46.

[16] I. Hutchby and S. Barnett. "Aspects of the sequential organization of mobile phone conversation". *Discourse Studies* 7.2 (2005), pp. 147–171.

[17] M. Kam, J. Wang, A. Iles, E. Tse, J. Chiu, D. Glaser, O. Tarshish, and J. Canny. "Livenotes: a system for cooperative and augmented note-taking in lectures". *Proc. of CHI '05.* Portland, Oregon, USA: ACM, 2005, pp. 531–540.

[18] M. V. Kleek, W. Styke, and D. Karger. "Finders / Keepers : A Longitudinal Study of People Managing Information Scraps in a Micro-note Tool". *Proceedings of the 2011 annual conference on Human factors in computing systems.* Vancouver, BC: ACM, 2011, pp. 2907–2916.

[19] S. Koço, C. Capponi, and F. Béchet. "Applying multiview learning algorithms to human-human conversation classification". *INTERSPEECH.* 2012.

[20] E. Laurier. "Why people say where they are during mobile phone calls" (2001).

[21] K. O'Hara, M. Perry, A. Sellen, and B. Brown. "Wireless world". Ed. by B. Brown and N. Green. New York, NY, USA: Springer-Verlag New York, Inc., 2002. Chap. Exploring the relationship between mobile phone and document activity during business travel, pp. 180–194.

[22] A. Piolat, T. Olive, and R. T. Kellogg. "Cognitive effort during note taking". *Applied Cognitive Psychology* 19.3 (2005), pp. 291–312.

[23] H. Sacks, E. A. Schegloff, and G. Jefferson. "A simplest systematics for the organization of turn-taking for conversation". *Language* (1974), pp. 696–735.

[24] G. Salton, A. Singhal, M. Mitra, and C. Buckley. "Automatic text structuring and summarization". *Information Processing & Management* 33.2 (1997), pp. 193–207.

[25] A. Weilenmann. "“I can’t talk now, I’m in a fitting room”: formulating availability and location in mobile-phone conversations". *Environment and Planning A* 35 (2003), pp. 1589–1605.

[26] S. Whittaker, P. Hyland, and M. Wiley. “FILOCHAT: handwritten notes provide access to recorded conversations”. *Proc. of CHI ’94*. Boston, Massachusetts, United States: ACM, 1994, pp. 271–277.

[27] S. Whittaker, R. Laban, and S. Tucker. “Analysing meeting records: An ethnographic study and technological implications”. *Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 101–113.

[28] S. Whittaker, S. Tucker, K. Swampillai, and R. Laban. “Design and evaluation of systems to support interaction capture and retrieval”. *Personal Ubiquitous Comput.* 12 (3 Jan. 2008), pp. 197–221.

[29] L. D. Wilcox, B. N. Schilit, and N. Sawhney. “Dynomite: a dynamically organized ink and audio notebook”. *Proc. of CHI ’97*. Atlanta, Georgia, United States: ACM, 1997, pp. 186–193.

Chapter 5

Conclusions

In this dissertation, we studied the valuation of personal information in two interaction contexts, web browsing and mobile communication. In the context of web browsing, we studied the monetary valuation of Personally Identifiable Information, or PII. In the context of mobile communication, we studied the process of annotation of mobile phone calls, i.e., the selection of the most important information from calls made by mobile phone users. Next we discuss the conclusions of the work herein presented, in relation research questions stated in the introduction of the thesis.

5.1 Conclusions In relation to Research Questions

Valuation of personal information in web browsing

Question 1: What monetary value do Internet users assign to their personal information while being online?

Beyond the actual average values that web users assign their PII, we consider specially interesting the comparative values between different PII types and quantities. It is worth noticing that users tend to value PII related to their offline identity higher than the PII related to their online identity. Valuation of items such as age, salary, phone number and address was consistently valued higher than other types of online PII. At the same time, valuation of PII more closely associated to offline identity, such as that related to Social or Finance websites, was valued higher than PII related to Search or News websites.

This indicates a disconnection between web user's perception of their *offline* and *online* privacy. It appears that privacy-related implications are easier to foresee and cause doubt in users when sharing their offline identity information, as it is more easily linked back to their personal identity. On the other hand, online PII is probably perceived to be easier to separate from personal identity. For web users, it is not evident that they can be identified by their browsing behavior alone. However, previous studies have provided evidence of the opposite [9].

Our work also indicates that web users lack awareness and understanding of the use that Internet-based companies make of their PII. In consequence, and unsurprisingly, they are not comfortable with the monetization of their PII. This can not be only attributed to the users' lack of interest in understanding the inner workings of the Internet economy, but probably also to the obscurity—intentional or not—of Internet based companies' ways to monetize their users' PII for supporting their business. We think that Internet companies should be more transparent in explaining to their users the costs involved in offering what is normally presented as "free" services. This might effectively contribute to reduce the distrust of web users, and to create a healthier climate for the information economy.

On the users' side, privacy has shown to be more an expressed concern than a real motivation for taking action [8]. Self-regulation is required from users, but incentives might be necessary to "nudge" them towards better privacy-protecting behavior. Our findings suggest that money is the primary good that users would be willing to accept in exchange for their PII, followed by the improvement of current services and the option to receive access to more free services. These findings back up the feasibility of transactional privacy and the creation of online personal information markets [13].

Valuation of personal information in mobile communication

Question 2: What factors influence the mobile phone users' need of taking notes?

We found that the most important factors that influence users to take notes from mobile phone calls are related to the call content itself. Proper names, numbers and question adverbs seem to correlate with the note taking behaviour. Additionally, some contextual factors such as the purpose of the call had some influence as well. Furthermore, we observed that during mobile phone calls, important information is constructed in a collaborative way, as opposed to being an unidirectional flow of information determined by the initiator of the call. Information exchanged during phone mobile phone calls then has potential for actions such as sharing, scheduling meetings and appointments, or adding up information to personal information archives.

Our findings suggests opportunities and provide insights for the development of automatic note-taking applications. For instance, an annotation system should leverage sensors for establishing call context, and process the contents of phone calls to detect patterns that might predict the need to take notes. As shown by the work of Bonin et al. [2], Machine Learning-based techniques can be used for this.

It makes sense to use mobile phones as the main tools for the task of automatic annotation. The main reason for this is that this is the most frequently used tool for taking notes during phone calls—at least for smartphone users. The second reason is that modern mobile devices have evolved to the point of being very capable devices with enough processing power for this kind of tasks. And finally, because it is the very same device that is being used for communication, avoiding the need to have more tools at hand.

Even though mobile phones are being used for much more than just making phone calls, voice communication is still one of the fundamental uses. This not only includes standard phone calls, but also network-based voice communication applications and voice messaging. We believe that our findings can be applied to these context as well, helping to advance the possibilities for using voice as a source of non-ephemeral data, whenever that is needed.

Question 3: Is valuation of personal information constant or variable across time?

When having a phone conversation, users tend to focus on the specific topics that initiated the process. If they review the information, they would consider that the pieces of information related to those topics are the most important, because that is what is fresh in their minds. After some time, however, the users frequently find that some of the remaining interactions contain other pieces that are also important, and which were overlooked when the first assessment was made. This can be noticed only when users shift their focus from the initial topics, something that can only happen when looking at the information in perspective after some time.

This is a phenomenon worth to be taken into account. Previously suggested solutions for the preserving information from conversations [6, 7, 5] focus on capturing the importance of information in real-time. While this seems to be a good approach, it is subject to the influence of the initial topic of the conversation. This influence is so strong that it is easy to let pass some unrelated pieces of information that can have importance and that are worth to be preserved for later use. With that approach, there is no way to re-asses the information *a posteriori*, under different light, with a different mindset.

Our findings imply that in order to avoid missing information from a real-time communication activity—a phone call, a chat, a meeting—, the user should be able to perform multiple *important assessment stages*: one could be done in real-time or just after the communication has taken place, and subsequent ones could be done at a later time. It is likely that a user can find more important information from several revisits. In other words, information from communication activities has an "afterlife": initial consumption of information is a just a step, not the end of the cycle. Users often may need to keep certain pieces of information for later use, even when at the moment they appear to be not so important.

With storage capacity and price being becoming less of a limitation[1], keeping large amounts of data is not a problem anymore. Big Data is in some way a consequence of this "keep everything" attitude. On a personal level, some practitioners now advocate for the quantification of the self [11, 12, 1] as a way to improve life quality. There are already some examples of this trend applied to personal communication [15]. Our findings align with these approaches, suggesting further ways to analyze personal communication data, and shedding some light on how to take a semi-automatic approach to PIM in the context of mobile communication.

Information valuation, as an important aspect of Personal Information Management, is not static at all. We consider that these dynamics should be taken into account not only for mobile phone call annotation solutions, but for general tasks that imply the assessment of value of personal information.

Question 4: Is personal information—in the case of mobile call transcripts—value "unanimous"? How does valuation change from the perspective of a neutral or context-free observer?

Our results on the evaluation of the three annotations methods suggest that automatically detecting phone noteworthy information from mobile phone calls and their metadata is a challenging task. One of the reasons for this is that there is a high subjectivity involved in determining what is the most important—i.e. noteworthy—information from calls. As seen in previous stages of our research, the concept of "important information" varies across time. Therefore, what is considered to be important at some point in time, might be not important at a later time and vice versa. This suggests that in the context of informal phone calls, the ground truth tends to be fuzzy and hard to rely on for training a Machine Learning algorithm.

[1]http://ns1758.ca/winch/winchest.html

Additionally, we observed that different people would hardly agree in what is noteworthy information from a phone call transcript. A phone call includes lots of information that is merely conversational and less informational, which probably adds complexity to the identifying the relevant information. This also means that the concept of information importance differs widely from person to person.

Despite this, our results suggest that different people might spot different relevant pieces of information, and joining all the pieces might be a good approximation to what the end user needs to annotate. We therefore see potential in the use of crowdsourcing for selecting important or noteworthy information from mobile phone calls.

5.2 Limitations

We found the average values that Internet users assign to their PII when browsing the web. It is not easy to say if such values are low or high, as this is highly subjective. For instance, we found that users value information related to their browsing behavior—clicks made on a certain website, for instance—at around 7€. This value is comparable, at the time of the study, to the price of a Big Mac meal in Spain[2]. Whether this value is high or low is relative.

Our study of economic valuation of online PII was conducted in Spain between July and September of 2011. This framing limits the scope of our findings, as the attitudes towards privacy vary depending on social, historical and cultural factors. It is possible that similar studies conducted in other geographical and cultural contexts might lead to different results.

Furthermore, it is likely that the media coverage of news related to privacy compromising technologies and practices in latest years[3] might have affected global awareness on privacy issues in general, and in particular with respect to online privacy. This might have undermined the trust of users for companies that are known to store personal information for business purposes. If this is indeed true, the values obtained by our study might be different if measured again today.

With respect to our study on mobile phone call annotation, it should be noticed that due to the study design, all the participants were smartphone (iPhone or Android) users. This might have biased some of our results, but it is worth noticing that the increasing adoption of smartphones[4] confirms the relevance of the study.

An additional limitation of our study design is that our dataset is comprised of information provided exclusively by the callers, unfortunately missing the point of view of the callee. Further research should be done to complement our findings while taking into account this important aspect.

5.3 Future work

With respect to personal information valuation during web browsing

Our work has shed light on the monetary valuation of PII during web browsing, but in an increasingly connected world, personal information is collected during many other

[2] Hence the name of the paper we published on the subject: "Your browsing behavior for a Big Mac: Economics of Personal Information Online"

[3] E.g. http://www.businessweek.com/articles/2013-06-07/what-you-need-to-know-to-understand-the-nsa-spying-scandal

[4] http://www.pewinternet.org/2013/06/05/smartphone-ownership-2013/

activities. Only considering mobile devices, phone calls, sensors, location and mobile app usage are just a few examples of personal information sources. Some previous work has already looked into valuation of location information [3, 4]. Staiano et al. [14] have gone further, studying other mobile phone information sources. We think there is still work to be done in this line of research.

Furthermore, new interaction technologies will require new studies on personal information valuation. The emerging use of wearable technologies implies the use of devices in closer proximity, possibly during longer periods of time and less noticeably. Conversely, the so-called Internet of Things is meant to provide processing power and connectivity to lots of objects around the user. Both technologies are meant to highly pervasive access to the users' lives, rising new research questions in relation to privacy and personal information valuation.

We have shown the potential for the use of the Experience Sampling methodology for the study of private information valuation in the context of web browsing. Future research on privacy valuation in these emerging contexts could make use of this methodology as well.

With respect to personal information valuation in mobile communication

Even when mobile phones are being used for much more than just making phone calls [10], voice data is widely used in VoIP and teleconferencing applications, among other applications. We think that our longitudinal study on phone call annotation might be applied to those communication scenarios, and even maybe in some text-based contexts. Further studies are needed to verify whether this is indeed true. Future research should focus on understanding the factors that are more prone to produce noteworthy material when having conversations in those scenarios, as well as the dynamic aspects of the information valuation.

Developing a complete automatic annotation tool is still left to be done. Even if noteworthy information could be detected, an ideal solution would use natural language processing to create new meaningful notes for later access. It should also be able to connect to usual applications such as calendars, contact lists or social networking apps, in order to provide a complete personal information management solution. Further work is needed, including prototype development and evaluation.

Finally, it is necessary to investigate the privacy issues involved in exploiting audio data from conversations as a source for indexable data. Telephones have been been in use for well over a hundred years, and phone users are not used to their calls being permanently or temporarily recorded. In fact, phone calls are protected by well established telephone recording laws[5]. That is not the case with technologies such as email, where everything is preserved by default. In consequence, privacy concerns are much higher for phone calls than for more recent communication methods. Further research should focus on finding ways to access phone call data while preserving users' privacy.

References

[1] *About the Quantified Self.* http://quantifiedself.com/.

[5]E.g. http://www.fcc.gov/guides/recording-telephone-conversations

[2] F. Bonin, J. San Pedro, and N. Oliver. “A Context-Aware NLP Approach For Noteworthiness Detection in Cellphone Conversations”. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 25–36.

[3] D. Cvrcek, M. Kumpost, V. Matyas, and G. Danezis. “A study on the value of location privacy”. *Proceedings of the 5th ACM workshop on Privacy in electronic society - WPES ’06*. New York, New York, USA: ACM Press, 2006, p. 109.

[4] G. Danezis, S. Lewis, and R. Anderson. “How Much is Location Privacy Worth?” *Online Proceedings of the Workshop on the Economics of Information Security Series (WEIS) 2005*. 2005.

[5] L. Degen, R. Mander, and G. Salomon. “Working with audio: integrating personal tape recorders and desktop computers”. *Proceedings of the SIGCHI conference on Human factors in computing systems*. CHI ’92. Monterey, California, United States: ACM, 1992, pp. 413–418.

[6] G. R. Hayes, S. N. Patel, K. N. Truong, G. Iachello, J. A. Kientz, R. Farmer, and G. D. Abowd. “The personal audio loop: Designing a ubiquitous audio-based memory aid”. *Mobile Human-Computer Interaction-MobileHCI 2004*. Glasgow, Scotland: Springer, 2004, pp. 168–179.

[7] D. Hindus and C. Schmandt. “Ubiquitous audio: capturing spontaneous collaboration”. *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*. CSCW ’92. Toronto, Ontario, Canada: ACM, 1992, pp. 210–217.

[8] C. Jensen, C. Potts, and C. Jensen. “Privacy practices of Internet users: Self-reports versus observed behavior”. *International Journal of Human-Computer Studies* 63.1-2 (2005), pp. 203–227.

[9] R. Jones, R. Kumar, B. Pang, and A. Tomkins. “”I Know What You Did Last Summer”: Query Logs and User Privacy”. *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. CIKM ’07. Lisbon, Portugal: ACM, 2007, pp. 909–914.

[10] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. “A survey of mobile phone sensing”. *Communications Magazine, IEEE* 48.9 (2010), pp. 140–150.

[11] I. Li, A. K. Dey, and J. Forlizzi. “Understanding my data, myself: supporting self-reflection with ubicomp technologies”. *Proceedings of the 13th international conference on Ubiquitous computing*. ACM. 2011, pp. 405–414.

[12] I. Li, Y. Medynskiy, J. Froehlich, and J. Larsen. “Personal Informatics in Practice: Improving Quality of Life Through Data”. *CHI ’12 Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’12. Austin, Texas, USA: ACM, 2012, pp. 2799–2802.

[13] C. Riederer, V. Erramilli, A. Chaintreau, P. Rodriguez, and B. Krishnamurty. “For sale: Your data, By: you”. *ACM Hotnets*. 2011.

[14] J. Staiano, N. Oliver, B. Lepri, R. de Oliveira, M. Caraviello, and N. Sebe. "Money Walks: A Human-centric Study on the Economics of Personal Mobile Data". *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '14. Seattle, Washington: ACM, 2014, pp. 583–594.

[15] S. Wolfram. *The Personal Analytics of My Life*. http://blog.stephenwolfram.com/2012/03/the-personal-analytics-of-my-life/.

ANNEXES

ANNEX I. Additional publications related to Chapter 3

This annex includes a paper an a patent that were published as part of the "Call-Notes" project:

J. P. Carrascal, R. de Oliveira, and M. Cherubini. "A Note Paper on Note-taking: Understanding Annotations of Mobile Phone Calls". *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services*. MobileHCI '12. San Francisco, California, USA: ACM, 2012, pp. 21–24

R. De Oliveira and J. Carrascal. *Method for collecting and storing annotations associated to a voice audio data.* US Patent 8,792,863. July 2014

A Note Paper on Note-Taking: Understanding Annotations of Mobile Phone Calls

Juan Pablo Carrascal*
Universitat Pompeu Fabra
jp.carrascal@upf.edu

Rodrigo de Oliveira
Telefonica Research
oliveira@tid.es

Mauro Cherubini*
Google
mauroc@google.com

ABSTRACT
Note-taking has been largely studied in contexts of work meetings. However, often people need to remember information exchanged in informal situations, such as during mobile phone conversations. In this paper we present a study conducted with 59 subjects who had their phone calls semi-automatically transcribed for later annotation. Analysis of the 621 calls and the subjects' annotation behavior revealed that phone recall is indeed a relevant user need. Furthermore, identifying patterns in phone calls such as numbers and names provide better indicators of annotation than variables related to the callers' profile, context of calls, or quality of service. Our findings suggest implications for the design of mobile phone annotation tools.

Author Keywords
phone call; annotation; context; mobile information

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces (D.2.2, H.1.2, I.3.6)

INTRODUCTION
Lots of information is exchanged every day during mobile phone calls. While a consistent part of this information could be ephemeral as supporting our social needs, another part of it might be important to remember as being functional to our lives. For example, we might receive a phone call to remember to buy some groceries on our way back home, to pick kids from school because our partner is busy, or we might agree to meet a colleague at a specific time in a restaurant.

It can be speculated that most people have a good memory for remembering important details shared during phone calls, while others require taking notes in order to avoid forgetting. To date however, there is little support of applications for this latter case and it is frequently experienced by many the cumbersome situation of having hands tight and needing to take notes while on the phone. Perhaps, this is due to the lack of studies that have tackled this issue.

*Research conducted while working for Telefonica Research.

MobileHCI'12, September 21–24, 2012, San Francisco, CA, USA.

Many studies in the past have focused on how people take notes during work-related meetings [1, 7, 3, 8]. However, how people take notes during daily mobile phone calls it is a different and less studied topic. More specifically, we have little evidence of what is important to remember during mobile conversations and which factors, whether contextual or demographic, play a role in this process.

RELATED WORK
In the recent years, scholars focused extensively on work-related meetings, during which participants also have the need to take annotations. Mobile phone calls can be considered as a two-people meetings but they are different from formal work-related gatherings. Typical mobile phone calls tend to be relatively short in time when compared to work meetings. Also, while in mobile phone calls parties tend to quickly share information with ad-hoc formalisms between greetings and salutations, work meetings tend to be rather structured following agendas and action points.

People normally take notes during meetings and in many other situations of daily life to record important issues and remember things that have been discussed [1]. Geyer and colleagues [3] note how: "personal notes primarily serve as a memory aid for individuals to remember important facts, actions, ideas, and decisions but are hardly useful for persons other than the author". In formal work settings, notes take a structured form and usually include action items. Often also formal minutes of the meetings are recorded to create a shared group memory and to make the meeting more efficient [3].

Despite the rise of Information Capture and Retrieval (ICR) technology for collaborative exchanges, the most common recording techniques used in meetings today are still pens, paper or notebooks, and sometimes laptops [7]. More sophisticated techniques comprise the recording of the audio and sometimes the audio-video trace of the meeting [3]. However, the drawback of these techniques is that they require a timestamped indexing of the content to alleviate retrieval of relevant information.

Notes that people take for themselves during meetings contain "personally important points and in particular details on action items that the note-taker needs to deal with later" [1]. Personal notes usually mention decisions, names, dates and actions [7]. Whittaker and colleagues [7] analyzed the content of people's notes and found that in 30% of the cases personal notes concerned comments that could supply context for action.

Whittaker et al. [6] have studied note taking as the primary way of recording what occurred during a meeting. Subjects in their study reviewed regularly their notes after the meeting (33%). The large majority of people in their sample (70%) reported difficulty in taking notes due to various reasons, like the failure to note facts that turned out to be important later, illegible names, lack of time, and notes that lacked the right level of summarization for *a posteriori* understanding. One of the most significant problem with personal notes is that taking notes reduces the ability of people to participate in the conversation [7].

ICT systems developed in the past aimed at addressing some of these issues. Hindus & Schmandt [4] presented a system that allowed people to mark interesting portions of an ongoing telephone conversation. Degen et al. [2] modified a hand-held tape recorder so that users could mark the audio while it was being recorded. Wilcox et al. [8], designed a system called Dynamite, which allowed users to attach keywords to recorded notes so to create an index of the content. Whittaker et al. [5] looked as well at how to improve voicemail by allowing users to visually inspect its content and enabling annotations.

All these findings come from studies of work meetings while little work has focused on understanding the role of annotations as support of daily mobile phone conversations. With this work we aim at shedding some light in this area by highlighting *annotation habits*, an initial overview of *how* people make these annotations, and some evidences of *problematic aspects* related to annotations made during mobile calls.

METHODOLOGY

We designed a quantitative experiment to collect a large sample of mobile phone calls, their annotations –if any–, and contextual parameters at the time of the calls.

Participants

A total of 59 subjects (41 male) participated actively in the user study, *i.e.* answered the pre-study questionnaire and made at least one phone call during the study. Their mean age was 31.05 years ($s = 7.4$), they were all living in Spain and reported being fluent in speaking Spanish –a requirement of the study. The sample was well geographically distributed (37 unique cities) and included only subjects that had at least the basic education (primary school: 3.4%), followed by 3.4% who finished secondary school, 78% that concluded technical school or obtained a bachelor degree, and 15.2% who had either a masters or a doctorate degree. The reported annual income suggests that all social classes were represented in the sample (27%, 19%, 25%, 20%, and 9% earned up to €10K, €20K, €30K, €40K, and €40K+ a year respectively).

Procedure

The study spanned over 50 days. It was conducted amongst participants who voluntarily registered after following advertisements in popular Web portals in Spain. We had to limit participation to people who owned and used an iPhone or Android mobile, since we only had VoIP applications available for these platforms. We invited participants via email, asking them to fill a registration questionnaire which, besides general demographics, asked relevant questions such as calling habits and general note-taking habits during phone calls.

We offered participants free calls to mobile or fixed phone lines inside the Spanish territory. To be able to make free calls, they had to install a VoIP application on their mobile phones and configure it to connect through our servers. Participants were allowed to make calls either using the VoIP application–in which case they would contribute it to our study–or the native phone application so as to prevent us from having access to the conversation. We explicitly explained them how to switch between these two options. Whenever participants used our VoIP application, a short message was played to both the caller and the callee informing them that the call was going to be recorded and transcribed. Transcriptions were first generated automatically and later manually inspected and corrected by an expert before being presented to participants.

We developed a web application which allowed participants to interact with call data. It displayed a list of phone calls, and for every call, its transcription, date, duration, status and callee number, as well as a related questionnaire. Participants were asked to:

1. Enter the web application and confirm that they wanted to contribute the calls. Participants were allowed to delete a call within a 24 hour period if they considered it to have sensible content. They could review call data to assist the decision.
2. To select and highlight pieces of text that they considered important, or worth remembering, from the transcriptions. In case there was no important text, to explicitly declare so.
3. To answer a questionnaire related to each phone call.

Measures

The questionnaire associated to each phone call included several questions: 1) Relationship with callee; 2) Who was with the caller at the time of the call; 3) Location of the caller at the time of the call; 4) Objective of the call; 5) Level of importance of the call; 6) Reasons for highlighting text (if applicable); 7) Whether the caller could take written notes a the time of the call; 8) How important was to take notes during the call; and 9) General questions about sound and transcription quality. Given that participants reported objectives of calls in free text, these were manually classified by two coders as either: “discuss topic”, “appointment”, “give/receive information”, “ask favor”, or “social”. Inter-rater reliability was highly acceptable ($K = .81, p < .001$).

RESULTS AND DISCUSSION

Patterns of phone calls and annotation habits in the study were consistent to self-reported data. During the 50 days of the study deployment, participants made 621 calls with a total duration time of 87,035 seconds ($\bar{x} = 140.15$; $s = 191.991$). Quality of the calls was considered acceptable ($q2 = 3$: acceptable, $q1 = 3$: acceptable, $q3 = 4$: good). Transcriptions of the calls and annotations yielded a total of 811,453 characters ($\bar{x} = 1,306.69$; $s = 1889.6$) and 44,744 characters ($\bar{x} = 72.05$; $s = 219.59$) respectively. The average number of calls per participant was 10.53 ($s = 8.69, min = 1, max = 32$), and they annotated an average of 4.61 of their calls ($s = 5.36, min = 0, max = 29$). Hence, about 44% of all phone calls were annotated, which is consistent with the participants' self-reported annotation habits captured by the pre-study questionnaire (34% and 46% indicated taking notes frequently using mobile phones and paper/pencil respectively). Likewise, participants called family members more often than friends, and called friends more often than work colleagues, which reveals the same order reported in the pre-study questionnaire. These findings support consistency between the participants' behavior in the study and how they perceive their behavior in real life.

Phone Recall Needs

Recalling information from calls is a general need and not necessarily a simple task. While 15% of the participants reported that this need rarely occurs, almost half of the sample agreed it occurs sometimes (48%), and over one third indicated it happens frequently (37%). No one reported the absence of this need. When evaluating the easiness to recall information obtained in phone calls, 37% said it is either easy or very easy, 36% reported it is neither easy nor difficult, and the remaining 27% agreed the recall task is at least difficult. These results suggest the importance of supporting recall of phone conversations.

Mobile phone is the primary tool to support recall. Several tools can be used to annotate phone calls. According to our sample, mobile phones and regular paper and pencil are the most important ones–46% and 34% respectively use them frequently for this task. Note that our sample is composed only by smartphone users, hence the popularity of mobile phones as the primary annotation source. Participants reported taking *a posteriori* notes of phone conversations using text-based notepads and audio-based memo applications (25% record audio notes for phone calls at least once a week).

Dynamics of Phone Annotation

If the call is important, expect important notes to be taken. As one might expect, the participants' evaluation of the importance of the calls strongly correlates to the importance of the corresponding annotations ($\rho = .50, p < .001$). This means that the higher one thinks is the importance of a call, the more likely important annotations will be made for it.

If the call is important, callers tend to get prepared for taking notes. While 65% of all calls happened when users had one hand holding the phone and the other hand free (usual setting when attending a phone call), this figure increased to 91% when call annotations were considered important. From this result, we raised the hypothesis that callers would make sure to have their hands free when making important calls because of the higher probability of taking notes during them. Indeed a significant association ($\phi = .11, p = .007$) was found between importance of the call (*i.e.* whether the call had at least some importance) and ability to take notes (*i.e.* whether the participant had at least one hand free to take notes). As there is no off-the-shelf solution for creating hands-free notes during a phone call, apparently people tend to change their behavior right before the call to ensure that at least one of their hands will be free for taking notes.

Transcribing the entire conversation is indeed not efficient. Only 5.5% of the characters in the transcribed calls were highlighted in the participants' annotations. This result is in agreement with previous work in the sense that providing full transcripts of conversations most likely overload users thus switching their recall problem for information retrieval [7].

Annotations have patterns and these are better indicators of note-taking than most of the variables observed in the study, i.e. *contextual, quality of service (QoS) and caller profile.* In the preliminary questionnaire, participants were asked about the kinds of information they find themselves trying to remember after a phone call. About 79% mentioned pieces of information that necessarily include numbers (*e.g.*, phone numbers, dates, prices, addresses) and 34% mentioned information related to names (*e.g.*, addresses, contacts). Therefore, we implemented three parsers to count numbers, names, and interrogative adverbs (*i.e.* why, where, how, when) in the phone call transcriptions respectively. According to Table 1, these pattern variables have significant medium correlations with the length of notes ($\rho = .33, \rho = .31, \rho = .30$ respectively), and these are higher than most correlations with other variables observed in this study–call length is the only exception ($\rho = .41$). In other words, the more numbers, names, and interrogative adverbs are mentioned in a call, the higher the probability to take notes and also the longer the annotations might be.

From the data shown in Table 1, we can highlight at least three interesting findings. First, none of the callers' demographic variables (*i.e.* gender, age, education and income) revealed significant correlations with the annotation variables.

The second interesting result is the fact that QoS variables are positively correlated to taking phone notes. Moreover, quality of call is positively correlated to both the note taking activity ($\rho = .18$) and duration of calls ($\rho = .20$). One possible explanation is that the better the quality of calls, the more time users engaged in a phone conversation, thus increasing the probability of taking notes.

And finally, contextual variables played distinct roles on the note-taking activity. Although the call place and callee information did not reveal any significant relationship with phone annotation, information about the caller's companion, the call objective, and the call length did. Phone calls next to work colleagues were positively correlated with taking notes ($\rho = .11$), probably due to the nature of the call objective (information: $\rho = .11$; appointment: $\rho = .08$). On the other hand, phone calls with no companion were inversely related to generation of notes ($\rho = -.09$). Stronger interaction effects were revealed by call length ($\rho = .35$) and social calls, i.e. calls reported by participants as *"just to chat"* or *"say hello"* ($\rho = -.26$). The former result could indicate that the lengthier the call, the higher chances of taking notes, whereas the latter result is an indication that social calls tend to have significantly fewer annotations.

In summary, the results of this study suggest that taking notes

Table 1. Correlations/Associations between annotation-related variables (Note Taken: Yes/No; Note Length in number of characters) and other variables related to the call task.

Variable	Source	Coefficient*	
		Note Taken	Note Length
Total numbers/call	Pattern	**.29**	**.33**
Total names/call	Pattern	**.28**	**.31**
Total adverbs/call	Pattern	**.25**	**.30**
Gender	Profile	.07	.06
Age	Profile	-.01	-.05
Education	Profile	-.02	-.01
Income	Profile	.02	.02
Recall frequency	Profile	**-.10**	-.07
Recall easiness	Profile	-.04	**-.12**
Note mobile frequency	Profile	-.02	.00
Note paper frequency	Profile	**-.11**	**-.18**
Call quality	QoS	**.18**	**.20**
Transcription quality	QoS	**.12**	**.08**
Call length (chars)	Contextual	**.35**	**.41**
Call length (seconds)	Contextual	**.29**	**.35**
Call who: mate	Contextual	.04	.04
Call who: family	Contextual	-.05	-.06
Call who: friend	Contextual	.04	.06
Call who: work	Contextual	-.03	-.04
Call with: alone	Contextual	**-.09**	**-.09**
Call with: mate	Contextual	.06	.07
Call with: family	Contextual	.05	.03
Call with: friend	Contextual	-.05	-.04
Call with: work	Contextual	**.11**	**.10**
Call from: home	Contextual	.04	.03
Call from: work	Contextual	.05	.05
Call from: commute	Contextual	-.07	-.05
Call why: discuss topic	Contextual	.04	**.08**
Call why: appointment	Contextual	**.08**	.06
Call why: info	Contextual	**.11**	.06
Call why: ask favor	Contextual	**.08**	**.10**
Call why: social	Contextual	**-.26**	**-.23**

* Correlations between ordinal and non-normal interval variables were assessed using Spearman's Rho (ρ). Associations between dichotomous variables were assessed using the χ^2 derived Phi coefficient (ϕ). Coefficients in bold are significant at $p < .05$.

during mobile phone calls is a common need, and that such need is not well satisfied by current off-the-shelf solutions. Participants of our study chose to be prepared whenever they felt that a call was important, leaving at least one hand free for taking notes. The importance of the notes taken was, unsurprisingly, correlated with the stated importance of the call. Advanced smartphones users used their devices more frequently than pen and paper for taking notes. However, the latter method was still widely used.

Another interesting finding was that participants annotated mostly information containing numbers and names, such as phone numbers, addresses, dates, or contacts. These patterns can be easily identified signaling the importance to annotate calls. Further semantical analysis might reveal more complex patterns and potentially refine our conclusions.

IMPLICATIONS FOR DESIGN

Our study provide evidence that users tend to have at least one of their hands free during calls they consider to be important. While being on the go, this probably implies limiting user's mobility. In order to overcome this problem, we emphasize the need to assist users in creating automatic annotations during mobile phone calls. We can speculate that for the callee this need might be even greater given that s/he cannot anticipate important incoming calls. Mobile phones shall be the most suitable devices to enable such solution as they were reported to be the primary phone annotation tools for the participants of our study–all smartphone users. The lack of viable solutions for automatic hands-free annotation of phone calls implies that currently the process must be accomplished offline, after the call has finished, thus increasing the possibility for important pieces of information to get lost.

An important feature for a practical solution is to avoid full call transcription and rather focus on important pieces of information towards preventing the user's information overload. According to our study, these pieces of information usually appear in patterns, such as phone numbers, dates, addresses, prices, shop/to-do lists, contact names, activities, among others. The application should be able to recognize these patterns and annotate them for later recall.

Finally, the mobile application should not only look for patterns in the call, but also leverage its embedded sensors to gather relevant contextual information for the note-taking activity, such as with whom the caller is (alone *vs.* with work colleagues) and what is the objective of the call (social *vs.* non-social). By identifying call context, call QoS–via analysis of the microphone signal, and patterns in the calls, the need to annotate a phone call might be detected and potential annotations inferred.

Acknowledgments. We thank Diego Urdiales, Isabel Ordas, Miguel Angel Santiago, Sergio Romero and all participants of the study. Telefonica Research participates in the Torres Quevedo subprogram (MICINN), cofinanced by the European Social Fund, for researchers recruitment.

REFERENCES

1. Bothin, A., and Clough, P. Participants ' personal note-taking in meetings and its value for automatic meeting summarisation. *Information Technology and Management* (2011), 1–19.
2. Degen, L., Mander, R., and Salomon, G. Working with audio: integrating personal tape recorders and desktop computers. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '92, ACM (New York, NY, USA, 1992), 413–418.
3. Geyer, W., Richter, H., and Abowd, G. D. Towards a smarter meeting record–capture and access of meetings revisited. *Multimedia Tools Appl. 27* (December 2005), 393–410.
4. Hindus, D., and Schmandt, C. Ubiquitous audio: capturing spontaneous collaboration. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, CSCW '92, ACM (New York, NY, USA, 1992), 210–217.
5. Whittaker, S., Davis, R., Hirschberg, J., and Muller, U. Jotmail: a voicemail interface that enables you to see what was said. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '00, ACM (New York, NY, USA, 2000), 89–96.
6. Whittaker, S., Hyland, P., and Wiley, M. Filochat: handwritten notes provide access to recorded conversations. In *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, CHI '94, ACM (New York, NY, USA, 1994), 271–277.
7. Whittaker, S., Tucker, S., Swampillai, K., and Laban, R. Design and evaluation of systems to support interaction capture and retrieval. *Personal Ubiquitous Comput. 12* (January 2008), 197–221.
8. Wilcox, L. D., Schilit, B. N., and Sawhney, N. Dynomite: a dynamically organized ink and audio notebook. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '97, ACM (New York, NY, USA, 1997), 186–193.

US008792863B2

(12) **United States Patent**
De Oliveira et al.

(10) **Patent No.: US 8,792,863 B2**
(45) **Date of Patent: Jul. 29, 2014**

(54) **METHOD FOR COLLECTING AND STORING ANNOTATIONS ASSOCIATED TO A VOICE AUDIO DATA**

(71) Applicants: **Rodrigo De Oliveira**, Barcelona (ES); **Juan Pablo Carrascal**, Barcelona (ES)

(72) Inventors: **Rodrigo De Oliveira**, Barcelona (ES); **Juan Pablo Carrascal**, Barcelona (ES)

(73) Assignee: **Telefonica, S.A.**, Madrid (ES)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 27 days.

(21) Appl. No.: **13/729,235**

(22) Filed: **Dec. 28, 2012**

(65) **Prior Publication Data**

US 2014/0187208 A1 Jul. 3, 2014

(51) **Int. Cl.**
H04W 4/12 (2009.01)

(52) **U.S. Cl.**
USPC **455/412.1**; 379/142.17; 379/202.01

(58) **Field of Classification Search**
CPC H04M 3/533; H04M 15/06; H04M 3/56; H04M 3/51
USPC 455/412.1; 370/142.17, 202.01, 265.03
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

7,602,892 B2 10/2009 Cragun
2011/0206198 A1 * 8/2011 Freedman et al. 379/265.03
2012/0321062 A1 * 12/2012 Fitzsimmons et al. ... 379/142.17

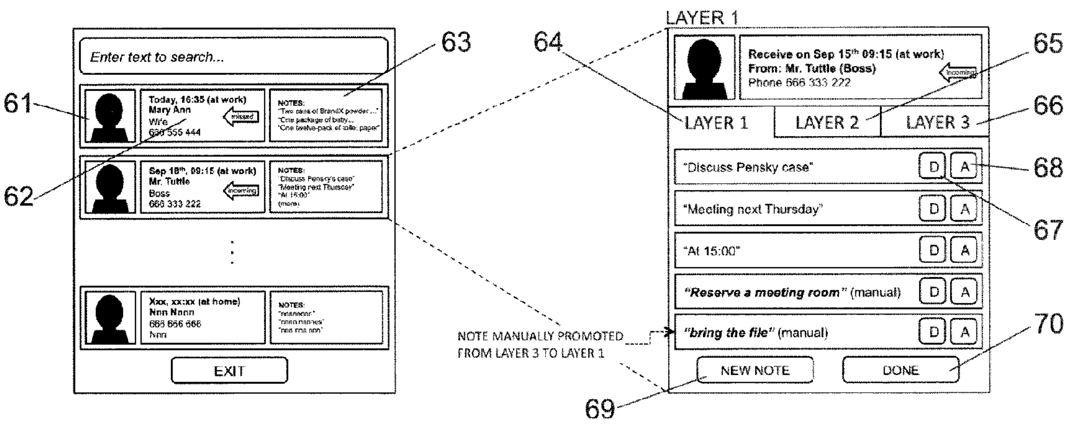
* cited by examiner

Primary Examiner — Phuoc H Doan
(74) *Attorney, Agent, or Firm* — Gary J. Gershik; Cooper & Dunham LLP

(57) **ABSTRACT**

Present invention refers to a method for collecting, organizing and storing annotations associated to a voice audio data. The method comprises providing the voice audio data; transcribing said voice audio data into text data; then, identifying in the text data a piece of information according to a pattern previously set; generating automatically an annotation containing the piece of information identified; assigning automatically a level of relevance for the annotation; asking a user for confirming the automatically assigned level of relevance; if the user does not confirm the automatically assigned level of relevance, assigning a second level of relevance instead of the automatically assigned according to a user input; and storing the annotation associated to the level of relevance assigned. A user can access later to these annotations to recall a phone call conversation or a voice message.

20 Claims, 6 Drawing Sheets



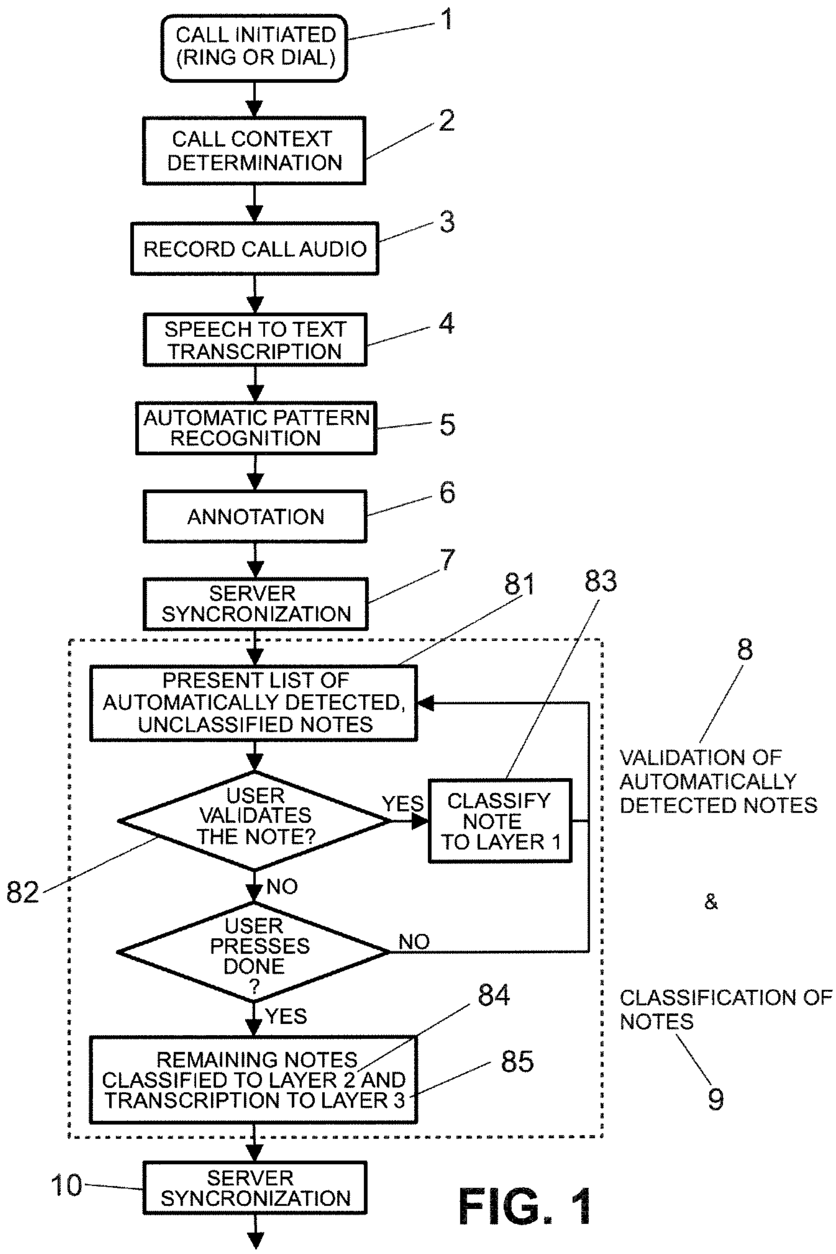

CALL INITIATED (RING OR DIAL)
1
CALL CONTEXT DETERMINATION
2
RECORD CALL AUDIO
3
SPEECH TO TEXT TRANSCRIPTION
4
AUTOMATIC PATTERN RECOGNITION
5
ANNOTATION
6
SERVER SYNCRONIZATION
7
81
83
8
PRESENT LIST OF AUTOMATICALLY DETECTED, UNCLASSIFIED NOTES
USER VALIDATES THE NOTE?
YES
CLASSIFY NOTE TO LAYER 1
VALIDATION OF AUTOMATICALLY DETECTED NOTES
&
82
NO
USER PRESSES DONE ?
NO
YES
84
CLASSIFICATION OF NOTES
REMAINING NOTES CLASSIFIED TO LAYER 2 AND TRANSCRIPTION TO LAYER 3
85
9
10
SERVER SYNCRONIZATION

FIG. 1

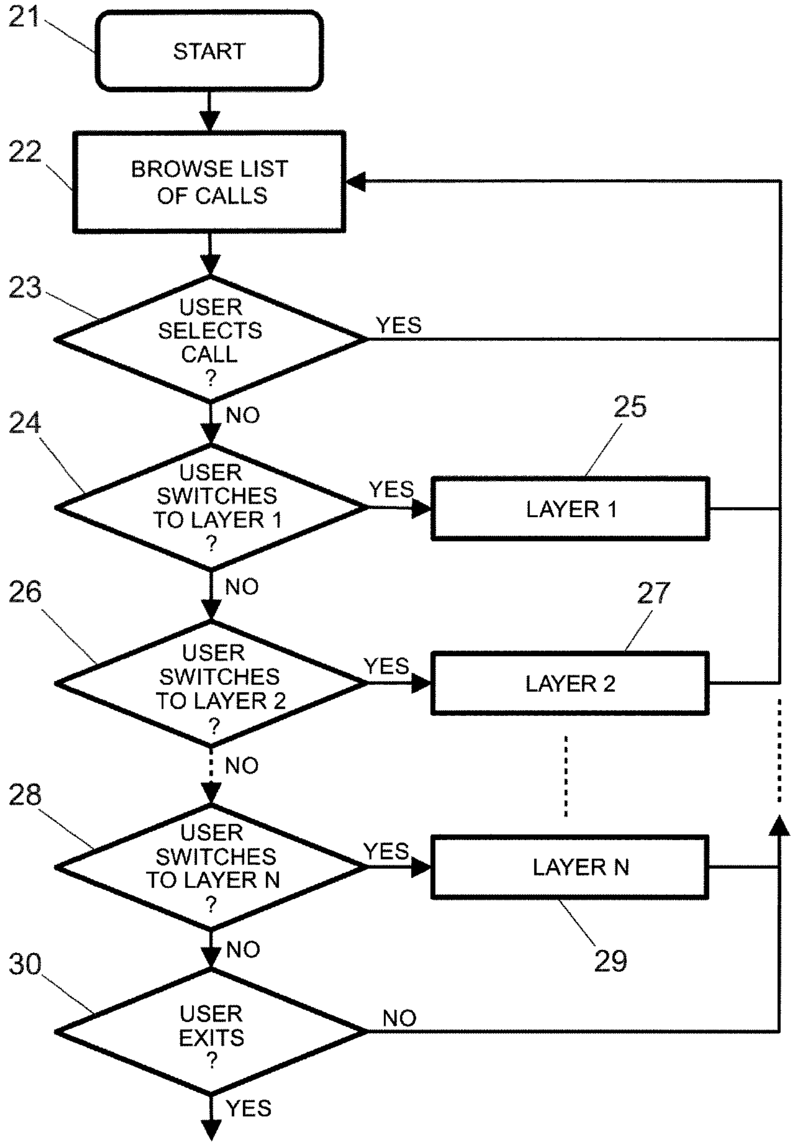

21
START
22
BROWSE LIST OF CALLS
23
USER SELECTS CALL ?
YES
NO
24
USER SWITCHES TO LAYER 1 ?
YES
25
LAYER 1
NO
26
USER SWITCHES TO LAYER 2 ?
YES
27
LAYER 2
NO
28
USER SWITCHES TO LAYER N ?
YES
LAYER N
29
NO
30
USER EXITS ?
NO
YES

FIG. 2

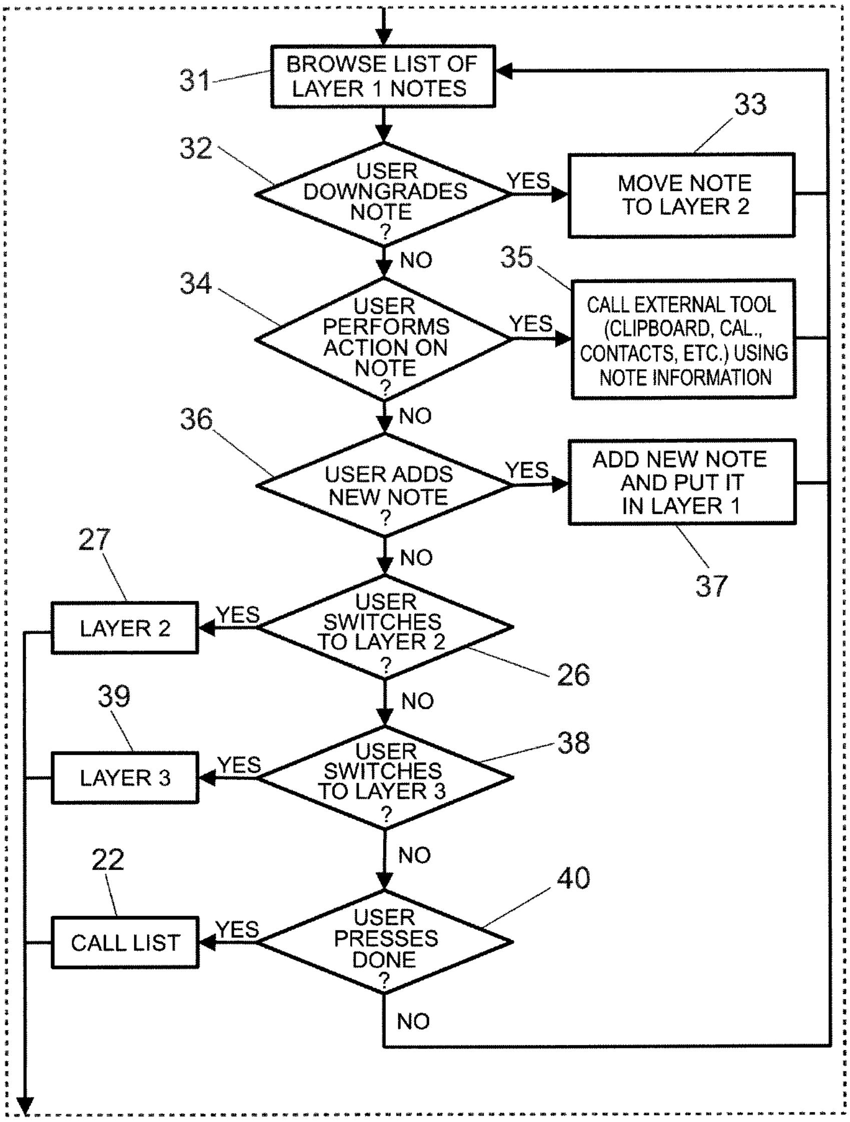

31
BROWSE LIST OF LAYER 1 NOTES
32
USER DOWNGRADES NOTE ?
YES
33
MOVE NOTE TO LAYER 2
NO
34
USER PERFORMS ACTION ON NOTE ?
YES
35
CALL EXTERNAL TOOL (CLIPBOARD, CAL., CONTACTS, ETC.) USING NOTE INFORMATION
NO
36
USER ADDS NEW NOTE ?
YES
ADD NEW NOTE AND PUT IT IN LAYER 1
37
NO
27
LAYER 2
YES
USER SWITCHES TO LAYER 2 ?
26
NO
39
LAYER 3
YES
USER SWITCHES TO LAYER 3 ?
38
NO
22
CALL LIST
YES
USER PRESSES DONE ?
40
NO

FIG. 3

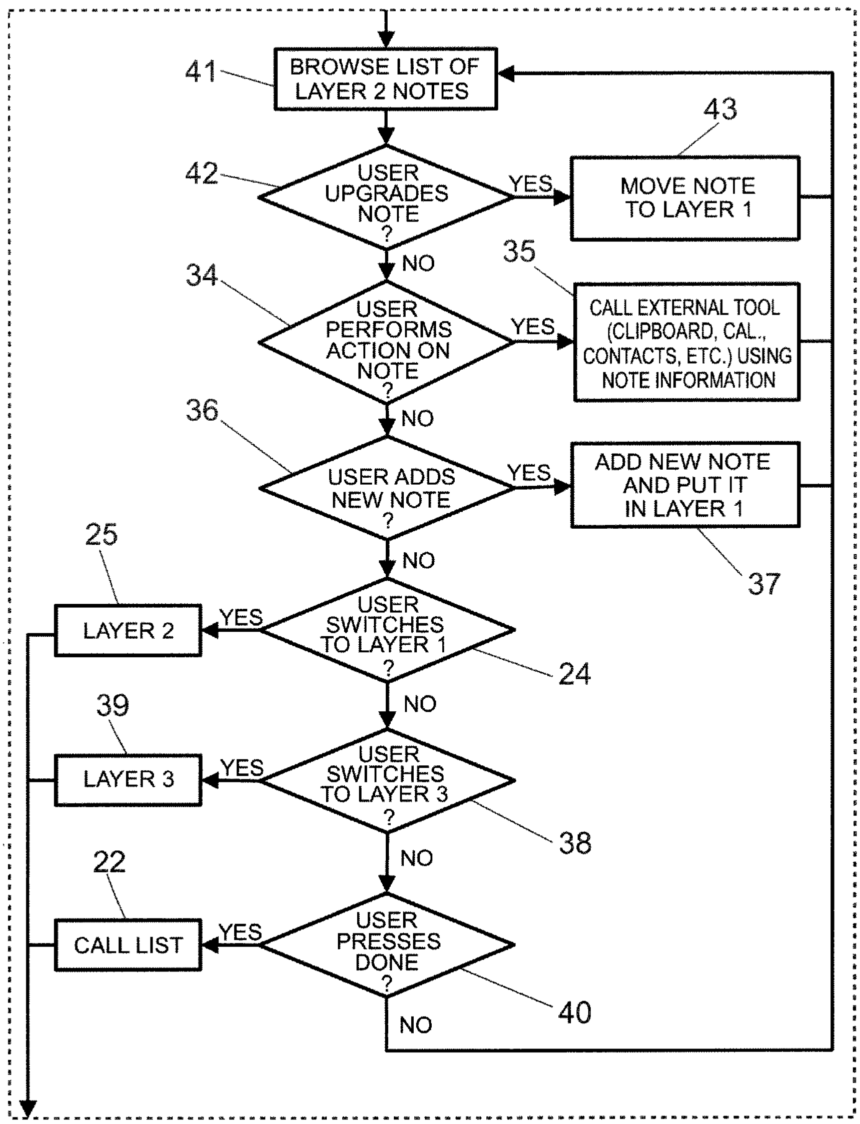
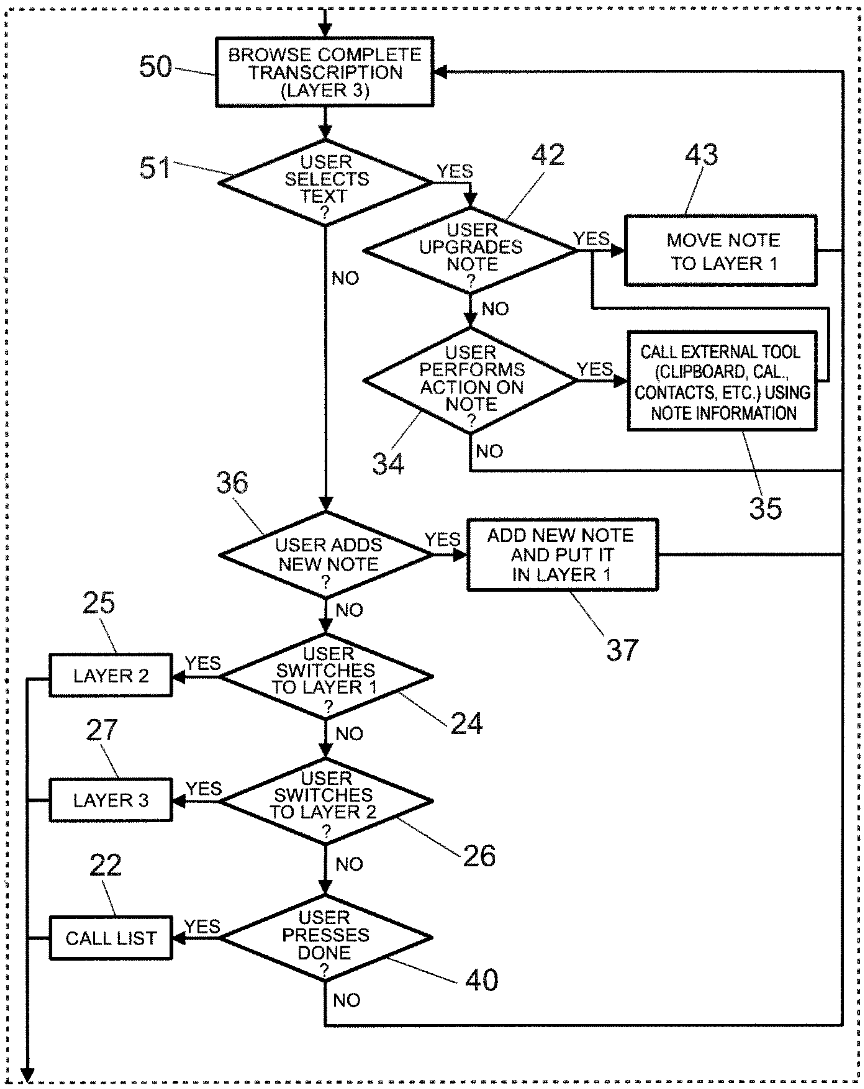
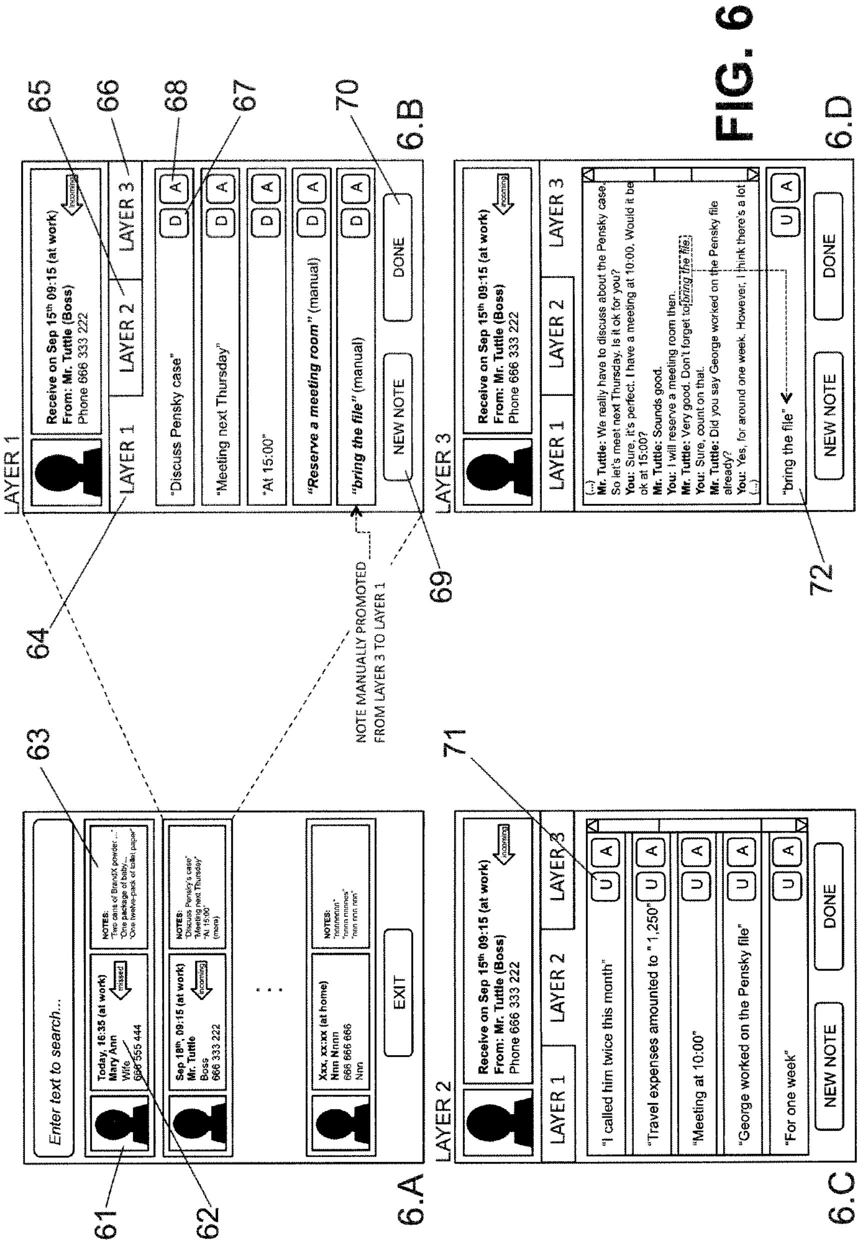

41
BROWSE LIST OF LAYER 2 NOTES
42
USER UPGRADES NOTE ?
YES
43
MOVE NOTE TO LAYER 1
NO
34
USER PERFORMS ACTION ON NOTE ?
YES
35
CALL EXTERNAL TOOL (CLIPBOARD, CAL., CONTACTS, ETC.) USING NOTE INFORMATION
NO
36
USER ADDS NEW NOTE ?
YES
ADD NEW NOTE AND PUT IT IN LAYER 1
37
NO
25
LAYER 2
YES
USER SWITCHES TO LAYER 1 ?
24
NO
39
LAYER 3
YES
USER SWITCHES TO LAYER 3 ?
38
NO
22
CALL LIST
YES
USER PRESSES DONE ?
40
NO

FIG. 4



50
BROWSE COMPLETE TRANSCRIPTION (LAYER 3)
51
USER SELECTS TEXT ?
YES
NO
42
USER UPGRADES NOTE ?
YES
NO
43
MOVE NOTE TO LAYER 1
USER PERFORMS ACTION ON NOTE ?
YES
NO
34
CALL EXTERNAL TOOL (CLIPBOARD, CAL., CONTACTS, ETC.) USING NOTE INFORMATION
35
36
USER ADDS NEW NOTE ?
YES
NO
ADD NEW NOTE AND PUT IT IN LAYER 1
37
25
LAYER 2
YES
USER SWITCHES TO LAYER 1 ?
NO
24
27
LAYER 3
YES
USER SWITCHES TO LAYER 2 ?
NO
26
22
CALL LIST
YES
USER PRESSES DONE ?
NO
40

FIG. 5



6.A
61
62
63
Enter text to search...
Today, 16:35 (at work)
Mary Ann
Wife
666 555 444
missed
Sep 18th, 09:15 (at work)
Mr. Tuttle
Boss
666 333 222
incoming
Xxx, xx:xx (at home)
Nnn Nnnn
666 666 666
Nnn
EXIT
NOTE MANUALLY PROMOTED
FROM LAYER 3 TO LAYER 1
LAYER 1
6.B
64
65
66
67
68
69
70
Receive on Sep 15th 09:15 (at work)
From: Mr. Tuttle (Boss)
Phone 666 333 222
LAYER 1
LAYER 2
LAYER 3
"Discuss Pensky case"
"Meeting next Thursday"
"At 15:00"
"Reserve a meeting room" (manual)
"bring the file" (manual)
D
A
NEW NOTE
DONE
LAYER 2
6.C
71
"I called him twice this month"
"Travel expenses amounted to "1,250""
"Meeting at 10:00"
"George worked on the Pensky file"
"For one week"
U
A
LAYER 3
6.D
72
(...)
Mr. Tuttle: We really have to discuss about the Pensky case. So let's meet next Thursday. Is it ok for you?
You: Sure, it's perfect. I have a meeting at 10:00. Would it be ok at 15:00?
Mr. Tuttle: Sounds good.
You: I will reserve a meeting room then.
Mr. Tuttle: Very good. Don't forget to bring the file.
You: Sure, count on that.
Mr. Tuttle: Did you say George worked on the Pensky file already?
You: Yes, for around one week. However, I think there's a lot
(...)
"bring the file"

FIG. 6

METHOD FOR COLLECTING AND STORING ANNOTATIONS ASSOCIATED TO A VOICE AUDIO DATA

TECHNICAL FIELD OF THE INVENTION

Present invention relates to the field of communications and specifically relates to meaningful annotations of voice communications as a phone call conversation or a voice message.

BACKGROUND OF THE INVENTION

Phone calls generate large amounts of information. The discussion of work-related topics, schedule of time and place of a meeting, annotation of shopping/to-do lists and contact details of a colleague, among others, often involve meaningful pieces of information which are hard to keep in mind after the conversation. Recalling information exchanged in phone conversations is not an easy task and it usually requires some method for memory aid. Traditionally, people would take hand notes in a paper while they were having a phone conversation, but as people use mobile telephones more and more often, it started to appear automatic methods for recording and summarizing phone calls. Personal notes primarily serve as a memory aid for individuals to remember important facts, actions, ideas, and decisions and the annotations may be as simple as highlighting passages or producing a new written product. It is often the case one has to take notes either during—online—or after—offline—cooperative situations, such as in meetings, lectures or phone calls.

Full recordings of phone calls were once considered to address this problem. Nevertheless, it has been demonstrated that it is not an efficient approach since browsing audio data is a lengthy activity that most people prefer to avoid.

Another later solution, avoiding recording full conversations, focuses on the fact that people often choose annotation tools for faster retrieval, even if that implies lower accuracy in the information recall process. Examples of these tools include audio recorders, mobile phone annotation applications, and the usual pen and paper approach. However, it is often the case when these resources are not available during a phone call, or even that either parties have both hands busy, thus preventing the note-taking activity. Furthermore, taking notes during phone calls implies an additional cognitive load that might lead to loss of conversation threads and break the natural information flow.

There are some services that try to address these issues by allowing users to perform certain actions during the call in order to create live annotations or set markers in the recorded calls. However, these solutions have some inconveniences; such as interrupting the natural information flow of a phone conversation, requiring users to perform specific actions during the call, or simply creating a blob of audio information with more data than it is actually needed.

In the recent years, methods for annotation in work-related meetings have been studied. Notes can be used as memory cues for participants to recall events of a meeting rather than being full recordings of the activity but both attention and active participation is required, and taking notes at the same time may become an additional cognitive load that reduces the person's ability to participate.

These needs have been lately supported by different solutions, including electronic annotation tools that leverage desktop computers or mobile devices, as well as the common paper and pen approach, still frequently used in the form of post-it notes, miscellaneous text files, or the corner of other printed documents.

A phone call could be considered as two-people meeting. Nevertheless, these activities are different from work-related meetings. For example, typical phone calls tend to be relatively shorter, they are frequently not planned beforehand and they lack the structure of a meeting, being plenty instead of a series of salutations and informal dialogs. Also, it has been observed that during phone calls, participants often have their hands busy, either by performing another activity (e.g. driving) or by holding the phone, documents or other objects. Despite these differences, specific annotations for phone calls have received little coverage in the prior art.

Finally, some related research efforts had been done in the process of automatic summarization of texts. Summarization aims to include all the concepts included in a body of text, while reducing the actual amount of text. Summarization then is not selective about the pieces of information that should be included in the final result: it must include every piece, but they must occupy less space. Annotation, on the other hand, aims to select very specific pieces from a body of information while ignoring the remaining ones. As example of an existing solution, the U.S. Pat. No. 7,602,892 "Telephony annotation services" provides a simple method for phone call annotation. This method requires users to remember a set of actions that must be taken to trigger the annotation process. It also requires users to interrupt the normal flow of the conversation to perform these actions, so, in consequence, users might lose pieces of information while performing these actions during a call (e.g. in order to setup audio markers or record a live audio note during a phone call). In addition, the nature of the method could significantly reduce precision in annotations (e.g. audio markers created in real-time by users should frequently have an offset time related to the actual important part of the call). And finally, because of its lack of precision, the resulting recorded information is excessive and would require further editing to obtain the actual annotations.

It is therefore, a lack in the prior art of a method or system to automatically annotates pieces related to phone conversations requiring very few—or none—user interaction at the moment of the call. A lack of a method able to automatically identify and annotate important pieces of information—in the long term—that tend not to be considered relevant in first instance whenever they are not part of the original objective of the call. Somehow, it is needed to organize automatically taken annotations avoiding requiring an immediate user interaction, but later the user should has the chance to decide about said organization by giving his approval or reassigning the relevance of the annotations automatically taken.

SUMMARY OF THE INVENTION

Present invention solves the aforementioned problem of annotating pieces of information related to an audio data, generally taken from a phone conversation in order to eventually recall the information by presenting a method for collecting, organizing and storing annotations associated to a voice audio data. Said annotations are automatically taken and organized according to certain relevance, but at some time the user decides to accept the organization or changing the relevance. The method is characterized by the following steps:

a) providing the voice audio data;

b) transcribing said voice audio data into text data;

c) identifying in the text data a piece of information according to a pattern previously set;

d) generating automatically an annotation containing the piece of information identified;

e) assigning automatically a level of relevance for the annotation;

f) asking a user for confirming the automatically assigned level of relevance;

g) if the user does not confirm the automatically assigned level of relevance, assigning a second level of relevance instead of the automatically assigned according to a user input;

h) storing the annotation associated to the level of relevance assigned.

A user accessing to the annotations stored may execute the steps of selecting one of the stored annotations and then changing the level of relevance, which had been assigned previously. Annotations require a validation or confirmation by the user before being stored. The method comprises the step of asking for a user input in order to confirm annotations automatically generated. If the user does not confirm the automatically assigned level of relevance, a second level of relevance is assigned instead of the automatically assigned according to a user input. Confirming an annotation may be accept or reject said annotation or also may be assign a certain level of relevance. So that, the user is prompted to interact with the annotations automatically generated.

The invention comprises grouping the annotation according to the level of relevance assigned. The level of relevance is assigned automatically to the annotation generated automatically, but the user can change the level of relevance and it is also possible to generate an annotation manually, which either is assigned automatically a level of relevance or the user assigns a level of relevance manually. These groups classified by their level of relevance can be displayed according to a layer scheme, referring each layer to a different level of relevance.

The voice audio data may be provided by a telecommunication network and specifically, the voice audio data provided may be derived from a phone call conversation or even a voice message. The annotations are stored, in one embodiment of the invention, associated to said phone call conversation or voice message. Then, all the phone call conversations or voice messages are displayed associated to the annotations and are available for the users.

In one embodiment of the invention, there are three levels of relevance a first level of relevance corresponding to annotations manually generated by a user and automatically generated annotations which level of relevance is reassigned to first level by a user, and automatically generated annotations which level of relevance is confirmed as first level by a user; a second level of relevance corresponding to automatically generated annotations which level of relevance is reassigned to second level by a user, and automatically generated annotations which level of relevance is confirmed as second level by a user; and a third level of relevance corresponding to the whole transcription of the voice audio data. Reassigning a level of relevance is to assign a different level of relevance different from the automatically assigned.

A particular embodiment of the invention comprises providing data related to a context of the voice audio data. The context may comprise a selection of at least one of temporal data, geographical data, GPS data, accelerometer data or name of the caller.

In another particular embodiment of the invention, the text pattern for identifying pieces of information comprises text data being numbers, question adverbs, proper names or geographical places. A text pattern-search algorithm is in charge of recognizing meaningful pieces of information according to the previously mentioned types or others.

Another aspect of the invention refers to a computer program product comprising computer program code adapted to perform the method of the invention when said program code is executed on a computer, a digital signal processor, a field-programmable gate array, an application-specific integrated circuit, a micro-processor, a micro-controller, or any other form of programmable hardware.

Finally, a digital data storage medium is provided for storing a computer program product comprising instructions causing a computer executing the program, to perform all steps of the method of the invention.

Proposed invention takes advantage of offering a user-interaction-free annotation of phone calls. It also provides benefits for a long-term preservation of important call information and a fast access to most important information while allowing browse and search on additional is notes information which might become more important in the longer term. All this means a more efficient use of note data than existing solutions.

DESCRIPTION OF THE DRAWINGS

To complete the description that is being made and with the object of assisting in a better understanding of the characteristics of the invention, in accordance with a preferred example of practical embodiment thereof, accompanying said description as an integral part thereof, is a set of drawings wherein, by way of illustration and not restrictively, the following has been represented:

FIG. **1**.—shows a flowchart comprising the steps of proposed invention.

FIG. **2**.—shows a flowchart comprising the steps of a user accessing to the information of proposed invention.

FIG. **3**.—shows a flowchart corresponding to one layer of one embodiment of the invention

FIG. **4**.—shows a flowchart corresponding to one layer of one embodiment of the invention

FIG. **5**.—shows a flowchart corresponding to one layer of one embodiment of the invention

FIG. **6**.—shows an example of an interface for browsing notes in an embodiment of the invention

DETAILED DESCRIPTION OF THE INVENTION

The invention describes a process for annotating information that has been discussed during a phone call. Once the annotations have been collected from an audio data, they are organized and stored in an efficient way that makes them available for the users to eventually recall the information.

Embodiments of present invention propose, given that the user is registered in a service with his telephone number, whenever the user dials or receives a phone call, the audio data of that call are recorded and transcribed into text. After this has been done, a pattern-search algorithm looks for certain pieces of information that indicate that useful information has been discussed during the call. Then, a natural language processing algorithm uses these patterns, some contextual information and the data in the transcription to automatically generate meaningful notes that might be useful for eventual recall.

When the user accesses call information (obtained as above described) for the first time, he will be presented with the list of automatically detected notes. The user can then organize the notes based on their relevance. In some embodiments, the notes are represented in different layers according to their relevance, and the layers are used to facilitate the access to note information by the user, who can browse and search information along the layers. An additional layer contains the whole call transcription.

Embodiments of the invention may contain a different number of layers. In the specific case of a three-layer implementation, the user selects which of the notes he approves, which are the notes that he considers being more important than the rest. Some embodiments of the invention also allow the user to manually add notes, which, because of their explicit importance, should be put in the first layer in order of relevance (layer 1). After this user intervention, call information is classified in three different but related categories:

Layer 1 notes: notes explicitly chosen, both by "approving" the automatically detected notes or manually taken by the user

Layer 2 notes: stores notes which were automatically detected but not approved

Layer 3 notes: stores the whole transcription of the phone conversation

FIG. **1** shows the steps of one embodiment of the invention. First of all, the user registers his phone number with a telephony service operator that provides the service implementing the method of the invention. The call can be initiated (**1**) in two scenarios:

Scenario 1 (server only): Phone calls are dialed or received by means of a mobile device (such as a mobile phone), a fixed phone, or desktop computer using a to Voice over IP application. Audio and text recording and processing is done in a server—in the cloud.

Scenario 2 (server+device): Phone calls are dialed or received by means of a mobile device (such as a mobile phone), or a desktop computer using a Voice over IP application. Audio and text recording and processing can be done either in the device (using a mobile or desktop application), or in the server. Answering an incoming call is not a prerequisite. If this is the case, the proposed method can also be applied to voice mail messages.

Next step is the determination of a call context (**2**). When a call is dialed or the phone device rings, common sensors embedded in the phone device (when available) and call-related data (e.g. party information and time of the call) are used to determine the context of the call (e.g. if it is a work call, location, companions at the time of the call, etc.). Then, the phone call is recorded and stored (**3**). According to the scenarios mentioned previously, audio data is recorded either in a cloud server or in the device. In fact, all the steps of the method of the invention may run on a cloud or in a device, as mobile telephone, depending on the specific embodiment. Both channels (caller and callee) are recorded separately. This activity ends as soon as the call hangs up or is interrupted for whatever reason.

The speech to text transcription (**4**), call audio data is analyzed and transcribed to text in the server or in the mobile or desktop application, depending on the scenario. After that, an Automatic Pattern Recognition algorithm (**5**) running either in the server, mobile device or desktop computer, look for common patterns that have been found to correlate to note taking. Patterns may be, but are not limited to, numbers (e.g. a phone number such as 6559900, an ordinal like "third"), follow-ups to question adverbs ("where", "when", "how", "who", "whom", "what", "which") and proper names (e.g. "John", "Catalunya", "Colombia").

The obtained patterns, as well as contextual data obtain in step 2, are used as a basis to perform semantic analysis and natural language processing in order to obtain meaningful notes (**6**). For instance, a phone number "6559900" might help obtaining the more meaningful note "John's phone number is 6559900"; the name of a place called "Catalunya" might lead to "Let's meet at Plaza Catalunya". A list with these annotations is built.

If the user employs a desktop or mobile application implementing the proposed method as happens in the scenario 2, call annotation information is synchronized (**7**) between such application and the server, and associated with the corresponding call log entry. This keeps the data up to date both in the server and the application.

Up here, the steps are triggered by the call (**1**), but the following steps (validating and classifying notes) can be either also triggered by the call or initiated by the user at any time.

The validation of automatically detected notes (**8**) depends on the two scenarios presented before. In the Scenario 1, after the call is finished, the user can enter a web application where call data is presented as well as a note validation prompt for the call. In the Scenario 2, the user is however presented only with a note validation prompt right after the call is finished. In both scenarios, the note validation prompt displays the list of the automatically generated notes (**81**).

The note validation prompt asks (**82**) the users for ranking the notes according to their relevance. The number of levels of relevance can be configured. A specific case with three levels of relevance just asks the user for accept or reject a note. Accepted notes are associated to a higher level (**83**) of relevance that rejected notes (**84**). Being the lowest level of relevance reserved for the whole transcription (**85**) of the audio data. The validation can be performed at any time after previous steps have finished. A user can validate notes immediately after the call or whenever s/he prefers to do it.

Additionally, the user is able to manually enter new notes at this point, so to add custom information that he might consider important while the call is fresh in his mind.

The classification of notes (**9**) depends an the level of relevance assigned by the user in the previous step; notes are classified in some embodiments in a multi layer annotation scheme. This way, notes considered to be more important are classified in Layer 1, notes ranked to be less important, or assigned with a lower level of relevance, are classified in Layer 2, notes even less important are in Layer 3 and so on. A last layer is added, containing the whole call transcription. This allows the user to browse the expanded call information if desired. Manually entered notes, being implicitly important, are classified as Layer 1 notes.

Finally, if the user employs a desktop or mobile application implementing the proposed method as described in the scenario 2, call and annotation information is uploaded to the server in order to keep an online backup of the data. This server synchronization (**10**) allows users to preserve call annotations even if the user decides to access the information from another device.

FIG. **2** represents the steps followed to manage and consume the data acquired. The process can be started (**21**), according to the scenarios previously mentioned, by:

i. opening the related mobile/desktop application, or

ii. accessing a web application by means of a web browser

Then, a list of calls is displayed to the user, and for each call, its related information (such as caller/callee, time, location or any other available contextual information) and a clickable summary of notes. When this summary is clicked for a specific call, the corresponding notes appear.

The search/browse note information step (**22**) comprises a user browsing the stored notes, being the information presented in a Multilayer Annotation Visualization Interface. It consists in a number of layers that are used to store the call notes. Lower-numbered layers contain more important notes and vice versa (as explained before). The user stays in step **22** unless he selects a call. Then, when a call is selected, the user is asked (**24**) about accessing to the layer containing the notes more relevant, if he accepts, said layer is displayed (**25**) to him. If not, next layer in order of relevance is asked (**26**) to be access, if he accepts said layer is displayed (**27**) to him. This process is repeated (**28**) until the user selects a layer "n" to be displayed (**29**) or until the user rejects accessing the last layer, then the user exits from the method. According to a specific three-layer implementation, the FIGS. **3**, **4** and **5** describe the process followed in each layer.

Users can reclassify the notes if they consider their importance have changed at some point in time. If a user considers a note's importance to have decreased, it can be downgraded, so it will be moved to a lower importance layer. If he considers a note's importance have increased, he can upgrade it, moving it to a higher importance layer. Thus, the information is to managed based on users' needs.

FIG. **3** describes the steps of layer 1, the layer containing the most relevant notes. Once the user has selected to enter in layer 1, the notes comprised by layer 1 are displayed to the user. First step is checking (**32**) if one of the notes has to be downgraded to a less relevant layer by instructions of the user. If the user has selected a note to be downgraded, then said note is moved to layer 2 (**33**). In the case that the user is not downgrading any note, next step is checking if the user wants to perform some action (**34**) on a note. When the user performs an action on a note, the method skip to perform said action as for example adding a contact to a contact list, a date to a calendar or email information (**35**). If the user does not perform any action on any note, then the next step is asking the user for adding (**36**) notes, if the user adds (**37**) a note, said note is automatically put in layer 1. If the user have not added any note, the method finishes checking the actions taken on the notes and depending on the selection of the user (**26**) (**38**) goes to other layer (layer 2 (**27**) or layer 3 (**39**)) or come back to the call list (**22**) in the case that the user considers that he has finished browsing (**40**).

FIG. **4** describes the steps of layer 2. Once the user has selected to switch to layer 2 (**27**), then a list of layer 2 notes is displayed (**41**) to the user. The first step comprises checking if a note has to be upgraded (**42**). If the user has selected a note to be upgraded to layer 1, then said note is moved (**43**) to layer 1. If the user has not selected any note to be upgraded, the next steps work as in FIG. **3**, plus the step of asking the user for switching to layer 1 (**24**), in that case layer 1 is displayed to the user (**25**).

FIG. **5** describes the steps of layer 3. Once the user has selected to switch to layer 3 (**39**), then a complete transcription of the audio data is displayed (**50**) to the user. In this layer the user can select a piece of text (**51**) and then there are two options: upgrading (**42**) a note containing the piece of text, which is automatically moved (**43**) to layer 1, or performing an action (**34**) on the note containing the piece of text. Both options lead to move the note to layer 1.

An example interface of such implementation with three layers, containing three different levels of relevance for the notes taken from the audio data, in a mobile device is shown in FIGS. **6**A, **6**B, **6**C and **6**D. FIG. **6**A shows a home screen, where the not of calls is displayed associated to a certain contact (**61**) of the user, context details (**62**) and notes (**63**) from layer 1.

FIG. **6**B represents the screen that is displayed when the user selects one call from the list displayed in FIG. **6**A. In this case the notes of layer 1 (**64**) are display, but the user can select to be displayed layer 2 (**65**) or layer 3 (**66**). All the notes are display associated to a couple of buttons to downgrade (**67**) a note to layer 2 or perform an action (**68**) on a note. At the bottom of the screen there are two more buttons for adding (**69**) new notes manually or to come back (**70**) to the list of calls of FIG. **6**A.

In FIG. **6**C, the user has selected "layer 2" (**65**). The notes assigned with a level of relevance according to layer 2 are displayed associated to a couple of buttons; one button for performing an action (**68**) on a note, as in FIG. **6**B, and one button for upgrading (**71**) a note and move it to layer 1.

In FIG. **6**D, the user has selected "layer 3" (**66**). The transcription of a whole conversation is displayed and the user has the option of selecting a piece of text (**72**) and then, executing the same actions than in layer 2, upgrading (**71**) or performing an action (**68**) on the selected text.

This layer-based sorting of notes might be changed according to the user's needs, such as ordering by call date and time, caller/callee name, etc.

Finally, three embodiments of the invention, according to the specific embodiment of three-levels multilayer scheme disclosed before, are described below to highlight the benefits of proposed invention in daily situations.

First Embodiment

This first particular embodiment comprises a young user A, who has a profile as an active smartphone user and communicates frequently with work colleagues and friends using his mobile phone. The steps of the invention, specifically for this embodiment:

- Activation: user A decides to subscribe to the service of the invention, called "Annotation Service" offered by his Communication Service Provider to help him remembering information that he frequently discusses during phone calls. Additionally, he installed the companion application to manage the annotation service and system from his mobile phone. He receives a call from a friend so the method is activated.
- Call context determination: the application estimates, by means of reading accelerometer and GPS data on user A's mobile, that he is at his office. According to the call log and information in his contact list, the caller is from his boss, user B.
- Phone call recording and storage: the mobile application records audio of both parties' channels (User A's and user B's). When the call is hung up, both audio files are stored in his mobile phone.
- Speech to text transcription: the audio file is transcribed into text in his mobile.
- Automatic pattern recognition: the Automatic Pattern Recognition algorithm is run on the transcription.
- Annotation: automatically detected patterns are leveraged to create meaningful notes. A list is compiled with these items:
  - i. "I called him twice this month"
  - ii. "Travel expenses amounted to € 1,250"
  - "Discuss Pensky case"
  - iv. "Meeting next Thursday"
  - v. "I have a meeting at 10:00"
  - vi. "At 15:00"
  - vii. "George worked on the Pensky file"
  - viii. "For one week"
- Server Synchronization: Automatically detected notes are stored and associated with the call log entry in his mobile and all information (call log and notes) is uploaded to the server.

- Validation of automatically detected notes: right after previous steps are completed, user A's mobile phone displays a Note Approval Prompt. He approves the notes "Discuss Pensky case" and "Meeting next Thursday".
- Classification of notes: the system classifies the notes mentioned in previous steps as Layer 1 notes,
- Server Synchronization: just after user A closes the application in his mobile, call data (transcription and annotations) are sent to the server. Now user A could use either the application on his mobile or a desktop application to access his calls' notes.

Now, once the information has been collected, organized and stored, to manage and consume the data acquired, the following steps are comprised in the process:

- The Note Approval Prompt goes away and user A keeps working on the Annotation System without closing it.
- Search/browse note information: user A reviews the three Layers to be sure all the information he needs is in the right place.
- Note upgrading and downgrading: while browsing Layer 2 he realizes he did not take note of the time of the meeting, so he upgrades the note "At 15:00". The system sends this note to Layer 1, while keeping the remaining notes in Layer 2.
- Manual annotation: while in Layer 1, user A adds the note "Reserve a meeting room", to complement the notes in Layer 1. He also knows that there is something else he needs to take into account from the previous conversation. He goes to Layer 3 and browses the transcription. Then select the text "bring the file" and upgrades it. This text is sent as a note to Layer 1.
- Taking action on notes: user A selects the option to send the note "Discuss Pensky case" to his calendar application. He repeats this step for the remaining notes to complement the meeting data in the calendar application.

Second Embodiment

The second embodiment comprises a user C having a senior profile as a processed meat distributor, who uses the fixed phone in his office to receive orders from local restaurants and small supermarkets. He also uses a desktop computer to manage orders, accounting, and his clients contact information.

- Activation: user C frequently receives calls asking for a variety of products, so he decided to hire the Annotation Service to his Communication Service Provider to better remember the high amount of orders he receives every week. Given that user C hired the service, every call he makes or receives is monitored by the annotation server. User C receives a call from a local restaurant owner, user D.
- Call context determination: the server detects the number of the call and determines it comes from user C's client user D. It also checks the time and determines it is done during working hours.
- Phone call recording and storage: the call is monitored and recorded in the server. After hanging up, the call is stored in the cloud.
- Speech to text transcription: the server analyzes and transcribed the speech data to text data.
- Automatic pattern recognition: the server runs the Automatic Pattern Recognition algorithm, and found the notes "fifty kilos", "Serrano", "twenty kilos", "which", "Riojano", and "Friday".
- Annotation: the semantic and natural language analysis extracts the following notes from based on the found patterns, the caller, and on the fact that the call was done during work hours:
  - i. "Fifty kilos of jamón Serrano"
  - ii. "Twenty kilos of chorizo. Kind: Riojano"
  - iii. "Due to next Friday"
  - iv. "For user D"
- Server Synchronization: given that in this scenario the user does not use a desktop or mobile application, it is not necessary to perform this step.
- Validation of automatically detected notes: by the end of the day, user C accesses the web application to check the calls he received that day. He browses the list of calls and opens the one received from his client user D. He clicks on the approval button of all notes, and manually enters a new note "Chorizos should be smoked". This note is automatically put in Layer 1.
- Classification of notes: the backend of the web application, based on user C's choices, classifies all the automatically detected notes, as well as the manually entered note as Layer 1 notes and the whole transcription as a Layer 3 note.
- Server Synchronization: again, there is no need for synchronization between an application and the server. However, at this point all call information is stored in the server, so it can be eventually accessed by the web application or other dedicated mobile or desktop applications if needed.

Now, once the information has been collected, organized and stored, to manage and consume the data acquired, the following steps are comprised in the process:

- Later, after shipping the ham and chorizos to user D, user C enters the web application again.
- Search/browse note information: he browses the call list in the application and find user D's call.
- Note upgrading and downgrading: user C downgrades all notes, so preventing them to appear in Layer 1, and sending them to Layer 2.
- Manual annotation: user C does not take any additional notes.
- Taking action on notes: user C selects the note containing the text "Fifty kilos of jamón Serrano", now in Layer 2, and selects the action "copy to clipboard". He pastes the content of the note in an inventory application. He repeats the process for the note, which contains "Twenty kilos of chorizo. Kind: Riojano".

Third Embodiment

This third embodiment show how the method can be used in situations when the callee is not available, thus requiring processing of the voice mail message left by the caller. This embodiment comprises a user E. The profile of user E is a young executive frequently discusses over the phone with a user F and work colleagues. He installed a companion application on her mobile. Due to the multiple meetings at work, he frequently cannot answer his phone, so he relies on voice mail.

- Activation: user E is in a meeting, thus he turns the ringer on her phone off. While at the meeting, user F calls him to ask him to buy some groceries on his way home. Since user E does not answer, user F leaves a message in his voice mail. This triggers the Annotation Service implementing the proposed method.

Call context determination: the call is detected to come from user E's home phone, and since it was work time and user E had her ringer off, the Annotation Service determines he was at work.

Phone call recording and storage: after user E's husband finishes recording the voice mail message and hangs the phone, the audio of the call (only one channel in this case) is stored.

Speech to text transcription: call audio data is transcribed into text.

Automatic pattern recognition: after the server has run the Automatic Pattern Recognition algorithm, it detects the words "two", "BrandX", "one", and "twelve".

Annotation: after the semantic analysis, these notes are generated by the Annotation Service:

i. "Two cans of BrandX powder milk"

ii. "One package of baby wipes"

iii. "One twelve-pack of toilet paper"

iv. "Michelle says hello"

Server Synchronization: the server pushes automatic annotation data to user E's mobile. It is associated to the corresponding call log entry.

Validation of automatically detected notes: after the meeting, user E sees that he has a lost call, which left a new voice mail message. Given that he has hired the Annotation Service, the application in his mobile also displays the call annotations so they are ready for revision in his mobile. He approves first three notes.

Classification of notes: first three notes are classified as Layer 1 notes, fourth note is classified as Layer 2 and the whole transcription receives Layer 3 classification.

Server Synchronization: new information, including note classification is uploaded and updated in the server.

Now, once the information has been collected, organized and stored, to manage and consume the data acquired, the following steps are comprised in the process:

User E opens the Annotation System application to remember what he has to buy before going home.

Search/browse note information: he browses the calls and goes to user F's call and opens its associated notes.

Note upgrading and downgrading: he does not upgrade nor downgrade any note at the moment.

Manual annotation: he adds a manual note: "milk, bread and eggs", since he knew he need to buy those as well. This note is automatically put in Layer 1.

Taking action on notes: since the list of groceries he has to buy is comfortably stored in the Layer 1 of the Annotation System application, he does not have to take any additional actions.

The invention claimed is:

1. A computer-implemented method for collecting, organizing and storing annotations associated to a voice audio data characterized by the steps of:

a) providing the voice audio data;

b) transcribing said voice audio data into text data;

c) identifying at least one text pattern from the text data according to a pattern previously set;

d) generating automatically an annotation containing the text pattern identified;

e) assigning automatically a level of relevance for the annotation;

f) asking a user for confirming the automatically assigned level of relevance;

g) if the user does not confirm the automatically assigned level of relevance, assigning a second level of relevance instead of the automatically assigned according to a user input;

h) storing the annotation associated to the level of relevance assigned.

2. The method according to claim **1** further comprising a user generating an annotation manually.

3. The method according to claim **2**, wherein there are three levels of relevance: a first level of relevance corresponding to annotations manually generated by a user and automatically generated annotations which level of relevance is reassigned to first level by a user, and automatically generated annotations which level of relevance is confirmed as first level by a user; a second level of relevance corresponding to automatically generated annotations which level of relevance is reassigned to second level by a user, and automatically generated annotations which level of relevance is confirmed as second level by a user; and a third level of relevance corresponding to the whole transcription of the voice audio data.

4. The method according to claim **2**, wherein the voice audio data is derived from a phone call conversation or a voice message.

5. The method according to claim **4** further comprising grouping the annotations according to the level of relevance assigned.

6. The method according to claim **5**, wherein there are three levels of relevance: a first level of relevance corresponding to annotations manually generated by a user and automatically generated annotations which level of relevance is reassigned to first level by a user, and automatically generated annotations which level of relevance is confirmed as first level by a user; a second level of relevance corresponding to automatically generated annotations which level of relevance is reassigned to second level by a user, and automatically generated annotations which level of relevance is confirmed as second level by a user; and a third level of relevance corresponding to the whole transcription of the voice audio data.

7. The method according to claim **6** further comprising providing data related to a context of the voice audio data.

8. The method according to claim **7** further comprising running the steps of the method in a mobile telephone.

9. The method according to claim **1**, wherein the voice audio data is derived from a phone call conversation or a voice message.

10. The method according to claim **9**, wherein the phone call conversation or voice message is displayed associated to the annotation stored.

11. The method according to claim **1** further comprising grouping the annotations according to the level of relevance assigned.

12. The method according to claim **11** wherein the groups of annotations are displayed according to a layer scheme, referring each layer to a different level of relevance.

13. The method according to claim **1** further comprising providing data related to a context of the voice audio data.

14. The method according to claim **13** wherein the context comprises a selection of at least one of temporal data, geographical data, GPS data, accelerometer data or to name of the caller.

15. The method according to claim **1**, wherein the text patterns comprise text data being numbers, question adverbs, proper names or geographical places.

16. The method according to claim **1**, wherein the voice audio data is provided by a telecommunication network.

17. The method according to claim **1** further comprising uploading the voice audio data to a server.

18. The method according to claim **1** further comprising running the steps of the method in a mobile telephone.

19. A computer program product comprising computer program code adapted to perform the method according to claim **1** when said program code is executed on a computer, a digital signal processor, a field-programmable gate array, an application-specific integrated circuit, a micro-processor, a micro-controller, or any other form of programmable hardware.

20. A digital data storage medium storing a computer program product comprising instructions causing a computer executing the program, to perform all steps of a method according to claim.

* * * * *

ANNEX II. Research activities surrounding this thesis

Related research projects

Ethics in persuasive computing

While conducting his research at Telefónica I+D in Barcelona, the author of this thesis co-authored a paper proposing strategies for the implementation of ethics in behavior design. It was published as a work in progress in CHI 2014, and it is included next.

R. de Oliveira and J. P. Carrascal. "Towards Effective Ethical Behavior Design". *CHI '14 Extended Abstracts on Human Factors in Computing Systems.* CHI EA '14. Toronto, Ontario, Canada: ACM, 2014, pp. 2149–2154

Towards Effective Ethical Behavior Design

Rodrigo de Oliveira*
Telefonica Research
Pl. de Ernest Lluch i Martín 5
08019 - Barcelona
oliveirar@google.com

Juan Pablo Carrascal
Universitat Pompeu Fabra
Carrer de Tànger, 122-140
08018 - Barcelona
jp.carrascal@upf.edu

* The author is currently affiliated with Google Inc., USA.

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.
ACM 978-1-4503-2474-8/14/04.
http://dx.doi.org/10.1145/2559206.2581182

Abstract

Many of today's persuasive systems are designed taking into account cognitive biases to foster positive changes in people's behavior (e.g. adopt greener attitudes). However, these biases are also exploited to shape the users' behavior in a way that not necessarily benefit them (e.g. user retention in a website). Scholars addressed this problem by developing design guidelines and methods for ethics in persuasive computing, but these approaches alone have proved to be inefficient since they require every designer to be aware, understand, and comply with the recommended ethical practices. We propose preventive approaches that shall support higher compliance, as well as a remediation-based approach that does not require compliance from every designer. These approaches aim to help users understand persuasive elements embedded in systems, as well as to take more rational decisions when interacting with them. We expect that using preventive and remediation-based approaches will more effectively implement ethics in behavior design.

Author Keywords

Persuasive computing; Ethics; Cognitive biases

ACM Classification Keywords

H.5.m [Information Interfaces and Presentation (e.g. HCI)]: Miscellaneous; K.7.4 [The Computing Profession]: Professional Ethics

Motivation

One of the greatest faculties of the human brain is the ability to reduce the enormous amount of information surrounding us to a proper size for processing and yet keeping relevant information as proved by millions of years of adaptation. However, this extraordinary gift is subject to a number of cognitive biases that can lead people to take decisions not in their best interest. Some examples include people's frequent poor dietary choices and sedentary lifestyle [6], which can cause several illnesses, such as diabetes and heart disease.

Understanding and changing these behaviors have been the main research topic of a large body of work in persuasive computing [12]. Many scholars have leveraged persuasive techniques to shape people's behavior for preferable outcomes. For example, Arroyo et al. [7] developed Waterbot, a system that can be installed on household faucets to motivate people to turn off the tap when the water is not being used. Green attitudes have also been promoted for sustainable uses of energy resources [21]. In preventive health, Oliveira el al. [11] developed a mobile social game to help patients become more adherent to their medication prescription.
Lee et al. [16] evaluated several persuasion techniques to promote healthier eating habits in the workplace, including the default bias and planning strategies. Consolvo et al. [9] used the addictiveness of game playing to fight obesity, And related to privacy, Wang et al. [20] proposed man-in-the-middle-like technologies to mitigate biases when sharing information in online social networks.

These are only a few examples highlighting how proficient our scientific community has become in fostering positive changes in people's behavior. However, in many cases it is not clear what a positive behavior change is. Should users be persuaded to spend more time checking their online social network feeds? Should they be influenced to buy more items in an e-commerce website? Designers of these technologies might think so, but it is not clear that these target behaviors are positive for users. Ethical[1] issues arise when technologies are designed to shape the users' behavior towards a target not intentionally defined by them. As pointed out by Smids [17], the voluntariness condition is key and must consider external influences and whether the user acts intentionally.

In persuasive computing, scholars have proposed guidelines and methodologies to support designers in ethical behavior shaping. Berdichevsky and Neuenschwander [8] were among the first researchers to propose guidelines, suggesting in their golden rule of persuasion that "creators of a persuasive technology should never seek to persuade a person or persons of something they themselves would not consent to be persuaded to do". In a similar attempt, Fogg [12] suggested 7 steps for designers to evaluate the ethical nature of a persuasive technology by examining its intentions, methods and outcomes. Discourse ethics was proposed to search for further guidelines in the field [18]. Several methodologies for ethical design have been proposed, including value sensitive design [10], persuasion profiling [14], and deep involvement with stakeholders and users [15]. And in personal informatics, systems have been designed to encourage and support self-reflection and self-behavior management [4].

Although very enlightening, the aforementioned principled approaches depend on designers' awareness, understanding, and commitment to ethical practices. However, there is a number of examples revealing that designers have not embraced such practices [1]. We believe there is an urgent need to combine today's

[1] In this paper we refer to *ethical* design as the process by which designers create persuasive technologies following guidelines and methodologies suggested for ethical behavior shaping [8, 12].

Persuasion Summary
Task: Flight ticket purchase

Biases
- decision-making and behavioral .... !
- probability and belief .... !
- memory errors .... !
- social .... !

Inappropriate design patterns
- bait and switch .... !
- disguised ads .... !
- faraway bill .... !
- forced continuity .... -
- forced disclosure .... !
- hidden costs .... !
- trick questions .... !
- other inappropriate patterns .... !

Examples of impact for the user
- purchase more expensive flights
- purchase additional insurances
- spend more time searching flights

! Present
! Present, but can be disabled
- Not present or disabled by default

Full report available at: [web address]
Evaluated by [organization] on [date]

Figure 1: Mock up of a persuasion summary label describing persuasive elements used by a website that sells flight tickets. The label could be presented on the website, in online application stores, or attached to the physical software box sold in physical stores. More details on the list of inappropriate design patterns can be found in [1].

principled-based approaches with more effective solutions to implement ethics in persuasive computing.

Proposed Approaches

In self-beneficial behavior shaping, designers apply various persuasive techniques to shape the users' behavior towards targets defined *by the users* (e.g., eat healthier, quit smoking). In these cases, designers are not required to remove biases, but they rather introduce new ones, supposedly "stronger", to help users change some of their undesired behaviors.

However, this does not apply to cases where users do not know how they should best behave *and* would like to take a more rational decision about it (e.g. deciding whether to use a certain service more often or not). While some designers have addressed this user need supporting self-reflection for unbiased decision making [4], others have designed technologies that shape the user's behavior for their own benefit [1]. From an analysis of previous work in the field and today's motivations for building persuasive technologies, we highlight three main issues that should be addressed in order to effectively support ethical behavior design:

1. Lack of Awareness: Today's approaches are based on providing awareness of guidelines and methodologies for ethical behavior design, mostly in scientific journals and conferences (e.g. [8, 12, 10, 14, 18, 15, 4]). However, not every designer has access to these channels or is proactively engaged in the research community.

2. Lack of Understanding: As noted by Torning et al. [19], ethical considerations have been often mentioned, but not clearly addressed. We further highlight that not only it is unclear how designers should best apply related theoretical concepts in commercial products, but also that consumers are not knowledgeable about how persuasive technologies shape their behavior. Hence why some designers came up with initiatives to collectively create consumer-targeted educational content [1].

3. Lack of Commitment: By disregarding ethics in behavior design and leveraging persuasive techniques to increase user base and user retention, designers were able to quickly increase their profits [13]. Hence, designers have a conflict-of-interest that prevent them from committing to ethical practices.

Next we propose three approaches to address these issues.

Enforced Prevention
Preventive approaches that *enforce* the use of ethical guidelines [8, 12] could leverage the influence of regulatory entities, such as the government or organizations with decision power to authorize and deny commercialization of technologies in a certain territory. Specialized committees could be created to evaluate ethics of persuasive technologies. These evaluations should yield consumer-targeted summaries with information about the persuasive techniques used in each technology studied by the committee. These summaries could be designed based on consumer labeling efforts capable of presenting complex information in a concise and easy to understand format, such as the "Nutrition Facts" label [2]. Figure 1 shows an example of our proposed design.

The main advantages of this approach include: (1) making consumers *aware* of ethical issues in persuasive technologies, (2) providing them with the tools to *understand and judge* between competing technologies, and (3) ensuring the designers' *commitment* to ethical practices by means of regularization. However, it is limited by the complexity and time demand for proposing, discussing, voting, and implementing regulation laws that prevent bad cases of behavior design.

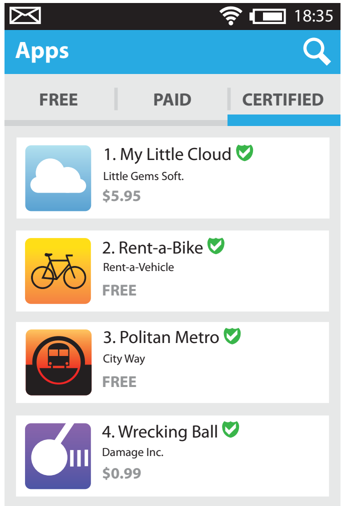


Figure 2: Example of active encouragement to ethical behavior design. The mobile search interface groups results by certified applications, i.e. apps evaluated by an expert committee and certified to meet minimum requirements for ethical behavior design. This special category gives higher visibility to designers of these apps, while also enabling consumers to enter a space where they feel safe when searching and downloading applications.

Actively Encouraged Prevention
The aforementioned approach can be envisioned in a scenario where designers are not obliged to adhere to ethical guidelines, but rather *actively encouraged* to do so. By active encouragement, we mean to *ease* and *motivate* the adoption of ethical design practices and the generation of consumer-targeted summaries.

In that sense, designers should have easy access to guidelines and methodologies for ethical behavior design, as well as to organizations that evaluate compliance to these ethical practices. In addition, consumers should be able to easily identify and check whether technologies meet requirements, e.g. with visual cues, like those from certified websites in the internet security domain [5].

In terms of motivation, designers' compliance to ethical practices could be rewarded by displaying certification stamps next to names of websites or applications—listed in search results—that were awarded the certification. Similarly, online application stores could include a specific category for searching "certified apps", thus supporting higher visibility of apps that comply with ethical design practices, as well as allowing consumers to enter a space where they feel comfortable searching and downloading applications (see Figure 2). Once these certifications become ubiquitous, we expect consumers to give preference for using certified persuasive technologies, and hence companies shall be more motivated to follow guidelines from persuasive computing. Further research is needed to identify appropriate incentives for the first adopters of these certifications, such as tax deductions, privileged governmental partnerships, among others.

Although actively encouraged prevention is not limited by the lengthy process of establishing regulation laws, it adds extra complexity for implementing different incentive schemes and does not guarantee commitment from every designer.

Remediation-based Approach
Preventive approaches are not perfect, and hence consumers might be always surrounded by technologies that do not comply with ethical design guidelines. We propose a remediation-based approach to empower users to reveal and remove—or at least mitigate—biases in persuasive technologies. We envision its implementation by groups of designers that implement adaptive solutions for technologies that misuse behavior design. Three activities shall be conducted by these groups:

1. Identify the most relevant biases in the given persuasive technology;
2. Provide interactive mechanisms that enable users to reveal these biases for self-awareness and educational purposes;
3. Provide intervention methods that empower users to remove or mitigate the effect of existing biases.

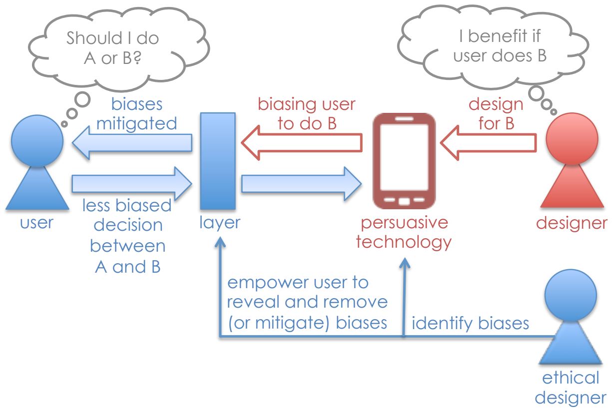


Figure 3: Schematic of our remediation-based approach.

An schematic representation of our approach is presented in Figure 3. It does not require commitment from every designer, like in [8, 12, 10, 14, 18, 15]. Smaller groups of ethically-conscious designers shall be able to implement it as an extra layer interfacing users and today's persuasive technologies. Next we present a few examples about how this remediation-based approach can be realized.

Example 1: Mobile. Consider a user that is thinking about temporarily disabling his/her phone's GPS sensor due to privacy concerns. Many of today's smartphones introduce biases that prevent this behavior. In some cases, the functionality is somewhat "hidden" in the settings' menu, requiring more effort on the user's side to find it (mental/time demand). In other cases, a discouraging message is displayed after disabling the sensor, further asking for the user's confirmation (loss aversion[2]). Our approach could be implemented by designers—not necessarily related to the phone maker—that *identify* the presence of the given biases, and develop a mobile application to empower users to *reveal* and *remove* these biases. In that sense, the mobile application could offer an easy and fast way for users to turn the GPS sensor on and off, such as through a shortcut button or a gesture based interaction.

Example 2: Web. Let us use the famous example of default bias in web forms [6] to exemplify our approach. Consider a user that is buying a flight ticket online using a desktop computer. After s/he makes the flight reservation, the website automatically adds a special customer service for €30, expecting that the default bias will discourage the user from opting-out the service. In this case, designers—not necessarily related to the site owners—could implement our approach by studying the website and identifying the presence of the given default bias; and implementing a browser plugin that can be installed in the users' computers, enabling them to *reveal* and *remove* the default bias. In order to remove it, the plugin could restructure the webpage including an extra modular step that inquiries users on whether they want or not to include the special service.

Example 3: Social Networks. Removing biases is not always possible. In these cases, our approach can be implemented by means of intervention methods that do not specify a target behavior. For example, should users browse social network websites for longer periods of time than they currently do? In this case, our approach could be implemented by revealing the main biases that increase user retention time, and enabling users to activate interventions that mitigate these biases. An example of intervention is to provide daily summaries about the user's time spent in these websites. Other interventions that go beyond self-reflection could prove to be more effective and should be systematically tested, thus opening different avenues of research in persuasive computing.

Conclusions and Future Work

In this paper we proposed preventive and remediation-based approaches to more effectively implement ethics in behavior design. We discussed the scope of each of these approaches, highlighting their advantages and disadvantages. We focused our proposal on a remediation-based approach that can be implemented by small groups of ethically-conscious designers, hence not requiring compliance from every designer, like in most of today's approaches [8, 12, 10, 14, 15].

Many questions remain open for future research and discussion. Besides evaluating our remediation-based

[2]Loss aversion bias affects how we think about our possessions, making the loss of giving up something more salient than the gains of acquiring something else [6]. The message for turning the GPS off focus on what the user loses when turning the sensor off, whereas no discouraging message is usually presented when turning the GPS on.

approach, we would like to investigate to what degree of automation it would be more successful. Although we give focus on a manual procedure for inspection of persuasive technologies conducted by a committee of designers, one could envision a hybrid approach that automatically identify certain biases for later manual inspection.

Another topic for discussion is the qualification required to be a designer of persuasive technologies. One could envision designers being certificated by authorities or organizations trusted for implementing appropriate guidelines and methods, in a similar way that it is done for other domains, like internet security [5] or quality management systems [3]. Alternatively, groups of designers could build their reputation by alerting users of biases present in the services they use everyday [1], and thus avoid going through a potentially long certification process. We believe both cases have potential when designers further provide interactive mechanisms that educate users in revealing biases in technologies, *and* empower users to remove or mitigate these biases.

We hope that our proposed approaches generate fruitful discussions in the HCI community to effectively implement practical solutions for ethical behavior design.

References

[1] Dark patterns. Available at: www.darkpatterns.org.
[2] FDA U.S. Food and Drug Administration. How to Understand and Use the Nutrition Facts Label. Available at: www.fda.gov/food/ingredientspackaginglabeling/labelingnutrition/ucm274593.htm.
[3] ISO - International Organization for Standardization. Available at: www.iso.org.
[4] Personal informatics. Available at: www.personalinformatics.org.
[5] VeriSign SSL certificates. Available at: www.verisign.com.
[6] Ariely, D. *Predictably irrational: the hidden forces that shape our decisions*, 1 ed. HarperCollins, Feb. 2008.
[7] Arroyo, E., Bonanni, L., and Selker, T. Waterbot: exploring feedback and persuasive techniques at the sink. In *Proc. of CHI*, ACM (2005), 631–639.
[8] Berdichevsky, D., and Neuenschwander, E. Toward an ethics of persuasive technology. *Communications of the ACM 42*, 5 (1999), 51–58.
[9] Consolvo, S., McDonald, D. W., Toscos, T., Chen, M. Y., Froehlich, J., Harrison, B., Klasnja, P., LaMarca, A., LeGrand, L., Libby, R., Smith, I., and Landay, J. A. Activity sensing in the wild: a field trial of ubifit garden. In *Proc. of CHI*, ACM (2008), 1797–1806.
[10] Davis, J. Design methods for ethical persuasive computing. In *Proc. of Persuasive*, ACM (2009), 6.
[11] de Oliveira, R., Cherubini, M., and Oliver, N. Movipill: improving medication compliance for elders using a mobile persuasive social game. In *Proc. of Ubicomp '10*, ACM (New York, NY, USA, 2010), 251–260.
[12] Fogg, B. J. *Persuasive technology: using computers to change what we think and do.* Morgan Kaufmann, 2002.
[13] Helft, M. The class that built apps, and fortunes. *The New York Times*, 7 May 2011. Retrieved from www.nytimes.com/2011/05/08/technology/08class.html?partner=rss&emc=rss&_r=0.
[14] Kaptein, M., and Eckles, D. Selecting effective means to any end: Futures and ethics of persuasion profiling. In *Proc. of Persuasive Technology*, Springer (2010), 82–93.
[15] Karppinen, P., and Oinas-Kukkonen, H. Three approaches to ethical considerations in the design of behavior change support systems. In *Proc. of Persuasive*, Springer-Verlag (Berlin, Heidelberg, 2013), 87–98.
[16] Lee, M. K., Kiesler, S., and Forlizzi, J. Mining behavioral economics to design persuasive technology for healthy choices. In *Proc. of CHI 2011*, 325–334.
[17] Smids, J. The voluntariness of persuasive technology. In *Proc. of Persuasive*, Springer-Verlag (Berlin, Heidelberg, 2012), 123–132.
[18] Spahn, A. And lead us (not) into persuasion?? persuasive technology and the ethics of communication. *Science and engineering ethics 18*, 4 (2012), 633–650.
[19] Torning, K., and Oinas-Kukkonen, H. Persuasive system design: state of the art and future directions. In *Proc. of Persuasive Technology*, ACM (2009), 30–37.
[20] Wang, Y., Leon, P. G., Scott, K., Chen, X., Acquisti, A., and Cranor, L. F. Privacy nudges for social media: an exploratory facebook study. In *Proc. of WWW* (2013), 763–770.
[21] Yun, R., Scupelli, P., Aziz, A., and Loftness, V. Sustainability in the workplace: nine intervention techniques for behavior change. In *Proc. of Persuasive*, Springer (2013), 253–265.

Mobile search

The author spent the summer of 2014 working as an intern in Yahoo! Labs in California, under the supervision of Karen Church and Beverly Harrison. His work there focused on mobile search, and produced a paper that was submitted to CHI 2015. It is included next.

J. P. Carrascal and K. Church. "An In-Situ Study on the Interaction between Mobile Apps & Mobile Search". *Proceedings of the SIGCHI conference on Human Factors in computing systems (Submitted)*. 2015

An In-Situ Study on the Interaction between Mobile Apps & Mobile Search

Juan Pablo Carrascal
Universitat Pompeu Fabra
Carrer de Tànger, 08018 - Barcelona, Spain
jp.carrascal@upf.edu

Karen Church
Yahoo! Labs
701 First Street, 94085 - Sunnyvale, CA, USA
kchurch@yahoo-inc.com

ABSTRACT
When trying to satisfy an information need, smartphone users frequently transition from mobile search engines to mobile apps and vice versa. However, little is known about the nature of these transitions nor how mobile search and mobile apps interact. In a 2-week mixed-method field study involving 18 participants we collected both quantitative data on mobile app and mobile search usage as well as qualitative insights on why certain interactions between apps and mobile search occur. Our results show that when people engage with mobile search they tend to interact with more mobile apps and for longer durations. Furthermore certain categories of apps are used more intensely alongside mobile search. We conclude with a concrete discussion on what these patterns mean for mobile search and how we might design mobile search experiences that take these app interactions into account.

Author Keywords
Mobile Search; Apps; Context; Empirical Study

ACM Classification Keywords
H.5.2 Information interfaces and presentation: Miscellaneous

INTRODUCTION
People adopt a range of ingenious methods when trying to satisfy their daily information needs [6], but for smartphone users in particular, the Internet has become the dominant means of finding information. A recent comScore report[1] shows that mobile devices accounted for 55% of Internet usage in the United States in January 2014. Native mobile apps made up the majority of that traffic (85%), while the remaining 15% came from mobile browsers. With a wealth of information seeking apps at their fingertips, mobile users can now issue queries directly from within these apps. For example, Yelp can be used to search for local businesses, Google Maps can be used to find directions, while IMDB can be used to look for movie reviews. Thus the boundaries between mobile search engines and mobile apps have become blurred.

Smartphone users also switch between mobile search and other mobile apps to satisfy their information needs. Brown et al. [4] describes one such switching interaction in which mobile search and mobile maps application are used interchangeable — with search being used to find information about certain places and maps being used to get directions to those places. It's likely that such switching interactions occur across a diverse range of everyday apps, i.e. beyond just maps applications. One of the goals of this work is to understand such switching behaviors.

Past research has also shown that mobile search can be triggered by our conversations and our surroundings [10, 7, 1, 27]. Furthermore, the results of mobile search can feed into next actions and plans [4]. Thus it's likely that some of these mobile search *triggers* and *actions* may relate to or be supported by interactions with mobile apps. For example, a text message from a friend may prompt a search or after searching for a local restaurant, a person calls to make a dinner reservation. The goal of this paper is to explore these interactions and to investigate if and how mobile search and app usage relates.

While patterns of mobile app usage [3, 23, 20, 12] and mobile search [2, 9, 15, 16, 29, 30] have been studied separately, the relationships and interactions between them have not been examined in-depth. A detailed study of in-situ app and search engine use will provide insights into how we can support more enriching cross app and search engine experiences, and assist mobile users in completing their daily information seeking tasks. To shed light on these interactions we conducted a 2-week mixed-method study involving 18 Android users. We combined interviews, a daily online diary and real-world mobile usage data to gather rich insights into mobile information seeking behaviors. Over the 2-week period we collected approximately 54,000 mobile app launches, over 900 mobile search sessions and over 500 diary entries. Our results highlight key differences between sessions of app interactions involving mobile search and sessions of app interactions without search.

Our core contributions are as follows:

- We define and introduce cross app and search interactions as an important and challenging research opportunity in the field of mobile HCI and mobile information retrieval.

[1] Mobile apps overtake PC Internet usage in U.S., See: cnnmon.ie/1cfHe0Z

Submitted to CHI 2015

- We characterize app and search interactions from a unique perspective, combining both real-world in-situ search and app usage with qualitative insights. We explore both temporal and topical aspects of these interactions.
- We discuss what our findings mean for future mobile information seeking services.

RELATED WORK

There are 3 key areas of related work: (1) studies of mobile information needs; (2) studies of mobile web and mobile search; and (3) studies of mobile app usage.

Research has shown that a significant proportion of daily information needs arise when users are away from familiar contexts [11, 6]. For example, 67% of needs occur when users are away from home or work [8]. Mobile needs are prompted by explicit contextual factors including activity, location, time and conversation. For example, Sohn et al. [24] found that 72% of needs are influenced by context, while Hinze et al. [14] found that geographical questions are more prevalent when mobile. The act of fulfilling information needs is also affected by contexts [5], with location being particularly influential [17]. Access to mobile internet and gender also play a role in the means chosen to address information needs [6]. Overall, research has shown that the internet is a dominant means of addressing mobile information needs[24] in particular among experienced mobile Internet users [13].

Other research looks beyond the needs that mobile users encounter and towards how and why people access and consume online content via their mobile device [18]. Qualitative studies have highlighted that the majority of mobile Web access occurs in stationary, familiar environments like home and work [22, 7]. The main motivations for general mobile Web usage are awareness and staying up-to-date [26], while mobile search motivations are more inclined to relate to curiosity [7]. Mobile search is a social act, often sparked by conversations [10] and conducted in the presence of other people [7]. Similar social influences have been found for local mobile search [1, 27].

Other research has taken the form of analyzing Web usage patterns of real mobile phone users [2, 9, 15, 16, 29, 30]. These studies use datasets from commercial search engines like Google as well as operator-specific search services. Overall these studies highlight that mobile queries tend to be shorter than their desktop counterparts, users tend to submit fewer queries per session compared to desktop users and that adult related content is very popular. Recent studies show signs of evolving patterns, highlighting that the increased popularity of smartphones is having an impact on search behavior. For example, mobile queries are becoming less homogenous, mobile query length is steadily increasing and interest in local content is rising. Finally, research from Google [16] and Bing [25] highlights differences in search based on devices like smartphones, tablets and desktop, e.g. iPhone users search in ways that are similar to desktop users [16].

Finally, there have been studies aimed at understanding how people engage with native mobile apps. Böhmer et al. [3] studied app use among over 4,100 Android phone users and highlight the prevalence of communications related apps throughout the day, with certain categories of apps being used more often at certain times, e.g. Clock in the morning. Rahmati et al. [23] conducted a year-long study of 34 college students using iPhones to investigate how users in different socio-economic groups adopt new smartphones and highlight that 40% of all app usage comes from a single application, with $> 90\%$ of usage from the top 10 applications. Tossell et al. [28] look at Web use on smartphones and find that native internet applications like Mail, Facebook, Maps and Weather are visited twice as much as browsers.

McGregor et al. [20] use a video-data collection method and record 100 days of iPhone use. They highlight the 'occasioned' nature of smartphone use, pointing to 3 distinct usage patterns: *micro-breaks*, *digital knitting* and *reading*. Lee et al. [19] explored if and how app usage relates to smartphone overuse and addiction in college studies. Most recently Ferreria et al. [12] look at app micro-usage and find that 40% of all app launches last less than 15 seconds and such short interaction happen mostly when a person is at home and alone.

Of particular relevance to this paper is work by Brown et al. [4] who explore web search and maps use. Using innovative video data collection of everyday smartphone use, they focus on what prompts the use of particular applications at specific times or in specific situations. They highlight 'occasioned search', that is search triggered by the environment or local events. Our goal is to expand upon these insights to investigate the triggers of mobile search in more detail and to explore the unique interaction between triggers and general mobile app usage. Brown et al. also highlight the use of the mobile Web and mobile maps interchangeable in information searches. In this paper we examine such switching interactions looking at search and the range of mobile apps smartphones users engage with daily, i.e. beyond just maps applications.

In summary, past work has explored mobile search and mobile app usage primarily in isolation. In this paper we present an in-depth investigation about how mobile search and app usage is interlinked. We build upon previous research, focusing in particular on the triggers, actions and app interactions in and around mobile search. Furthermore, we study this behavior from a unique perspective, combining both real-world in-situ search and app usage with qualitative insights.

STUDY METHOD

We conducted a 2-week in-situ field study in June and July 2014 where we collected both objective and subjective data. Objective data was collected via two apps installed on participant's smartphones, which tracked their actual mobile app and mobile search usage. Subjective data was gathered at the start and end of the study in initial and final in-person interviews as well as daily via an online diary where participants reviewed and described specifics of their mobile searches.

Participants

We recruited 18 participants (10 male; 8 female) from 8 different cities in the Greater San Francisco Bay Area using a professional recruiting agency. Participants ranged in age between 18-48, with an mean age of 32.7 ($s = 9.8$) and had a diverse set of occupations including students, administrative assistants, social workers, managers, chemists, homemakers and construction workers. Their education levels ranged from high school to college degree. All participants were active users of mobile search engines like Google and Yahoo. All of them owned Android smartphones as their main communication device, the majority of which were Samsung Galaxy phones (14). Participants were compensated for the time.

Procedure

The study was conducted between June and July 2014 and was comprised of three stages:

1. Initial interview and app installation

The initial interview was semi-structured and lasted from 30 minutes to 1 hour, depending on the scope and diversity of participant's mobile search and app usage. The interviews covered (a) their daily mobile device habits; (b) their general mobile search engine use; (c) concrete examples of their most recent mobile search, focusing on aspects such as what triggered the search in the first place and any actions that took place as a result of the search; and (d) apps they use for searching and situations in which they choose an app over a search engine for satisfying their needs. To conclude the interviews, we installed two custom apps on their smartphones:

AppLogger: a logging app that ran as an android service in the background of the participants phone and kept track of all their mobile app usage. This tool collected time stamped usage data, specifically: which app was currently active, the time it was launched, and how long that app stayed active in the foreground.

MSearch: a mobile search app which embedded the search functionality of a well known commercial search engine. This app collected time stamped search data including search queries and search interactions (i.e. clicks) Participants moved the app icon to their homescreen during the initial interview to ensure the search app was easily accessible.

2. Two-week mobile search and online diary study

Participants were asked to use the MSearch app for all their mobile search needs over the 2-week study period and to review their mobile searches in a Web-based online diary tool. All queries and clicks generated through our search app were sent to a server for processing. Queries and clicks were grouped into sessions using a session delta of 5 minutes and these search sessions formed the basis of the participants' diary. The online diary was designed to capture the motivations, triggers and actions surrounding their mobile searches while minimizing the time burden on the participants. To achieve this balance the diary presented participants with a maximum of 3 of their mobile search sessions per day, which were selected at random. Following is a list of the questions we asked each user to answer for each search session. In parenthesis we indicate if a question was multiple choice (closed), or freeform text (open):

1. What information were you looking for? (open)
2. What were you doing at the time of the search? (open)
3. Where were you at the time and did your location influence the search? If so, how? (closed/open)
4. Who were you with at the time and did the people around you influence the search? If so, how? (closed/open)
5. Did you share the information you found with other people? If so, how? (closed/open)
6. How important was it to find the information you searched for? (closed, 5-point Likert scale)
7. How urgent was it to find the information you searched for? (closed, 5-point Likert scale)
8. Could you find the information? If not, what alternative approaches did you try to find the information? (closed)
9. What did you do with the information you found during the search? Did you take any additional actions? (open)

Note that participants were encouraged to access their online diary each evening. Participants were also sent a daily email for the duration of the study to remind them about the study.

3. Final interview and app removal

At the end of the 2-week study, participants attended a final in-person interview. Prior to the interview, we reviewed their mobile search and mobile app usage log data as well as their diary entries to extract specific usage behaviors we wanted to follow up on or get more details for. We probed participants about two or three of their reported mobile searches and asked them about the triggers and actions associated with those searches. We also asked about any changes in their mobile app usage in the past two weeks, e.g. deleting of any apps. To conclude, both the AppLogger and MSearch apps were uninstalled from the participants phone. All in-person interviews were audio recorded and transcribed.

RESULTS

Quantitative data from both mobile search and mobile app usage was analyzed both per-session and in aggregate to understand the types and nature of mobile search interactions; the types of mobile apps that were used and for how long; and the interaction between mobile app usage and mobile search. *Qualitative data* from the transcribed interviews and the diary responses were analyzed using grounded theory-based affinity analysis. This approach is commonly used to organize and group large quantities of subjective data into a logical set of themes or categories. We extracted over 3,000 individual quotes from the qualitative data, which made up individual data items in our analysis. These data items were then iteratively reviewed and grouped by two researchers to find repeating themes across participants.

Basic Descriptive Statistics

The results reported in this section are based on the mobile app and search usage logs gathered between 23rd June and 13th July 2014. Although the mobile search part of the study lasted 14 days, the *AppLogger* ran for an average of 16 days ($s = 2.57$) due to scheduling final participant interviews. Before we can explore the interactions between mobile app usage and mobile search, we much first analyze app usage and mobile search separately.

Table 1. Number of unique apps, app launches and average usage time of apps group by app category

Category	# Launches	% Launches	# Unique Apps	Avg Dur(secs)	App Examples
Social networking	9714	17.98	17	107.6	Facebook, Twitter, Instagram, OkCupid
Browser & Search	7727	14.30	10	112.9	Chrome, Google Search, Firefox
SMS/Texting	6136	11.36	3	63.0	Build in Messaging
Phone & Audio	5456	10.10	10	25.8	Dialer, Google Voice, Skype, Voxer
Email	3928	7.27	4	54.0	Gmail, Yahoo mail, SolMail
Games	2780	5.15	73	264.1	Slotmania, CandyCrush, Words
Contact	2574	4.76	4	53.0	Contacts, Mr. Number
Music & Audio	2468	4.57	16	74.0	Pandora, iHeartRadio, SoundCloud
Photography	2172	4.02	19	36.4	Shutterfly, Gallery, Camera, PicCollage
System & Settings	1904	3.52	12	29.5	Software updates, Android Settings
Tools & Utilities	1777	3.29	45	30.8	Calculator, Clean Master, Media Storage
Instant Messaging	1659	3.07	9	79.3	Snapchat, ChatOn, Kik, Trillian
Media & Video	1338	2.48	10	275.8	YouTbe, Netflix, BitTorrent
Productivity	1169	2.16	24	61.4	Calendar, Quickoffice, LastPass, Gnotes
Travel & Local	1035	1.92	13	111.6	Maps, Muni Alerts, Yelp
Entertainment	370	0.68	19	239.7	Meme Generator, Series Guide, IMDB
Unknown	346	0.64	23	142.5	*Not applicable*
Shopping & Retail	254	0.47	12	138.3	eBay, Groupon, Macy's, LivingSocial
Finance	252	0.47	15	58.1	Wells Fargo, Chase, PayPal, Mint
Health & Fitness	228	0.42	10	156.7	S Health, Fitit, Push Ups, My Diet Coach
Education	181	0.34	1	22.2	SMCC
Lifestyle	169	0.31	12	65.8	dscout, 7-Eleven, AIDS Walk
Weather	146	0.27	5	19.2	The Weather Channel, Weather Widget
News & Magazines	124	0.23	7	73.0	Flipboard, yahoo, Glamlife
Business	71	0.13	7	112.5	VPN Client, CamCard, Job Search
Personalization	28	0.05	7	129.9	Zedge, Live wallpaper, travel wallpaper
Sports	9	0.02	1	47.8	theSocre
Books & Reference	7	0.01	6	49.4	Dictionary, Wikipedia, Audible

Mobile App Usage

Over 189K mobile phone usage events were logged over the entire study duration with an average of 10,515.3 events per participant ($s = 6124.7$). Of these, 54,022 corresponded to app launches, with an average of 3,001.2 app launches per participant ($s = 2,003.9$). The remaining events were device events such as turning the display on/off, unlocking the phone and accessing the homescreen. A total of 394 unique mobile apps were launched across our 18 participants over the course of the study, with an average of 52 unique mobile apps launched per participant ($s = 20.6$).

Compared to prior work, which shows that users spend an average of 1 hour using mobile apps [3, 21], our participants spend significantly more time interacting with apps. The average time our participants spent interacting with mobile apps is > 4.5 hours ($s = 248.2$ mins). In contrast, Table 1 highlights that individual mobile app usage is often short, lasting an average of just 90 seconds ($s = 241.5$), slightly longer than Böhmer et al.'s 71 second average. Overall we found that 73.5% of app uses were one minute or less in duration and approx. 40% of uses are 15 seconds or less which is in line with findings by Ferreria et al. [12].

To get a better sense of the types of apps used, we manually classified the 394 unique apps into corresponding Google Play categories. We did this mapping by searching the Play store based on app name. Similar to the Böhmer et al. study we made some minor changes to the Google Play categorization. Specifically, we opted for one high-level games category instead of multiple micro-categories of Games (e.g. arcade, brain & puzzle, etc.). We opted to have a separate Browsers & Search category and a separate Systems & Settings category for handling the default Android settings apps. Finally we opted to break-out communications-related categories such as Contacts, SMS/Texting and Instant Messaging. Table 1 highlights that the top 3 categories are *Social networking*, *Browsers & Search Engines* and *SMS*, while almost 55% of app usage in our study relates to *Communications*.

Table 2. Distribution of search topics across all 882 unique queries

Topic	# Queries	% Queries
Entertainment	144	16.3
Trivia	144	16.3
Local	133	15.1
Shopping & Coupons	106	12.0
Travel & Commuting	84	9.5
Technology	49	5.6
Health & Fitness	45	5.1
General Information	42	4.8
Cooking, recipes & ingredients	28	3.2
Sport	24	2.7
Auto	24	2.7
Misc	17	1.9
Search & Navigational	12	1.4
Stocks & Finance	12	1.4
News & Weather	10	1.1
Employment	6	0.7
Education	1	0.1

Mobile Search Usage

Participants issued 882 unique queries through the *MSearch* app resulting in 2794 webpage visits. These webpage visits include both click-throughs and follow-on links. This corresponds to an average of 50.1 unique mobile search queries ($s = 47.5$) and 158 webpage visits ($s = 122.2$) per participant over the 2 week study period.

A mobile search session is a sequence of queries and search interactions (i.e. clicking on the next page of results of clicking on a individual search result) issued by a single user within a small time period. Using a standard mobile search session delimiter of 5 minutes [15], we identified a total of 843 search sessions. We found an average of 1.6 unique queries per session ($s = 1.18$). Approximately 78% of sessions resulted in at least one click-through (647 sessions) while the average number of webpage visits (click-throughs and follow-ons) per session is 4.9 ($s = 5.7$).

Compared to previous studies of mobile search patterns we observe that our participants issued longer search queries. The average number of words per query is 3.55 ($s = 2.1$), compared to 3.05 reported in a 2013 Bing study [25]. Finally to get a better sense of the types of queries our participants issued, two researchers manually classified all 882 unique search queries into a set of search categories. Cohen's kappa was used to measure intercoder reliability. After two iterations of independently coding 200 unique search queries and discussing any conflicts, a Cohen's Kappa value of 0.74 was reached. The remaining dataset was then divided, and each coder independently coded his or her part of the dataset. The resulting classification is shown in Table 2. We find that participants in our study predominately issued mobile search queries related to *Entertainment*, *Trivia*, *Local*, *Shopping & Coupons* as well as *Travel & Commuting*

The Online Diary

Participants contributed a total of 535 diary entries, equivalent to 29.7 entries per user over the course of the study. According to their answers, most of the searches were conducted at home (70.1%), followed by work/school (7.7%) or commuting (7.7%). Other locations mentioned included bars, cafes or restaurants, someone else's home, the gym, shopping and running errands. In 26% of the cases, participants stated that their location at the time influenced their search.

In contrast to past research on mobile search, we found that participants reported that they were mostly alone when searching (51.2%). The rest of the time they were with their partner or spouse (22.4%), relatives (12.9%), friends (7.3%), work colleagues or school mates (3.9%) or others (2.2%). In 15.5% of the time, the person accompanying the participant had an influence on searching.

Most of the time the information sought was found (80.6%). In 17.4% of cases the participants considered the information to be partially found. When participants couldn't find the information, they either did nothing (29.4%) or relied on alternative methods including: asking somebody (23.5%), using other search engines (23.5%), reading the information somewhere—in a map, etc.—(2.9%), or something else (20.6%). Finally, in 17.4% of cases, the information found was shared with somebody else.

Cross App and Mobile Search

To explore the relationship between mobile search and mobile app usage we must first define cross app and search interactions. Figure 1 presents the sequence of actual mobile app and search interactions for a 1 hour period (12-1pm) on a single day for a participant in our study. It shows that a user's day is typically comprised of a number of *"sessions"*. We define a session as a sequence of interactions that occur without the device being in standby mode, *i.e.* the display switching off, for longer than 30 seconds.

Some of these sessions involve the user turning their display on without actually launching any apps. For example to check the time or to check for any missed notifications, both of which would be visible on their lock screen and/or homescreen of the phone. In other sessions, the user launches and interacts with specific mobile apps. Sometimes this may be a single app launch, while in other cases this can involve opening a chain of apps. The diagram highlights that this particular user interacts with her device regularly throughout the hour but more intensely in the first 30 minutes. She primarily uses SMS but also plays a number of games (e.g. Pet Rescue Saga), browses the Internet and interacts with Facebook. We define sessions in which at least one app was launched as *"App Sessions"*.

The figure also highlights that the participant engaged with mobile search on two occasions. First submitting a query for 'advance screenings san francisco' and visiting a website (advancescreenings.com). Minutes later she submits a query 'rotten tomatoes' and visits the associated rotten tomatoes website. These queries (Q) and search clicks (C) are identified in bold in the figure. Thus some *"App Sessions"* involve the user engaging with mobile search, while other app sessions do not. We define these as $AppSessions_{search}$ and $AppSessions_{nonsearch}$ respectively.

Note that this particular use interacts with SMS before, during and soon after her mobile searches. This implies that her overarching task — deciding on a movie — involves cross app and search interaction. The goal of this work is to understand more about these cross app and search interactions. In the following section we dive into our analysis.

Are Search and Non-Search Sessions Different?

To investigate the mutual influence and interaction between mobile search and mobile app usage, we explore differences between $AppSessions_{search}$ and $AppSessions_{nonsearch}$ in terms of app launches, unique apps used and session duration. To do this we extracted a list of all *app sessions* in our dataset, that is sessions that include at least one launch of at least one mobile app. We excluded the first app session for every participant from all subsequent analyses since it corresponds to the installation of the *MSearch* app which was done during the initial interview. We found a total of $12,307$ app sessions in our dataset. Our *MSearch* app was used in 913 sessions, and in remaining $11,394$ it was not. We therefore split the dataset in two: 913 $AppSessions_{search}$ and 11,394 $AppSessions_{nonsearch}$ for comparison purposes.

Figure 2 shows the distribution of $AppSessions_{search}$ (red) and $AppSessions_{nonsearch}$ (blue) grouped by the total number of app launches. The figure suggests that app sessions involving mobile search include a higher number of app usages compared to non-search sessions. A non-paired Welch's t-test confirmed that there is a significant difference between the two app session groups ($t = 8.14, p < 0.01$, Cohen's $d = 0.72$). Note that Cohen's d measures the magnitude of the difference. Figure 2 shows a similar pattern for the

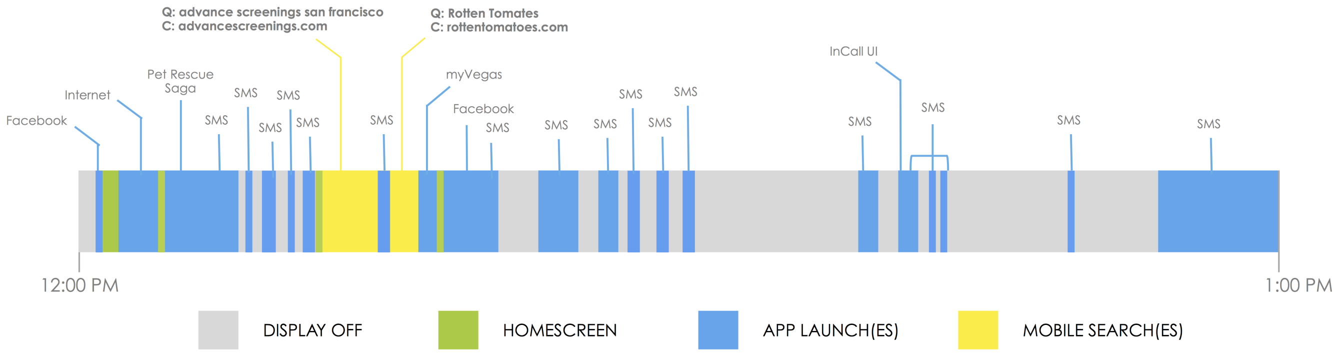


Figure 1. Participant timeline for a subset of a single day showing sequences of mobile device interactions in the form of app launches, home screen interactions and mobile searches.

number of unique apps used in search *vs.* non-search app sessions. That is when users engage in $AppSessions_{search}$, they interact with a higher number of unique mobile apps. We tested this and again found this difference to be significant ($t = 18.68, p < 0.01$, Cohen's $d = 1.18$). This implies there is greater diversity in their behaviors during search sessions.

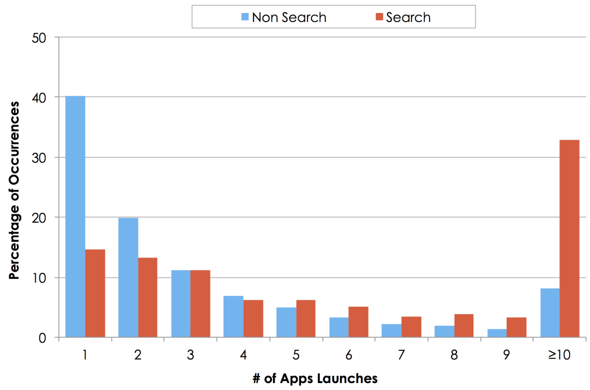


Figure 2. Number of app launches in $AppSessions_{search}$ **vs.** $AppSessions_{nonsearch}$

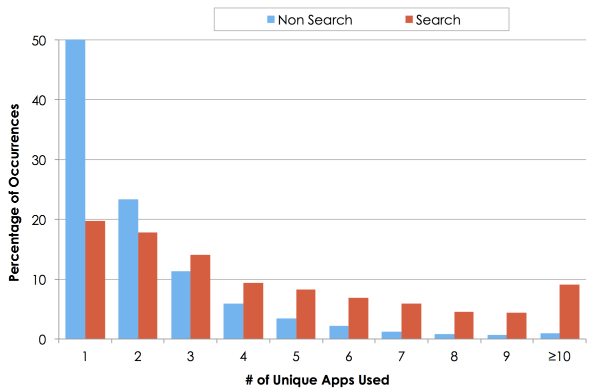


Figure 3. Number of unique apps used in $AppSessions_{search}$ **vs.** $AppSessions_{nonsearch}$

Finally we explored temporal differences, firstly in terms of overall session duration. Figure 4 highlights that app sessions involving search typically last longer than app sessions with no search. The average duration of $AppSessions_{search}$ is 1473 seconds (median=367, $s = 5947.05$) compared to 348.2 seconds (median=52, $s = 1275.6$) for $AppSessions_{nonsearch}$. This difference was also found to be significant ($t = 6.28, p < 0.01$, Cohen's $d = 0.57$) between the $AppSessions_{search}$ or $AppSessions_{nonsearch}$ datasets. Overall, these results indicate that $AppSessions_{search}$ tend to be more app-intensive than $AppSessions_{nonsearch}$. To investigate why these differences might occur, we turn our analysis to app interactions within these sessions to see if certain apps or certain categories of apps tend to be used more intensively in and around mobile search activity.

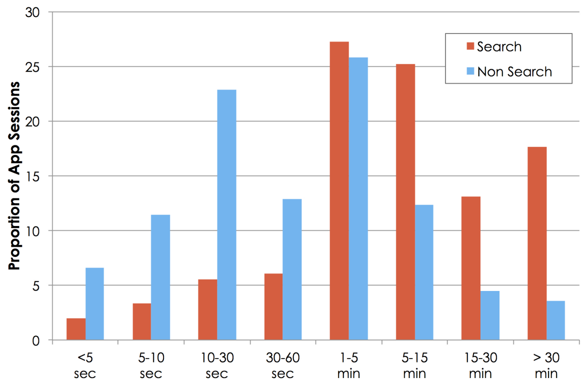


Figure 4. Session duration of $AppSessions_{search}$ **vs.** $AppSessions_{nonsearch}$

How are Search and Non-Search Sessions Different?

In this section we investigate differences in usage of different categories of apps between $AppSessions_{search}$ and $AppSessions_{nonsearch}$. We opted to make these comparisons at a category level as opposed to single app level for two reasons: (1) because of individual preferences, different participants used different apps from the same category (*e.g.* Gmail and the Android email client, the native Internet browser and Google Chrome); and (2) we wanted to limit the influence of apps that are only used by single participants.

To perform the comparisons, we extracted both the frequency of app launches and the usage time of every category of app on a per-session basis. Then we compared the frequency and duration of categories of apps in $AppSessions_{search}$ and $AppSessions_{nonsearch}$. Comparisons were assessed with a non-paired Welch's t-test. Table 3 highlights that we found a significantly higher launch frequency and significantly longer usage time in $AppSessions_{search}$ compared to $AppSessions_{nonsearch}$, for certain app categories. We find that categories of apps including Browsers, Email, SMS, Social Networking, Shopping & Retail and Entertainment were all used more intensively — both in terms of number of app launches and duration of app usage — when people engaged with mobile search. This implies that when users are in *information seeking* mode, they engage more heavily with other apps.

Table 3. T-test (non-paired) results for comparison of launch frequency and usage time of app categories between search and non-search sessions

	Launch frequency		Usage duration	
Category	**t***	**Cohen's d**	**t***	**Cohen's d**
Browsers	15.10	1.29	23.19	1.60
Email	3.80	0.16	8.80	0.57
Entertainment	2.62	0.23	4.12	0.35
Finance	1.98	0.10	2.02	0.10
Games	5.15	0.42	6.06	0.44
Media & Video	1.98	0.15	3.02	0.21
Photography	2.22	0.10	2.84	0.14
Shopping & Retail	2.60	0.20	3.90	0.23
SMS/Texting	3.17	0.23	4.50	0.26
Social networking	3.81	0.29	3.87	0.26
Tools & Utilities	2.80	0.30	4.94	0.33
Unknown	2.91	0.16	2.93	0.21
Weather	2.87	0.13	3.39	0.17

*All coefficients are significant at $p < .05$.

Triggers & App Interactions

Through qualitative analysis of diary and interview data, we identified 6 key triggers. While understanding the prompts of mobile search is not new, our goal here is to shed light on the relationship between triggers and app interactions. The 6 triggers we identified were either: external and internal. External triggers are sensory stimuli, that is things you can see, touch or hear. In contrast internal triggers are connected with our thoughts, emotions, body, or pre-existing habits. External triggers included (1) media, e.g. watching tv, listening to music, reading a newspaper; (2) conversations, face to face as well as as conversations across messaging, email and social networking app; (3) Tangible external triggers, that is noticing something material in their physical surroundings; (4) Activities & Events; (5) Physiological signals like hungry, stress, etc. and finally (6) State of Mind, e.g. random thoughts.

To understand more about how app interactions might trigger mobile searches we extracted the first apps launched within a search session (excluding our search app) and find that Facebook (8.3%), VEVO (4.8%), Calendar (3.2%), Messaging (3.1%) and Gmail (3%) were the top apps launched first within search sessions. We decided to take a manual look at the data to see if we could make any connections between the first app launched in a session and the query (mobile search) immediately succeeding that app interaction.

P5, for example uses SeriesGuide which helps people manage (re)watching their favorite TV shows. Looking at the data we found that P5 issued a query 'Friday night dinner tv wiki' immediately after an interaction with SeriesGuide. Thus the topic of P5's query after interacting with SeriesGuide, an entertainment app was also entertainment related. P1 issued two media & music related queries 'video2mp3' and 'worldstar hiphop' immediately after launching his music player. P18 for example, issued the query 'closer Hollywood sign' while vacationing in Los Angeles immediately after interacting with Google Maps. That is she issued a query with local intent after interaction with a location-based maps application. P8 interacted with Yelp and immediately afterwards searched for 'ca dmv wait time'. Again this highlights a connection between a location-based app and a location-based query. Finally P4 queried for 'bcbg generation jelly thong sandals' a type of footwear after interaction with Macy's, the department store app. Again her interaction with a shopping app appears to have sparked a search for a specific product.

While these interactions are not generalizable across all our users, nor are they indicative of the interactions of all smartphone users, they do provide anecdotal evidence of real connections between apps and mobile search. It appears that interactions with apps can in fact prompt mobile searches.

Apps Before & After Search

Next we dig deeper into mobile search sessions and explore differences in app usage *before* and *after* mobile search. We find 8391 app launches within $AppSessions_{search}$, 2715 (32%) happen *before* the first launch of our app within sessions, and 4746 (57%) occur after the first launch of our app within sessions. The remainder 11% relate to launches of the *MSearch* app. Additionally, we found that out of the 913 $AppSessions_{search}$, 338 (36.3%) sessions *began* with a launch of the *MSearch* app. This indicates that for $AppSessions_{search}$, there is frequently more app activity after mobile search than before.

To explore differences in pre and post search interactions, we extracted and compared the frequency of app launches per app category before the first appearance of the *MSearch* app and after the last appearance of the *MSearch* app, on a per-session basis. This allowed us to obtain additional insights about how certain categories of apps might be related to the search triggers and actions. We assessed this by a paired t-test comparing the launch frequency of every app category before and after the search activity. We found significant differences for some categories, in particular for *after* mobile search interactions (Table 4). These differences indicate that in search sessions, certain app categories tend to be launched more frequently *after* mobile search has taken place. Overall we found the use of browsers is more frequent after mobile search. Likewise communications apps like Email, SMS and Phone & Audio are used more often after a mobile search than before.

Based again on data from interviews and diary responses we identified 9 key mobile search actions, namely: (1) consuming content, i.e. watching, reading, listening; (2) sharing information/content; (3) keeping information which involved saving it digitally or taking down notes as well as making

Table 4. T-test (paired) results from comparing launch frequency of app categories before and after the search activity

Category	t^*	Cohen's d
Browsers	6.63	0.31
Email	4.38	0.20
Games	2.74	0.13
Phone & Audio Communication	2.35	0.11
Photography	3.65	0.17
SMS/Texting	2.62	0.11
Social networking	2.06	0.10
System & Settings	2.25	0.11
Tools & Utilities	3.57	0.17
Unknown	2.49	0.11

*All coefficients are significant at $p < .05$.

Table 5. Top 10 apps (in terms of launch frequency) in sessions that involve 2 or more launches of the $MSearch$ app

No	App	Freq	Perc
1	Msearch	1001	18.0
2	Facebook	572	10.3
3	Internet	306	5.5
4	Email	266	4.8
5	Messaging	196	3.5
6	GO SMS Pro	174	3.1
7	Chrome	174	3.1
8	Messages	165	3.0
9	Gmail	164	3.0
10	System	136	2.4

mental notes; (4) buying goods/services; (5) booking something; (6) planning something; (7) visiting somewhere; (8) contacting a person, business or place; (9) making/doing, e.g. cooking a recipe. Again drawing from the qualitative data, we find that some of these actions indeed map directly to app usage. For example:

- *P4 "I downloaded new music to my phone before proceeding to the gym"*
- *P8 "I ended up Youtube-ing it, and I watched the video"*
- *P2 "I called store and confirmed they have what I'm looking for and will stop by after work Tuesday to buy the item".*
- *P1 "I posted the picture of a Mohawk warrior on Facebook."*
- *P6 "I sent information back and forth to sister by Voxer."*

To date we've shown key differences in $AppSessions_{search}$ compared to $AppSessions_{nonsearch}$. We have highlighted the prevalence of certain categories of apps in and around mobile search and provided insights on how search and apps relate. In this section we have focused on app interactions before and after mobile search. In the following section we provide some insights into the often complex nature of switching between search and other apps.

Switching between Apps in Search

We found that 337 out of the 913 total $AppSessions_{search}$ involved 2 or more launches of the $MSearch$ app. This implies that that participants may have used mobile search, switched to another mobile app, and later switched back to mobile search. Table 5 shows the top 10 apps (in terms of app opens), within these 337 sessions. We find that most of these top apps relate to communications, primarily email, texting and social networking.

In the interviews, participants shed light on their sometimes very complex switching between apps and services to either address their information needs or to share their findings. P8 for example is an avid coupon user and actively uses apps like Groupon and LivingSocial to find discounts and deals. He explained an interesting example of how he searches and interacts when looking for concert tickets: *"I will actually go to Ticketmaster, see the price and then open up another window or another Groupon like that and then if I get one [groupon code], I'll copy it. Get out of there and go back to Ticketmaster to where it usually has a little icon where it says you can enter a coupon code, so paste it."*. At times when he can't find tickets, he'll also visit sites like craigslist. This highlights the disjoint and fragmented sequences of interactions involved. Groupon gives him discount codes, Ticket Master enables the purchase, while search engines allow him to find upcoming concerts in the first place.

P6 described sequences of interactions using Voxer with her sister *"Voxer is her favorite, and so, she had me sign up for Voxer so that she can communicate with me, which is actually kind of cool because while I'm searching for stuff, we can talk back and forth or we could text or we can send pictures between us"*. In an attempt to find information on a local beach she explained how her searching the Web and communicating with Voxer were interlinked, *"I will give her the information and she'll look at it, and then she'll text me something back and then we'll both look at it, so we were just going back and forth with ideas..."*.

Search: Apps vs. Search Engines

In the interviews we asked participants if and why they use other apps like Yelp, Maps and IMDB, etc for searching. In general it appears that search engines are used for broad information needs, while apps are used for more specific questions. One participant likened the interaction to starting with search and drilling down with apps. Search engines were also seen as offering more options, enabling participants to cast a wide net. This was particularly useful for shopping related needs where options and prices are important, e.g. *I would look at the search first because then it gives me an idea because sometimes there are some things that talk about prices, like what range and where you would find that price from what app or what location.* Heimonen[13] found a similar trend comparing search to known websites.

DISCUSSION

In this paper we have taken in-depth look at the complex cross app and search interactions surrounding mobile information seeking. Our results indicate that users who engage in search activity interact more with other mobile apps and for longer durations within search sessions. We found significant differences in the categories of apps used within search sessions compared to non search sessions. Browsers, email, SMS, social networking, shopping & retail and entertainment related apps were all used more intensively when people engaged with mobile search.

Using both qualitative insights and quantitative analysis we highlight that some app interactions lead to searches, while other app interactions are used by participants to take action after a search. Anecdotal evidence highlights that categories of app usage and search topics are linked, e.g. an interaction with Yelp leads to additional local search queries issued via a search engine. All in all mobile users use a range of apps and services to find answers to their daily needs.

Our results also show that there is more app activity *after* mobile search than *before*, suggesting that participants tend to start their app sessions with the *intention* of searching. This implyies that interactions with the material world tend to create more information needs and information seeking behaviors than virtual interactions. Our insights into the categories of apps used before and after mobile search, as well as complex switching between apps and search, point to an overarching theme of *task completion*. In our earlier examples, we find that for one participant the task was was going to a concert with his girlfriend. For another participant it was going to a beach with her family. In an attempt to complete these tasks, our participants used multiple information sources and often shared the information found with others to make joint decisions on day to day things. This involves using a range of apps both for the *finding* and *sharing* phases of these tasks. This highlights that tasks are more important that individual mobile app usage and future mobile search experiences should take these interactions into account.

Existing mobile information seeking services are trying to bring mobile users closer to task completion. Commercial search engines like Google already include actions like calling a business directly from a local search result or getting directions to a place. Yelp includes connections with OpenTable, allowing people to make a restaurant reservation from within Yelp. Google's recent partnership with Uber allows Uber customers to book an Uber directly from within the maps application. Such actions will help mobile users to complete their tasks more effectively, but our findings highlight a need for more of these actions, e.g. click-to-coupon or click-to-share.

It's also important to highlight that we have explored mobile search and mobile app interactions within *sessions*, using a 30 second display-off window as a delimiter. But it was clear based on interviews with our participants, that people conduct searches that can span multiple sessions to address a given need, and these sessions can span differing hours, days, weeks, even months. This was mentioned particularly with reference to searching' for bigger events like buying a car or planning a vacation. These tasks take more time and involve more research before a decision is made.

A behavior that's related to these searches for these bigger events in life is '*keeping* & *sharing*'. We were surprised to find that our participants often kept notes / track of the results of their mobile searches. Methods for keeping track ranged from both mental and written notes to digital notes in the form of emails and messages, screenshots of websites or search results as well as mobile bookmarks. In some cases this keeping was done for themselves while in other cases it was shared with others. Sharing was particularly popular in the run up to decisions and purchases which often involved loved ones and family. At other times keeping was done as a means of coming back to the search at a later more convenient point. Overall these insights highlight the need for more collaborative, shared mobile search experiences that support collaborative note-taking and bookmarking.

While many of these insights would not have been possible without a detailed analysis of both the quantitative and qualitative data from our study, there are of course some limitations to our approach. We built a custom mobile search app for the purpose of the study, thus we were only aware of searches conducted within our $MSearch$ app. Our participants may have searched from within other search engines or other apps during the study period. So in studying the interactions between apps and search it should be noted that we may be missing some such interactions. However, that said our analysis shed significant light on the nuances of these interactions. We should also note that our study participants live in the the Greater San Francisco Bay area. While we made every effort to recruit a diverse sample of participants, we are aware that mobile search and app usage patterns may differ in other parts of the world. Thus we would encourage future work in this space in other cities and other countries around the world.

CONCLUSIONS

In this paper we build upon and extend past research on mobile search and mobile app usage, focusing for the first time on the complex interactions and relationships between native apps and mobile search engines. By taking a step back from search and exploring the interactions in and around search, we have explored mobile information seeking in a new light. Rather that reporting detailed suggestions for improving future search experiences, we instead provide a detailed understanding of cross app and search use. As the landscape of smartphones continues to evolve and the lines between apps and search engines continue to blur, we would argue that more studies of this nature will be needed. In particular we would encourage future research to consider search and app interactions across multiple sessions and multiple timeframes to see if interesting patterns emerge.

REFERENCES

1. Amin, A., Townsend, S., Ossenbruggen, J., and Hardman, L. Fancy a drink in canary wharf?: A user study on location-based mobile search. In *Proceedings of INTERACT '09:*, Springer-Verlag (2009), 736–749.

2. Baeza-Yates, R., and Velasco, J. A study of mobile search queries in japan. In *Query Log Analysis: Social and Technological Challenges, WWW 2007 Workshop* (2007).

3. Böhmer, M., Hecht, B., Schöning, J., Krüger, A., and Bauer, G. Falling asleep with angry birds, facebook and kindle: a large scale study on mobile application usage. In *Proceedings of Mobile HCI'11*, ACM (2011), 47–56.

4. Brown, B., McGregor, M., and Laurier, E. iphone in vivo: Video analysis of mobile device use. In *Proceedings of CHI'13*, ACM (2013), 1031–1040.

5. Chua, A. Y. K., Balkunje, R. S., and Goh, D. H.-L. Fulfilling mobile information needs: a study on the use of mobile phones. In *Proceedings of ICUIMC '11*, ACM (2011), 92:1–92:7.

6. Church, K., Cherubini, M., and Oliver, N. A large-scale study of daily information needs captured in situ. *ACM Transactions on Computer-Human Interaction (TOCHI) 21*, 2 (2014), 10.

7. Church, K., and Oliver, N. Understanding mobile web and mobile search use in today's dynamic mobile landscape. In *Proceedings of MobileHCI'11*, MobileHCI '11, ACM (2011), 67–76.

8. Church, K., and Smyth, B. Understanding the intent behind mobile information needs. In *Proceedings of IUI'09*, ACM (2009), 247–256.

9. Church, K., Smyth, B., Cotter, P., and Bradley, K. Mobile information access: A study of emerging search behavior on the mobile internet. *ACM Transactions on the Web (TWEB) 1*, 1 (2007), 4.

10. Cui, Y., and Roto, V. How people use the web on mobile devices. In *Proceeding of WWW '08*, ACM (2008), 905–914.

11. Dearman, D., Kellar, M., and Truong, K. N. An examination of daily information needs and sharing opportunities. In *Proceedings of CSCW'08*, ACM (2008), 679–688.

12. Ferreira, D., Gonçalves, J., Kostakos, V., Barkhuus, L., and Dey, A. K. Contextual experience sampling of mobile application micro-usage. In *Proceedings of MobileHCI'14*, ACM (2014).

13. Heimonen, T. Information needs and practices of active mobile internet users. In *Proceedings of Mobility '09*, ACM (2009), 1–8.

14. Hinze, A. M., Chang, C., and Nichols, D. M. Contextual queries express mobile information needs. In *Proceedings of MobileHCI '10*, ACM (2010), 327–336.

15. Kamvar, M., and Baluja, S. A large scale study of wireless search behavior: Google mobile search. In *Proceedings of CHI'06*, ACM (2006), 701–709.

16. Kamvar, M., Kellar, M., Patel, R., and Xu, Y. Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices. In *Proceedings of WWW'09*, ACM (2009), 801–810.

17. Komaki, D., Hara, T., and Nishio, S. How does mobile context affect people's web search behavior?: A diary study of mobile information needs and search behaviors. In *Proceedings of AINA*, IEEE (2012), 245–252.

18. Lee, I., Kim, J., and Kim, J. Use contexts for the mobile internet: A longitudinal study monitoring actual use of mobile internet services. *International Journal of Human-Computer Interaction 18*, 3 (2005), 269–292.

19. Lee, U., Lee, J., Ko, M., Lee, C., Kim, Y., Yang, S., Yatani, K., Gweon, G., Chung, K.-M., and Song, J. Hooked on smartphones: an exploratory study on smartphone overuse among college students. In *Proceedings of CHI'14*, ACM (2014), 2327–2336.

20. McGregor, M., Brown, B., and McMillan, D. 100 days of iphone use: Mobile recording in the wild. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, ACM (2014), 2335–2340.

21. Nielsen. An era of growth: The cross-platform report. http://bit.ly/1pcoQFU, March 2014.

22. Nylander, S., Lundquist, T., and Brännström, A. At home and with computer access: why and where people use cell phones to access the internet. In *Proceedings of CHI'09*, ACM (2009), 1639–1642.

23. Rahmati, A., Tossell, C., Shepard, C., Kortum, P., and Zhong, L. Exploring iphone usage: the influence of socioeconomic differences on smartphone adoption, usage and usability. In *Proceedings of MobileHCI'12*, ACM (2012), 11–20.

24. Sohn, T., Li, K. A., Griswold, W. G., and Hollan, J. D. A diary study of mobile information needs. In *Proceedings of CHI'08*, ACM (2008), 433–442.

25. Song, Y., Ma, H., Wang, H., and Wang, K. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *Proceedings of WWW'13*, ACM (2013), 1201–1212.

26. Taylor, C. A., Anicello, O., Somohano, S., Samuels, N., Whitaker, L., and Ramey, J. A. A framework for understanding mobile internet motivations and behaviors. In *Proceedings of CHI'08 extended abstracts*, ACM (2008), 2679–2684.

27. Teevan, J., Karlson, A., Amini, S., Brush, A. B., and Krumm, J. Understanding the importance of location, time, and people in mobile local search behavior. In *Proceedings of MobileHCI'11* (2011).

28. Tossell, C., Kortum, P., Rahmati, A., Shepard, C., and Zhong, L. Characterizing web use on smartphones. In *Proceedings of CHI '12*, ACM (2012), 2769–2778.

29. Vojnovic, M. On mobile user behaviour patterns. In *International Zurich Seminar on Communications*, IEEE Communications Society (2008).

30. Yi, J., Maghoul, F., and Pedersen, J. Deciphering mobile search patterns: a study of yahoo! mobile search queries. In *Proceedings of WWW'08*, ACM (2008), 257–266.

Other research and development activities

During the time spent working on this dissertation, the author was also involved in several research projects in tasks that included application development, data processing and analysis, article writing, and presentation at conferences. Next, a brief description of the projects and an explanation author's contributions are included.

Movement of visitors in the Louvre Museum

This study was lead by Yuji Yoshimura and has produced two papers so far:

Y. Yoshimuraa, F. Girardinb, J. P. Carrascala, C. Rattic, and J. Blata. "New tools for studying visitor behaviours in museums: a case study at the Louvre". *Proceedings of the International conference of Information and Communication Technologies in Tourism (ENTER2012).* 2012

Y. Yoshimura, S. Sobolevsky, C. Ratti, F. Girardin, J. P. Carrascal, J. Blat, and R. Sinatra. "An analysis of visitors' behavior in The Louvre Museum: a study using Bluetooth data". *Environment and Planning B* (2014)

This project was aimed at providing insights on the movement of visitors inside the Louvre Museum. This was done by means of the analysis of the visitors' bluetooth footprints obtained by sensors located ant key points inside the museum. The author of this dissertation contributed with the processing and analysis of the data, with the development of tools for querying the dataset, and finally with the writing of some sections of both papers.

SOS project

The SOS, or Signal Orchestration System, is aimed to help the orchestration of learning activities in the classroom. It uses an Orchestration Manager for the teacher and several wearable devices for the students. This project is directed by Davinia Hernández-Leo and has produced a number of papers in technology-enhanced learning conferences and journals. The author of this thesis proposed and developed a new interface for the Orchestration Manager and contributed to a paper based on the SOS that was accepted at the Eight European Conference on Technology-Enhanced Learning (EC-TEL) 2013 in Paphos, Cyprus:

D. Hernández-Leo, R. Nieves, J. Carrascal, and J. Blat. "Signal Orchestration System for Face-to-Face Collaborative Learning Flows". English. *Scaling up Learning for Sustained Impact.* Ed. by D. Hernández-Leo, T. Ley, R. Klamma, and A. Harrer. Vol. 8095. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 560–564

The author also presented the SOS in a demo contest during the conference, obtaining the second place among the the most voted by the conference attendees and obtaining the "TEL Demo special recognition".

Wearable accessories for free play

The author of this dissertation contributed with sound-related advice to a paper that is part of Dr. Andrea Rosales' doctoral dissertation. Dr. Rosales proposes the use of technology enhanced wearable devices for encouraging free play in school-aged children. The paper centers around wearable accessories that allow children to use gestures and movement to play with sound:

A. Rosales, S. Sayago, J. P. Carrascal, and J. Blat. “On the evocative power and play value of a wearable movement-to-sound interaction accessory in the free-play of schoolchildren”. *Journal of Ambient Intelligence and Smart Environments* 6.3 (Jan. 1, 2014), pp. 313–330

TuneMap

TuneMap started as a hack made by the author of this thesis, Guillermo Malón and Alberto González for the Music Hack Day in Barcelona in 2012. It was initially a Web application that allowed the user to explore music by browsing a world map that includes a visualization of the distribution of music artists by city. The project was later continued by the author alone. His work lead to the publication of a paper describing the project and using its dataset to explore the possible relation between latitude and artist density:

J.-P. Carrascal. “TuneMap: an interactive geolocated music information browser”. *Proceedings of the Re-New Digital Arts Festival, Copenhagen, Denmark, 2013.* re-new digital arts forum - ISSN 2245-7801. 2013, pp. 174–179