When Whereabouts is No Longer Thereabouts: Location Privacy in Wireless Networks

THÈSE N^{O} 4928 (2011)

PRÉSENTÉE LE 28 JANVIER 2011

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
LABORATOIRE POUR LES COMMUNICATIONS INFORMATIQUES ET LEURS APPLICATIONS 1
PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Julien FREUDIGER

acceptée sur proposition du jury:

Prof. J. Huang, président du jury
Prof. J.-P. Hubaux, directeur de thèse
Prof. B. Faltings, rapporteur
Dr V. Niemi, rapporteur
Prof. G. Tsudik, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2011

Il n'y a de lumière que dans le noir.

Christian Bobin

Abstract

Modern mobile devices are fast, programmable and feature localization and wireless capabilities. These technological advances notably facilitate mobile access to Internet, development of mobile applications and sharing of personal information, such as location information. Cell phone users can for example share their whereabouts with friends on online social networks. Following this trend, the field of ubiquitous computing foresees communication networks composed of increasingly inter-connected wireless devices offering new ways to collect and share information in the future.

It also becomes harder to control the spread of personal information. Privacy is a critical challenge of ubiquitous computing as sharing personal information exposes users' private lives. Traditional techniques to protect privacy in wired networks may be inadequate in mobile networks because users are mobile, have short-lived encounters and their communications can be easily eavesdropped upon. These characteristics introduce new privacy threats related to location information: a malicious entity can track users' whereabouts and learn aspects of users' private lives that may not be apparent at first. In this dissertation, we focus on three important aspects of location privacy: location privacy threats, location-privacy preserving mechanisms, and privacy-preservation in pervasive social networks.

Considering the recent surge of mobile applications, we begin by investigating location privacy threats of location-based services. We push further the understanding of the privacy risk by identifying the type and quantity of location information that statistically reveals users' identities and points of interest to third parties. Our results indicate that users are at risk even if they access location-based services episodically. This highlights the need to design privacy into location-based services.

In the second part of this thesis, we delve into the subject of privacy-preserving mechanisms for mobile ad hoc networks. First, we evaluate a privacy architecture that relies on the concept of mix zones to engineer anonymity sets. Second, we identify the need for protocols to coordinate the establishment of mix zones and design centralized and distributed approaches. Because individuals may have different privacy requirements, we craft a game-theoretic model of location privacy to analyze distributed protocols. This model predicts strategic behavior of rational devices that protects their privacy at a minimum cost. This prediction leads to the design of efficient privacy-preserving protocols. Finally, we develop a dynamic model of interactions between mobile devices in order to analytically evaluate the level of privacy provided by mix zones. Our results indicate the feasibility and limitations of privacy protection based on mix zones.

In the third part, we extend the communication model of mobile ad hoc networks to explore social aspects: users form groups called "communities" based on interests, proximity, or social relations and rely on these communities to communicate and discover their context. We analyze using challenge-response methodology the privacy implications of this new communi-

cation primitive. Our results indicate that, although repeated interactions between members of the same community leak community memberships, it is possible to design efficient schemes to preserve privacy in this setting.

This work is part of the recent trend of designing privacy protocols to protect individuals. In this context, the author hopes that the results obtained, with both their limitations and their promises, will inspire future work on the preservation of privacy.

Keywords: location privacy, wireless networks, location-based services, ad hoc networks, rationality, game theory, communities

Résumé

Les outils de communication sont de plus en plus rapides, programmables et mobiles. Ces avancées technologiques facilitent notamment l'accès mobile à Internet, le développement d'applications mobiles et le partage d'informations personnelles. Les utilisateurs de téléphones portables peuvent par exemple partager leurs positions géographiques avec des amis sur les réseaux sociaux en ligne. Les futurs réseaux de communication seront sans doute encore plus omniprésents dans les activités quotidiennes. Ainsi, leur ubiquité permettra d'autant plus le partage d'informations personnelles. Ce type de réseaux utilisera entre autres des communications sans fil ad hoc et distribuées. De ce fait, les réseaux de communication transporterons de plus en plus de données personnelles.

Néanmoins, il devient aussi plus difficile de contrôler le partage des informations personnelles. La confidentialité des données personnelles est donc un élément critique des réseaux de communication. Les solutions de sécurité traditionnelles des réseaux filaires ne sont pas toujours adéquates pour la protection des données dans les réseaux mobiles à cause de la mobilité des utilisateurs et de l'écoute aisée des communications sans fil. Ces caractéristiques entraînent de nouvelles menaces, en particulier à l'encontre des données de localisation. Une entité malveillante peut suivre les déplacements de certains utilisateurs et ainsi découvrir des aspects de leurs vies privées qui ne sont pas apparents. Dans cette thèse, nous étudions ces menaces et explorons différentes approches afin de protéger la confidentialité des données de localisation.

Considérant la popularité croissante des applications mobiles, nous commençons par l'étude de la confidentialité des données dans les services de localisation. Nos recherches permettent d'identifier le type et la quantité de données de localisation qui suffisent afin de statistiquement obtenir l'identité et les points d'intérêt des utilisateurs d'un service de localisation. Ces résultats clarifient le danger et ainsi encouragent l'adoption de mesures de protection. Nos expérimentations indiquent que le risque est réel pour les utilisateurs, même s'ils révèlent peu de données de localisation. Ceci souligne le besoin de prendre en compte la sphère privée dans la conception de services de localisation.

Dans la deuxième partie de la thèse, nous proposons plusieurs mécanismes de protection des données dans les réseaux de communication ad hoc. D'abord, nous évaluons une architecture de sécurité basée sur le concept de *zones de confusion* utilisées afin de confondre une entité malicieuse qui chercherait à suivre les déplacements d'individus. Ensuite, nous montrons qu'il est nécessaire de coordonner la création de zones de confusion à l'aide de protocoles de sécurité. Nous considérons des approches de types centralisé et distribué. Etant donné que chaque individu peut avoir différents critères de protection, nous développons un modèle basé sur la théorie des jeux et une approche distribuée afin d'analyser la capacité d'individus rationnels à coordonner leur efforts. Ce modèle permet notamment de prédire la stratégie d'agents rationnels qui désirent minimiser leurs coûts et maximiser leur confidentialité et

ainsi de concevoir des protocoles plus efficaces. Finalement, nous évaluons analytiquement le niveau de protection obtenu à l'aide des zones de confusion en utilisant un modèle dynamique d'interaction entre les utilisateurs mobiles. Nos résultats identifient le potentiel ainsi que les limites de la protection à base de zones de confusion.

Dans la troisième partie, nous explorons les aspects sociaux des réseaux de communication: les utilisateurs peuvent former des groupes appelés communautés basés sur leurs intérêts, leur proximité ou encore leurs relations sociales. Ces communautés peuvent notamment être utilisée dans les réseaux ad hoc afin d'interagir avec des personnes qui ont des intérêts communs. A l'aide d'une méthodologie de défi-réponse, nous analysons les implications sur la vie privée de ce genre de communications. Nous observons notamment que des interactions répétées entre membres d'une même communauté permettent à une entité malicieuse d'identifier les membres de la communauté.

Ce travail vise à développer des solutions technologiques qui protègent les données personnelles des utilisateurs de nouvelles technologies. Dans ce contexte, l'auteur espère que les résultats obtenus, avec leurs limites et leurs promesses, puissent inspirer de futurs travaux sur la protection de la vie privée.

Mots-Clés: protection des données de localisation, réseaux sans fil, réseaux ad hoc, services de localisation, rationalité, théorie des jeux, communautés

Acknowledgements

These years of research proved to be an exceptional learning experience both at the professional and personal level. I greatly benefited from discussions with wonderful people I met along the way.

I am deeply grateful to my supervisor Prof. Jean-Pierre Hubaux for giving me the opportunity to do research in a stimulating environment. With his guidance and support, Jean-Pierre helped me articulate and develop my ideas as well as control my tendency for distraction.

I would like to thank the members of my thesis committee Prof. Boi Faltings, Dr. Valtteri Niemi, Prof. Gene Tsudik and Prof. Jeffrey Huang for their effort in reviewing this dissertation.

Special thanks to my co-authors for considerably contributing to this thesis, Hossein Manshaei, Reza Shokri, Maxim Raya, Nevena Vratonjic, Murtuza Jadliwala, Prof. Mark Felegyhazi, Mathias Humbert, Prof. Jean-Yves Le Boudec, Prof. Peter Marbach, Prof. David C. Parkes, Jacques Panchard, Panos Papadimitratos, Carmela Troncoso and Prof. Claudia Diaz.

These years of research were particularly enjoyable thanks to my colleagues at EPFL, Marcin, Nevena, Hossein, Reza, Maxim, Mark, Murtuza, Mathias, Igor, Naouel, Adel, Maciej, Imad and Michal for challenging my opinions, discussing ideas and reviewing my work. Special thanks to my office-mates Jacques and Hossein for the daily PhD adventure and to Marcin, Nevena, Olga, Zoltan, Wojciech, Adriana, Michal and Kasia for enlightening lunch times.

I am indebted to Hervé, Yves and Marc-André for providing a great computing infrastructure, Angela, Danielle and Patricia for your support with administrative issues, and Holly for pursuing my English education.

I also want to pay tribute to my friends for sharing many great moments; Yannick, Damien, Marc S. and Marc M. for memorable trips; Sébastien and Marius for co-organizing the beer tasting and the pétanque tournament; Projecto Cocktail for the salsa nights; the Bolivian team for an amazing trip to latin america's heart; and my flat-mates Nicolas, Damien, Jeremie, Julien G., Paul, Fabiola, Sarah and Brandon.

I am grateful to my family for their support and encouragement during all my studies, to my mother Yajaira, my father Jean-François and my brothers Thomas and Nicolas. Quisiera agradecer tan bien a mi familia en Venezuela por su apoyo moral.

Finally, I want to thank Alexandra for her love and for sharing some of the greatest moments during these last five years.

Contents

Introduction

L'enfer, c'est les autres.
Huis clos - Jean-Paul Sartre

Wireless communication devices are increasingly integrated into everyday activities. It has only taken two decades for cellular networks to evolve from a marginal feature to a worldwide phenomenon. Cell phones help more than 5 billion people communicate by making calls, exchanging short messages (SMS), emails or photographs, downloading videos and updating their online social network profile. As currently experienced with the surge of applications for handheld devices, mobile communication devices provide tremendous benefits for end users and service providers. Following this trend, the field of ubiquitous computing foresees an era where computing devices will be everywhere and wirelessly inter-connected using using WiFi, Bluetooth, radio-frequency identification (RFID) or near-field communications (NFC). Such pervasive communications offer opportunities to interact in new ways and facilitate the sharing of information.

One important feature of mobile devices is their ability to collect contextual information: cell phones can use sensors such as a microphone, a GPS receiver, or WiFi wireless interface in order to identify users' environment in the real world. This understanding of users' *context* enables mobile devices to infer what users may be looking for and offer, for example, customized services. Location information is an important piece of contextual data commonly available in modern handsets: users can share their location information with third-parties via infrastructure-based communications in return for location-based services [65, 197]. In the future, most information shared on the Internet may have a location coordinate. Another aspect of contextual data is the neighborhood of communication devices: if mobile devices could wirelessly interact, they would know the set of devices in proximity and increase their environment awareness. Technologies, such as WiFi and Bluetooth, are likely to fuel the adoption of so-called peer-to-peer wireless communications. In addition to classic cellular communications, peer-to-peer wireless communications have the potential to enable new services based on the location of users and the devices in proximity, such as vehicular networks [113] and pervasive social networks [1, 3, 4, 10, 189].

These technological advances are admirable in many respects, but also raise fundamental privacy implications. Until recently, users exploited the natural obscurity provided by real-world constraints in order to protect their privacy. However, new communication technologies change the medium of interaction and may thus create new privacy threats [14]. In particular, when users share their precise location information with location-based services, operators of those services learn users' visited locations. Even if users access those services episodically, revealed locations may enable the identification of users interests. Similarly, peer-to-peer wireless communications based on WiFi in ad hoc mode (e.g., WiFi Direct [11] or Nokia

Instant Community [32]) or similar technology (e.g., FlashLinQ [144]) reveal the presence of a given device to eavesdroppers. Unlike communications with location-based services, such communications happen frequently and are easy to observe. This thrust of peer-to-peer wireless technologies will intensify the amount of personal data sent over the air. Third-parties may thus track users' whereabouts with high granularity based on the broadcasted wireless messages. In other words, if users access location-based services or broadcast wireless messages to nearby devices, malicious entities could learn users' whereabouts, thus threatening *location privacy*. Location information only reveals where users are or have been. Yet, a persistent collection of location data can expose aspects of users' private lives that may not be apparent at first. For example, an eavesdropper may obtain users' identities and profile their interests based on users' presence or absence at specific locations. As communication technologies become pervasive, users and network designers face new challenges to limit the undesirable spread of personal information.

In view of this problem, the subject of this thesis is the design of privacy mechanisms that protect the location privacy of users in current and upcoming wireless networks. We explore emerging location privacy threats, investigate the use of privacy-preserving mechanisms and analyze their performance in various conditions. Before describing our approach and contributions, we introduce the notion of privacy in the following section.

Understanding Privacy

Society values privacy. In nearly every nation, constitutional rights seek to protect privacy as it is considered a fundamental right [13, 24, 163, 209]. Multiple reasons motivated legislators to protect citizens' privacy. They boil down to one major concern. In many cases, knowing information *about* an individual means having power *on* that individual. Such asymmetric situation could be misused by governments, corporations or isolated persons that collected private information. Yet, privacy is a concept in disarray [205] as new technologies increasingly simplify the sharing of personal information. Although the development of technology already raised privacy concerns in the past, with photography or landline phones for example [219, 220], the scale of the problem is much larger with the rise of the personal computer, of the Internet and of pervasive communications. End users increasingly share personal data with global service providers such as Google and Yahoo including search and location-based queries.

Some consider that the potential benefits of upcoming technologies may outweigh the possible loss of privacy. Studies show that a larger connectivity in a communication network improves the efficiency of the network to transmit information [201]. In particular, it simplifies the sharing of information, the acquisition of knowledge and the active participation in society. For example, the online encyclopedia Wikipedia is the unique result of global sharing of knowledge. In this setting, privacy can even be considered detrimental as it may be perceived as an antisocial conduct, conflicting with the efficiency of society by making it, for example, more difficult to control misbehavior. In addition, people sometimes suffer from their inability to relate in a meaningful way and may feel isolated from the rest of the world. Pervasive communications have the potential to alleviate this issue by enabling new and perhaps better ways to interact. For example, social media such as Twitter and Facebook provide new ways to socialize and new kinds of publicity [36]. Such tools are often described as privacy invasive applications. Yet, for most people, features outweigh potential privacy concerns. Social media is addictive because of the reinforcement of feeling connected and influential. These tools can

help people exist in a hyper-connected world through the creation of online alter egos and share with a global audience.

Still, there is value in not opening entirely to new technologies: isolation creates the solitude necessary to interact with ourselves. Solitude is crucial to develop reflection, build opinions and ultimately have a private life. After all, the essence of private life is to allow people to actually have something to share, in order to have a public life [133]. Privacy in other words, is essential in the process of identity construction. Hence, users should be able to share private information with others and maintain the *control* of their information, i.e., how the information is used and by whom. In that way, the usage of private information remains in context, used exclusively in the way the data owner had in mind.

The delicate balance between solitude and social life emerged during the Age of Enlightenment expressing that "La liberté des uns s'arrête là où commence celle des autres".[1] Since then, society makes space for individuals because of the social benefits this space provides. Yet, with new communication technologies such as the Internet, personal information is often shared publicly by default: all activities take place in public. In this setting, users of the Internet actively contribute to the globalization of surveillance: their personal data can be collected by third parties, processed in order to profile their preferences and exploited for unintended use. Hence, it becomes crucial to understand privacy in order to design communication protocols that actually help people better communicate.

Contextual Privacy

Privacy is usually defined as the right to be left alone. In information systems, it refers to the ability of users to control the spread of their personally identifiable information (PII). In practice, the definition of privacy is often misinterpreted because of the peculiarities of real world scenarios. Privacy is often treated as a unitary concept, one essential common feature applicable to all scenarios. For example, some consider privacy as secrecy, bringing forth the "nothing to hide" argument [206]. Under this argument, users should not be allowed to perform anonymous actions (i.e., to hide their identity while performing an action). Instead, transparency should be encouraged because innocent people have nothing to hide and do not need anonymity. This prevents the use of anonymity for negative purposes, such as hate speech or Internet trolling. Yet, in other contexts, anonymity is essential to preserve free speech so that people express ideas without facing risks. For example, minorities can express their opinions without risking blame from the majority. Hence, the nothing-to-hide argument speaks to some problems, but not to others and is not a good interpretation of the definition of privacy.

Instead, scenarios in which privacy is an issue may consist of different yet related components, where no feature is common to all. Hence, privacy should be seen as a *multi-dimensional concept* [204, 207], drawn from a common pool of similar elements, very much like the physical resemblance of members of a family [221]. Even if practical situations share common features, each of them may have different privacy issues. Hence, privacy challenges change from scenario to scenario, and every detail matters.

The conceptualization of each privacy problem should thus begin from its particular context [164]. The context helps understand privacy threats and design privacy protections. In particular, the context defines the type of data to protect, how this data is shared and with

[1] One's freedom ends where that of another begins.

whom. Consider, for example, the problem of location privacy in wireless networks: mobile devices share their location with third parties in return for personalized services. The associated threat is the risk that unauthorized third parties link location information to users' identities in order to track their whereabouts, thus jeopardizing their *location privacy*. Duckham and Kulik [82] formally defined location privacy as a special type of information privacy which concerns the claim of individuals to determine for themselves when, how, and to what extent, their location information is communicated to others.

Privacy Protection

To attain privacy, *legal* and *technological* means must be put in place. Privacy regulations can enforce privacy protection at a large scale using legal actions, such as data protection directives and fair information practices. This approach is mostly reactive: regulations are defined after technology is put in place. In addition, corporations have an incentive to slow down the regulation process in order to reduce the potential costs to comply with new regulations and to maintain higher degrees of freedom. Another aspect of the problem is that the legal system is not consistent across different countries or regions, whereas, communication technology is mostly the same everywhere. This leads to situations where laws protecting privacy are different from country to country. Consider for example the ongoing debate on the status of GPS locations [84]: in some US states, location tracking necessitates a warrant; in other states, its legal status is still unclear.

To avoid this issue, privacy-enhancing technologies (PETs) [55, 109, 143] can be incorporated into the design of new communication protocols. PETs protect privacy by eliminating or obfuscating personal data, thereby preventing unnecessary processing, misuse or involuntary loss of data, without affecting the functionality of the information system [215]. Their objective is to make it difficult for a malicious entity to link information to specific users. For example, in the case of location privacy, the adversary should not be able to link a location to a user. In order to obfuscate personal data, PETs often rely on cryptographic primitives, such as anonymous authentication and encryption. For example, private information retrieval [63] is a cryptographic tool that allows a user to retrieve an item from a server's database without revealing which item it is retrieving. In mobile networks, PETs are often executed on mobile devices. Resource constraints of mobile devices might not allow for the use of costly cryptographic techniques. An alternate solution consists in using statistical obfuscation methods such as data obfuscation. For example, a user can hide the item it retrieved from a database by simply retrieving a large number of items. In contrast with the strong protection provided by cryptography, statistical methods provide best-effort protection.

In some systems, cryptographic techniques may be insufficient to provide privacy. In computer networks, because of the way networking protocols operate, a detailed analysis of the timing of messages (e.g., time of transmission and reception) can reveal significant information about communicating parties even if messages are encrypted. For this reason, anonymous communications techniques usually rely on statistical mixing of data traffic in addition to cryptographic techniques. The same problem arises in wireless networks where wireless communications can be observed by passive eavesdroppers. Hence, in this dissertation, we use a combination of cryptographic and statistical obfuscation methods to provide location privacy in wireless networks.

The Future of Privacy

The ongoing debate about the future of privacy sees two main trends: some believe that privacy is "dead", whereas others believe that privacy remains fundamental.

Some advocate *post-privacy optimism*: they foresee a transparent society where new institutions and practices provide benefits that compensate for the loss of privacy. [41, 114]. If society thrives on differences instead of conformism, openness could facilitate the expression of personality and the externalization of identity over new communication technologies.

In contrast, *post-privacy negativists* associate the end of privacy with the end of civilization [37, 125, 167]. They argue that the global exposure of all actions in a transparent society may prompt self-censorship in the form of the *chilling effect* [196]: some conduct would be suppressed by fear of penalization and this process would fuel the standardization of society. Similarly, the difficulty in guaranteeing symmetric access to surveillance technologies would lead to the exploitation of information by a few, thus making *social sorting* [151] a reality: surveillance technology could automate categorization of people in order to influence their life choices.

Approach

In this thesis, we believe in a society that protects individuals: privacy should be actively protected to regulate the advancement of technology and to maintain technology at the service of people. People should have access to tools to control their privacy and determine how to participate online. In other words, we consider that the goal of communication technology is to enable people to better communicate and share more with the right people. This task will certainly become increasingly complex as technologies surpass human capabilities [141] and become more intertwined with society.

As previously discussed, the design of privacy protocols should start from the context of the considered scenario. In communication networks, essential pieces of this context are performance and security properties of communication protocols. For example, in location-based services, performance depends on the number of communications and computations, while security depends on authentication and encryption mechanisms. In this dissertation, we focus on the overhead introduced by privacy-preserving mechanisms. We notably investigate the influence of cost on the performance of privacy-preserving mechanisms and the effect of *rational behavior* by cost-averse entities on the achievable privacy. The consideration of cost in our analysis may lead to the design of more efficient privacy-preserving protocols.

Information required by communication protocols should be designed to disclose the *minimal amount of information*. For example, network identifiers used in communication protocols can aid the traceability of users, thus jeopardizing location privacy. Similarly, in location-based services, multiple queries containing location information could be linked to a specific user if a network pseudonym (e.g., IP address) uniquely identifies the user.

Privacy-preserving techniques can be used in conjunction with *security* mechanisms to provide other services such as revocation and authentication. In this dissertation, we show that protecting privacy does not impede *security*. We develop privacy-preserving mechanisms that enable misbehavior detection and establishment of *trust* [185]. We rely on cryptographic primitives to guarantee security in privacy-preserving communication networks.

We consider privacy-preserving mechanisms that rely on the properties of mobile networks and on cryptographic primitives. In particular, the proposed mechanisms exploit the

randomness of users' mobility to inhibit tracking users' whereabouts. We consider distributed privacy-preserving protocols, as they are particularly suited for decentralized communication networks such as ad hoc networks. We rely on analytical tools, such as optimization theory, game theory and mean field theory to model interactions of multiple mobile devices protecting their location privacy. These tools enable us to predict the result of complex interactions and evaluate the effectiveness of privacy-preserving mechanisms.

Contributions

We consider the **problem of location privacy in existing and emerging networks**. We identify *location privacy threats* and evaluate *location privacy preserving mechanisms*. A clear understanding of location privacy threats is required to design efficient privacy-preserving mechanisms. Also, privacy-preserving mechanisms change location privacy threats. We consider several system models capturing existing location-based services, emerging wireless networks and social networks.

Our contributions are as follows:

1. We investigate location privacy threats in existing location-based services. Although location-based services are already widely deployed, there is a growing concern over their privacy implications. We evaluate the relation between the quantity and type of location information and the ability of the adversary to obtain users' real identity and interests. Our analysis quantifies privacy risks in sharing location information with third parties. We show that, in some cases, third parties can recover personal user information after a small number of location-based queries and that users with a regular routine are particularly prone to such threats. These results question the ability of privacy-preserving mechanisms to obfuscate highly correlated information such as users' whereabouts.

2. We evaluate state-of-the-art security and privacy architectures of peer-to-peer wireless networks. Because these architectures rely on authentication and network identifiers, devices' locations can be tracked. We investigate the use of mix zones to anonymize location traces while preserving functionalities of peer-to-peer wireless services. Mix zones, like most privacy-preserving algorithms, aim at *engineering an anonymity set* by mixing the actions of a set of wireless devices. We propose several protocols to deploy mix zones in wireless networks and, in particular, a distributed solution - PseudoGame - that takes rationality into account in order to optimize the trade-off between privacy and cost. We compare different protocols and show that strategic behavior can lead to efficient mixing strategies at a reasonable cost.

3. We study a communication primitive that allows users to rely on communities in order to share information in an ad hoc fashion and enhance social interaction. Communications induced by this primitive may reveal users' community membership to eavesdroppers. We formalize this problem using the concept of community privacy. We propose communication protocols to efficiently identify communities of users in a privacy-preserving fashion. These protocols are similar to secret handshakes with a focus on cost reduction and unlinkability. We evaluate the achievable privacy using a challenge-response formalization and analytically derive the relation between anonymity and unlinkability. Our results shed light on the achievable unlinkability of secret handshake schemes.

Outline

This thesis has three parts, all considering location privacy problems from a different angle:

Part I discusses location privacy threats in existing location-based services. It considers a system where mobile devices episodically share their location with third-parties via infrastructure-based communications. In **Chapter 1**, we investigate privacy risks involved in this type of sharing.

Part II studies location privacy in peer-to-peer wireless networks. We consider a system model where mobile devices broadcast at high frequency wireless messages that can be easily eavesdropped (e.g., WiFi in ad hoc mode or Bluetooth). We introduce in **Chapter 2** the network model and the privacy architecture based on mix zones used throughout Part II. We consider centralized approaches to deploy mix zones in **Chapter 3**. In **Chapter 4**, we consider distributed approaches to create mix zones in which privacy and cost are traded-off by rational agents. We compare with simulations the proposed privacy protocols. In **Chapter 5**, we push further the analysis of the distributed approach by providing a framework to *analytically* evaluate the privacy obtained with mix zones.

Part III explores the relation between peer-to-peer wireless networks and social networks. We consider a system model where a social network is on top of peer-to-peer wireless networks, creating so-called pervasive social networks. In **Chapter 6**, in addition to location privacy, we consider the data and community privacy issues of such communication networks.

Part I

Location Privacy in Modern Communication Networks

Chapter 1

Location Privacy Threats in Location-based Services

> Everyone has three lives: a public life, a private life, and a secret life.
>
> Gabriel García Márquez

1.1 Introduction

In traditional cellular networks, users share their location with their network operator in order to obtain voice and data services pervasively. With the emergence of data services, users increasingly share their location with other parties such as providers of location-based services (LBSs). Specifically, users first obtain their location by relying on the localization capability of their mobile device (e.g., GPS or wireless triangulation), share it with LBSs and then obtain customized services based on their location. Yet, unlike cellular operators, LBSs are mainly provided for free and generate revenue with location-based advertisement. Hence, there is a key difference between the business models of LBS providers and cellular operators: LBS providers aim at profiling their users in order to serve tailored advertisement.

Subscribers of cellular networks know that their personal data is contractually protected. On the contrary, users of LBSs often lack an understanding of the privacy implications caused by the use of LBSs [101]. Some users protect their privacy by hiding behind pseudonyms that are mostly invariant over time (e.g., usernames). Previous works identified privacy threats induced by the use of LBSs and proposed mechanisms to protect user privacy. Essentially, these mechanisms rely on trusted third-party servers that anonymize requests to LBSs. However, such privacy-preserving mechanisms are not widely available and users continue sharing their location information unprotected with third-parties. Similarly, previous work usually considers the worst-case scenario in which users continuously upload their location to third-parties (e.g., traffic monitoring systems). Yet, with most LBSs, users do not share their location continuously but instead, connect *episodically* to LBSs depending on their needs and thus reveal a few location samples of their entire trajectory. For example, a localized search on Google Maps [31] only reveals a location sample upon *manually* connecting to the service.

In this work, we consider a model that matches the common use of LBSs: we do not assume the presence of privacy-preserving mechanisms and consider that users access LBSs on

a regular basis (but not continuously). In this setting, we aim at understanding the privacy risk caused by LBSs. To do so, we experiment with real mobility traces and investigate the *dynamics* of user privacy in such systems by measuring the *erosion* of user privacy. In particular, we evaluate the success of LBSs in predicting the true identity of pseudonymous users and their points of interest based on different samples of mobility traces. Our results explore the relation between the *type* and *quantity* of data collected by LBSs and their ability to de-anonymize and profile users. We quantify the potential of these threats by carrying out an experimentation based on real mobility traces from two cities, one in Sweden and one in Switzerland. We show that LBS providers are able to uniquely identify users and accurately profile them based on a small number of location samples observed from the users. Users with a strong routine face higher privacy risk, especially if their routine does not coincide with that of others. To the best of our knowledge, this work is among the first to investigate the erosion of privacy caused by the sharing of location samples with LBSs using real mobility traces and to quantify the privacy risk.

1.2 Related Work

Location-based services [10, 31, 89, 150] offer users to connect with friends, to discover their environment or to optimize their mobility. In most services, users share their location episodically when connecting to the service. Some services such as road-traffic monitoring systems require users to continuously share their location. In this work, we consider users that manually share their location and thus only reveal samples of their mobility.

The IETF Geopriv working group [22] aims at delivering specifications that will help implementing location-aware protocols in a privacy-conscious fashion. It proposes to use independent location servers that deliver data to LBSs according to privacy policies defined by users. In other words, it provides user control over the sharing of their location data. In this Chapter, we complement the IETF proposal by enabling to quantify the privacy threat induced by the sharing of specific location data with LBSs.

Privacy-preserving mechanisms impede LBSs from tracking and identifying their users [28, 107, 117, 118, 120, 157, 228]. In general, the proposed mechanisms either run on third-party anonymizing servers, or directly on mobile devices. Most mechanisms alter the user identifier or the content of location samples. For example, *anonymization* techniques remove the identifier from the user requests, and *obfuscation* techniques blur the location information. The effectiveness of privacy-preserving mechanisms is usually evaluated by measuring the level of privacy [77, 199, 202, 210]. However, privacy-preserving mechanisms are rarely used in practice. One reason may be that users do not perceive the privacy threat because they are not intensively sharing their location. In this work, we aim at clarifying the privacy threat when users reveal samples of their mobility manually and do not make use of privacy-preserving mechanisms.

Without privacy-preserving mechanisms, pseudonymous location information enables the identification of mobile users [27, 30, 138, 159]. Beresford and Stajano [27] identified all users in continuous location traces by examining where users spent most of their time. Using GPS traces from vehicles, two studies by Hoh *et al.* [119] and Krumm [138] found the home addresses of most drivers. De Mulder *et al.* [159] could identify mobile users in a GSM cellular network from pre-existing location profiles by using statistical identification processes. These works rely on continuous location traces precisely capturing users' whereabouts in order to

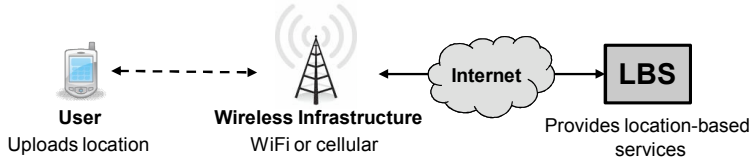


Figure 1.1: System model. Users episodically upload their location wirelessly to LBSs.

identify users. Yet, in most scenarios, users share location samples episodically and thus the privacy threat in practice remains unclear. Partridge and Golle [104] could identify most of the US working population using two location samples, i.e., approximate home and work locations. It remains unclear however whether users of location-based services face this threat. Hence, in this work, we consider real life scenarios of users sharing location information with LBSs and investigate the privacy risk. In [153], Ma *et al.* study the erosion of privacy caused by published anonymous mobility traces and show that an adversary can rapidly relate location samples to published anonymous traces. In this work, we push further the analysis by considering an adversary without access to anonymized traces.

1.3 Preliminaries

We present the assumptions regarding LBSs and the associated privacy threats.

1.3.1 Network Model

We study a network (Fig. 1.1) that involves mobile users equipped with wireless devices, third-parties running LBSs and a wireless infrastructure. Wireless devices feature localization technology such as GPS or wireless triangulation that lets users locate themselves. The geographic *location* of a user is denoted by $l = (lat, lon)$, where lat is the latitude and lon is the longitude. The wireless infrastructure relies on technology such as WiFi, GSM or 3G to let users connect to the Internet. LBSs are operated by independent third-parties that provide services based on the location of mobile users.

Cellphone users send their location together with a service request to LBSs through the wireless infrastructure. For each request sent, users may identify themselves to the LBS using proper credentials. In general, we assume that users are identified with *pseudonyms* (i.e., fictitious identifiers), such as their username, their HTTP cookie or their IP address: some services may require users to register and provide the corresponding username and password, whereas others may use HTTP cookies to recognize users.

LBSs provide users with services using the location information from requests. LBSs store the information collected about their users in a database. As defined in [203], each location sample is called an *event* denoted by $< i, t, l >$, where i is the pseudonym of a user, t is the time instance at which the event occurred, and l is the location of the user. The collection of events from a user forms a *mobility trace*.

1.3.2 Threat Model

LBS operators passively collect information about the locations of pseudonymous users over time. For example, an LBS can observe location samples of user A over the course of weeks

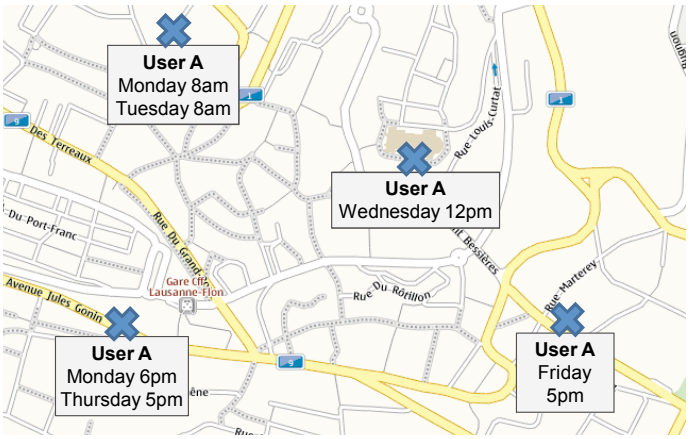


Figure 1.2: Threat model. The LBS learns multiple locations of user A over several days. It can then infer the activities of user A and possibly obtain his true identity.

(Fig. 1.2). A priori, the LBS also knows the map over which users move and has access to geographic information systems that provide details such as points of interest in maps [64]. In addition, the LBS may use the increasing amount of location information available from other sources such as public transportation systems [190].

We consider that the LBS aims at obtaining the true identify of its users and their points of interest. To do so, the LBS studies the collected information. Even if communications are pseudonymous, the spatio-temporal correlation of location traces may serve as a *quasi-identifier* [28, 30, 70, 104]. This is a significant threat to user privacy as such information can help the LBS to profile users.In this work, we investigate the ability of LBSs to identify users and their points of interest based on the collected information.

1.4 Privacy Erosion

In this section, we describe the process of *privacy erosion* caused by the use of LBSs. To do so, we first discuss how users tend to share their location with LBSs. Then, we explain how LBS operators can obtain the true identity of their users and their points of interest from the collected information.

1.4.1 Collection of Traces by LBSs

Depending on the provided service, LBSs collect location samples about their users. For example, typical services (such as Foursquare, Google Maps, Gowalla, or Twitter) offer users to connect with friends, to discover their environment or to optimize their mobility. Depending on their needs, users access such LBSs at different times and from different locations, thus revealing multiple location samples. In general, users of these services do not continuously share their locations. Instead, users have to manually decide to share their location with their mobile devices. We classify most popular LBSs into three broad categories and describe the information shared by users in each case.

Localized Search

Many LBSs enable users to search for local services around a specific location (e.g., localized Google Search [31]). Localized searches offer mobile subscribers spontaneous access to nearby services. Hence, a user location acts as a spatial query to the LBS. For example, users can obtain the location of nearby businesses, products, events, restaurants, movie theaters or other local information depending on the type of information provided by the LBS.

Such localized searches help users navigate unfamiliar regions and discover unknown places. Thus, users episodically connect to LBSs, revealing samples of their mobility. LBSs obtain statistical information about the visited locations of mobile users and learn popular locations and user habits. Yet, LBSs do not learn the actual activity of mobile users (e.g., the name of the visited restaurant) as they do not know the decision of the user about the provided information.

Street Directions

Another popular use of LBSs consists in finding a route between two locations. Typically, a user shares its location with an LBS and requests the shortest route to another location (e.g., Google Maps).

Users of such services usually reveal their current location and a potential destination. Hence, users may leak their home/work locations to LBSs in addition to samples of their mobility. This enables LBSs to obtain statistical information about the preferred origins and destinations of mobile users.

Location Check-ins

A novel type of location-based service offers users to check-in to specific *places* in return for information related to the visited location [89]. For example, it can be used to check into shops, restaurants or museums. It allows users to meet other users that share similar interests and to discover new aspects of their city through recommendations [150].

With such services, users not only precisely reveal their location (GPS coordinates), but also their intention. Indeed, the LBS can learn the current activity of its users. Users can check-in to public places, but also private homes.

In summary, depending on the provided service, users share different *samples* of their mobility. In order to take this into account, in the following, we consider that LBSs obtain various *type* of location samples out of users' whereabouts.

1.4.2 Attacks by LBSs

The spatial and temporal information contained in mobility traces may serve as location-based quasi-identifiers [30, 70]: an LBS may obtain the true identity and points of interests of its pseudonymous users from the collected mobility traces.

Location-Based Quasi-Identifiers

Quasi-identifiers were introduced by Delenius [70] in the context of databases. They characterize a set of attributes that in combination can be linked to identify to whom the database

refers (see [161] for more). In [30], Bettini *et al.* extend the concept to mobile networking. They consider the possibility to identify users based on spatio-temporal data and propose the concept of location-based quasi-identifiers. They describe how a sequence of spatio-temporal constraints may specify a mobility pattern that serve as a unique identifier. For example, as already mentioned, Golle and Partridge [104] identify location-based quasi-identifiers by showing that home and work locations uniquely identify most of the US population. Hence, if an LBS learns users' home and work locations, it can obtain their identity with high probability.

In this work, we assume that the LBS succumbs to the temptation of finding the identify of its users. To do so, we consider that the LBS uses the home and work locations of users as location-based quasi-identifiers.

Inferring Home and Work Locations

Previous works investigated the problem of characterizing and extracting important places from pseudonymous location data. These works propose various algorithms to infer important locations based on the spatial and temporal evidence of the location data. We group the existing works in two categories. In the first category [16, 119, 138], the authors use *clustering* algorithms to infer the *homes* of mobile users. For example in [138], Krumm proposes four different clustering techniques to identify the homes of mobile users in vehicular traces: traces are pseudonymous and contain time-stamped latitudes and longitudes. Similarly, in [119], Hoh *et al.* propose a k-mean clustering algorithm to identify the homes of mobile users in anonymous vehicular traces: traces do not have pseudonyms, but contain the speed of vehicles, in addition to their location. In the second category [147, 148], Liao *et al.* propose *machine learning* algorithms to infer the different type of *activities* from mobile users data (e.g., home, work, shops, restaurants). Based on pedestrian GPS traces, the authors are able to identify (among other things) the *home and work locations* of mobile users.

We rely on previous work to derive an algorithm that exclusively infers the *home* and *work* locations of mobile users based on spatial and temporal constraints of pseudonymous location traces. The algorithm operates in two steps: first, it clusters spatially the events to identify frequently visited regions; second, it temporally clusters the events to identify home and work locations.

The spatial clustering of the events uses a variant of the k-means algorithm as defined in [16]: it starts from one random location and a radius. All events within the radius of the location are marked as potential members of the cluster. The mean of these points is computed and is taken as the new centre point. The process is repeated until the mean stops changing. Then, all the points within the radius are placed in the cluster and removed from consideration. The procedure repeats until no events remain. The number of points falling into a cluster corresponds to its weight and is stored along with the cluster location. Clusters with a large weight represent frequently visited locations.

Based on the output of the spatial filtering, the algorithm then uses temporal evidence as a criterion to further refine the possible home/work locations. In practice, users have different temporal patterns depending on their activities (e.g., students). The algorithm considers simple heuristics that apply to the vast majority of users. For example, most users spend the night at home and commute in the beginning/end of the day. In order to apply the temporal evidence, the algorithm considers all events in each cluster and labels them as home or work. Some events may remain unlabeled if they do not match any temporal criterion. The algorithm considers two temporal criteria. First, the algorithm checks the duration of

stay at each location. To do so, it computes the time difference between the arrival of a trip at a certain location, and the departure time of the following trip. A user that stays more than 1 hour in a certain location over night is likely to have spent the night at home. Hence, the algorithm labels events occurring at such location as home events. Second, the algorithm labels events occurring after 9am and before 5pm as possible work events. Finally, for each cluster, the algorithm checks the number of events labelled home or work and deduces the most probable home and work locations.

Inferring User Points of Interest

Usually, LBSs use the content of queries to infer the points of interest of their users. Yet, LBSs may further profile users by analyzing the location of multiple queries and inferring users' points of interest.

We use the spatial clustering algorithm defined above to obtain the possible points of interest of users that we call *uPOIs*: a uPOI is a location regularly visited by a user. For each identified uPOI, we store the number of visits of the user and derive the probability P_v^i that a user i visits a specific uPOI, i.e., the number of visits to a uPOI normalized by the total number of visits to uPOIs.

Metrics

The real home/work addresses are unavailable in our data sets. Hence, we apply our algorithm to the original mobility traces and derive a baseline of home/work locations. We then evaluate the probability of success of the LBS by comparing the baseline to the outcome of our algorithm on the *samples* of location data collected by the LBS. In other words, we compare the home/work location pairs predicted from the sampled traces with the baseline. In practice, it is complicated to obtain the real home and work locations of users (i.e., the ground truth) in mobility traces without threatening their privacy. Because no real ground truth is available, this approach does not guarantee that we have identified the real home/work locations. Yet, it allows us to compare the effectiveness of the attack in various conditions.

The probability P_s of a successful identification by the LBS is then:

$$P_s = \frac{\text{Number of home/work pairs correctly guessed}}{\text{Total number of home/work pairs}} \tag{1.1}$$

This probability measures the ability of LBSs to find the home/work locations from sampled traces and thus uniquely identify users. This metric relies on the assumption that home/work location pairs uniquely identify users [104] (in Europe as well): it provides an upper-bound on the identification threat as home/work location pairs may in practice be insufficient to identify users especially in the presence of uncertainty about the home/work locations.

We also evaluate the normalized *anonymity set* of the home and work pairs of mobile users. To do so, we compute the number of home/work locations that are in a certain radius from the home/work location of a certain user i. For every user i, we define its home location as h_i and its work location as w_i. For each user i, we have:

$$A^i_{home} = \frac{1}{|h|}\sum_{j\neq i} 1_{|h_j - h_i| < R_A} \tag{1.2}$$

$$A^i_{work} = \frac{1}{|w|}\sum_{j\neq i} 1_{|w_j - w_i| < R_A} \tag{1.3}$$

where R_A specifies the radius considered for the anonymity set.

We measure the ability of LBSs to infer uPOIs by considering for each user i, the number of uPOIs correctly inferred. For every user i, we have:

$$P^i_{uPOI} = \frac{\text{Number of uPOIs correctly guessed}}{\text{Number of uPOIs}} \tag{1.4}$$

We also use the notion of Kullback-Leibler divergence [140] to measure the ability of the adversary to guess the probability of each user visiting specific uPOIs. For every user i, we have:

$$D_{KL}(P^i_v||Q^i_v) = \sum_j P^i_v(j) \log \frac{P^i_v(j)}{Q^i_v(j)} \tag{1.5}$$

where P^i_v is the actual probability that user i visits specific uPOIs and Q^i_v is the probability guessed by the adversary.

1.5 Evaluation

We present our methodology to evaluate the erosion of privacy caused by LBSs.

1.5.1 Setup

We start from data sets of real mobility traces. The data sets contain the location of users at a high granularity. Because users usually reveal only a few location samples to LBS operators, we artificially reduce the information available to the LBSs by selecting a few events from the traces. Then, we consider various de-anonymization attacks on the location traces. In practice, we load mobility traces in Matlab and apply the algorithm described in Section 1.4.2. We repeat every analysis 100 times and consider the average.

1.5.2 Mobility Traces

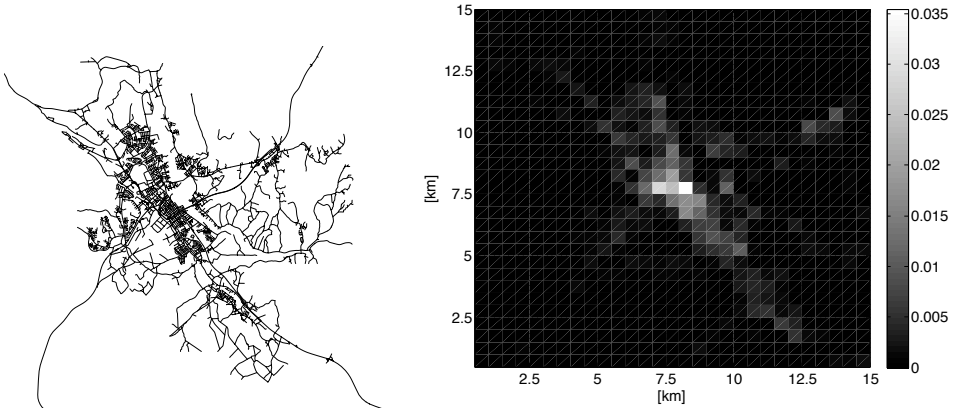
There exist several publicly available data sets of human mobility. For example, there are mobility traces of taxis [180], of student mobility in campus [83], or of sport activities [179]. Yet, most of these data sets have a limited applicability to our problem because the mobility of users is tied to specific scenarios (e.g, taxis, campus).

In this work, we consider two data sets representing normal activities of users in cities. These mobility traces contain several *trips* for each user. A trip defines a trajectory of a user going from one source location to a destination (e.g., a user commuting from home to work). Users move on a map following road constraints.

Borlange Data Set

The city of Borlange is a middle-sized ($15 \times 15\text{km}^2$) Swedish city of approximately 46000 inhabitants. Borlange has 3077 road intersections interconnected by 7459 roads (Fig. 1.3 (a)). The data set was collected over two years (1999-2001) as part of an experiment on traffic congestion that took place there.[1] About 200 private *cars* (with one driver per car) within a 25 km radius around the city center were equipped with a GPS device. At regular

[1]The data set is available at http://icapeople.epfl.ch/freudiger/borlange.zip .



(a) (b)

Figure 1.3: Borlange data set. (a) Map of Borlange, Sweden. The city has 46000 inhabitants and spreads over $15 \times 15\text{km}^2$. (b) Spatial histogram showing the density of users per cell $c(z)$.

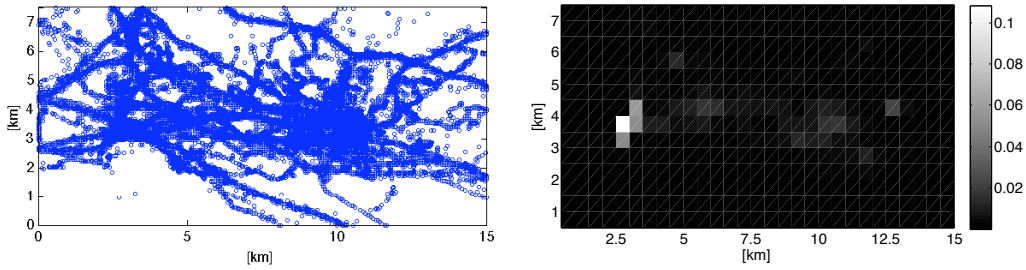
intervals (approximately every 5 seconds), the position, time and speed of each vehicle was recorded and stored. Mostly because of GPS accuracy issues, many observed trips did not match the Borlange map. The data was thus manually verified and corrected using road fitting algorithms for a subset of 24 vehicles resulting in a total of 420814 "clean" trips (see [92] for more details). This data set was obtained by civil engineers and used to analyze the route choices of mobile users.

Lausanne Data Set

The Lausanne area in Switzerland is a region of $15 \times 7\text{km}^2$ of approximately 120000 inhabitants (Fig. 1.4 (a)). In September 2009, Nokia began running a data collection campaign in Lausanne area. Around 150 *users* are equipped with GPS-enabled Nokia phones that record their daily activities and upload them on a central database. Among other things, the phones measure the GPS locations of users at regular intervals (approximately every 10 seconds). In July 2010, we took a snapshot of the database containing traces of 143 users tracked over 12 months.[2] Note that the database contains traces of pedestrians, but also of users in cars, buses and trains. It has thus a larger diversity in terms of mobility patterns than the Borlange data set. We focus on the traces that start and finish in the Lausanne area and obtain around 106600 trips.

In order to evaluate the statistical relevance of the mobility traces, we compute statistics of mobility in the data sets. We divide the whole region of Borlange/Lausanne into square cells of equal size (500×500m). and evaluate the distribution of users' visits in each cell. We define a variable C_z that counts the number of events among all users that happen in each cell z. For each cell, we compute the empirical probability that an event falls into the cell z, $c(z) = \frac{C_z}{\sum_x C_x}$. In Figure 1.3 (b), we show the density map (i.e., the set of cells with their corresponding $c(z)$) for the Borlange data set. We observe that the activity of users is concentrated in a few regions. We observe a similar distribution in the Lausanne data set (Fig. 1.4 (b)). Yet, in the latter, there is a small bias towards one location (the EPFL

[2]The data set is not publicly available.



(a) (b)

Figure 1.4: Lausanne data set. (a) Map of Lausanne area, Switzerland. The city has 120000 inhabitants and spreads over $15 \times 7\text{km}^2$. (b) Spatial histogram showing the density of users per cell $c(z)$.

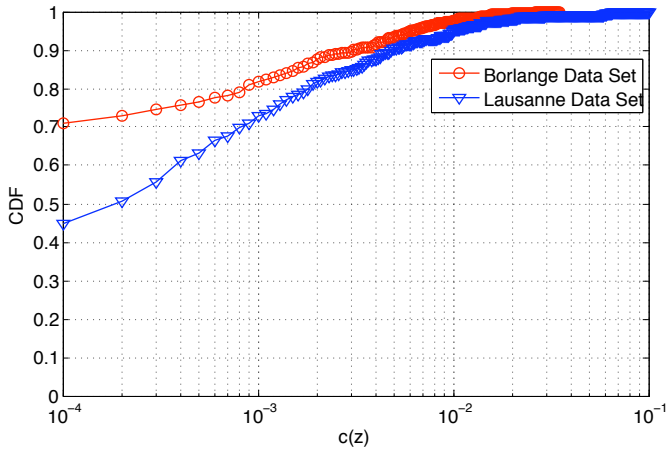


Figure 1.5: Empirical CDF of $c(z)$ in semi-log scale. We observe a linear behavior for both data sets indicating a heavy-tailed distribution of user density in the network.

campus), indicating that many users from the experiment share the same work place. The Lausanne data set reflects scenarios in which many users share the same work place, for example, downtown of a large city.

In Figure 1.5, we show the empirical Cumulative Distribution Function (CDF) of $c(z)$ for both data sets in semi-log scale. We observe that the CDF increases linearly, indicating a heavy-tailed distribution of user density. This confirms that some cells have a density much above the average. Our observations about the heavy-tailed distribution match existing results in the literature on mobility traces [57, 179, 218] and confirm the statistical relevance of the data sets. Intuitively, the heavy-tailed distribution may indicate that users are easily identifiable as they share few locations.

1.5.3 Modeling the Collection of Traces by LBSs

We start from mobility traces containing location samples at high granularity. As described in Section 1.4.1, the *type* and *quantity* of location information collected by LBSs depends on

the services and their usage. To take this into account, we select a few events from the entire traces in various ways. Each selected event effectively represents a *query* to LBSs.

Uniform Selection (UF)

We select events uniformly at random from the set of all possible events of each user. This captures scenarios in which users are likely to use an LBS anytime and anywhere.

Home/Work Selection (HW)

We distinguish between three types of events: *home*, *work* and *miscellaneous*. Home and work events refer to queries made from home and work, respectively, whereas miscellaneous events refer to other visited locations. Based on these type of events, we select location samples uniformly in each set corresponding to *home/work* events with probability ρ and miscellaneous events with probability $1-\rho$. A large ρ captures scenarios in which users access LBSs mostly from home and work (e.g., street directions), whereas a small ρ captures scenarios in which users access LBSs mostly on the go (e.g., localized search or location check-ins).

Points of Interest Selection (PO)

We distinguish between two types of events: *cPOIs* and *miscellaneous*. cPOI events refer to queries made from regions of a city with many points of interest (e.g., POIs of the city such as restaurants and shops), whereas miscellaneous events refer to other visited locations. Based on these type of events, we select location samples uniformly in each set corresponding to cPOI events with probability ρ. A large ρ captures scenarios in which users access LBSs mostly from popular locations (i.e., localized search), whereas a small ρ captures scenarios in which users access LBSs mostly in unpopular areas such as residential areas.

Preferred Selection (PF)

We distinguish between two types of events: *preferred* and *miscellaneous*. Preferred events refer to queries made from locations frequently visited by each user (i.e., uPOIs such as gyms and friends' places), whereas miscellaneous events refer to other visited locations. Based on these type of events, we select location samples corresponding to preferred events with probability ρ. A large ρ captures scenarios in which users access LBSs mostly during their routine, whereas a small ρ captures scenarios in which users access LBSs mostly in unfamiliar areas.

We tune the selection type using probability ρ. Note that home/work selection strategy with $\rho = 0.5$ is different from the uniform selection strategy: with $\rho = 0.5$ in home/work selection, home/work events and miscellaneous events have the same probability to be chosen, whereas with the uniform selection, all events have the same probability to be chosen. We consider various number of queries λ in order to model the quantity of data collected by LBSs. For example, a number of queries $\lambda = 60$ means that 60 samples of all location samples of each user are shared with the LBS.

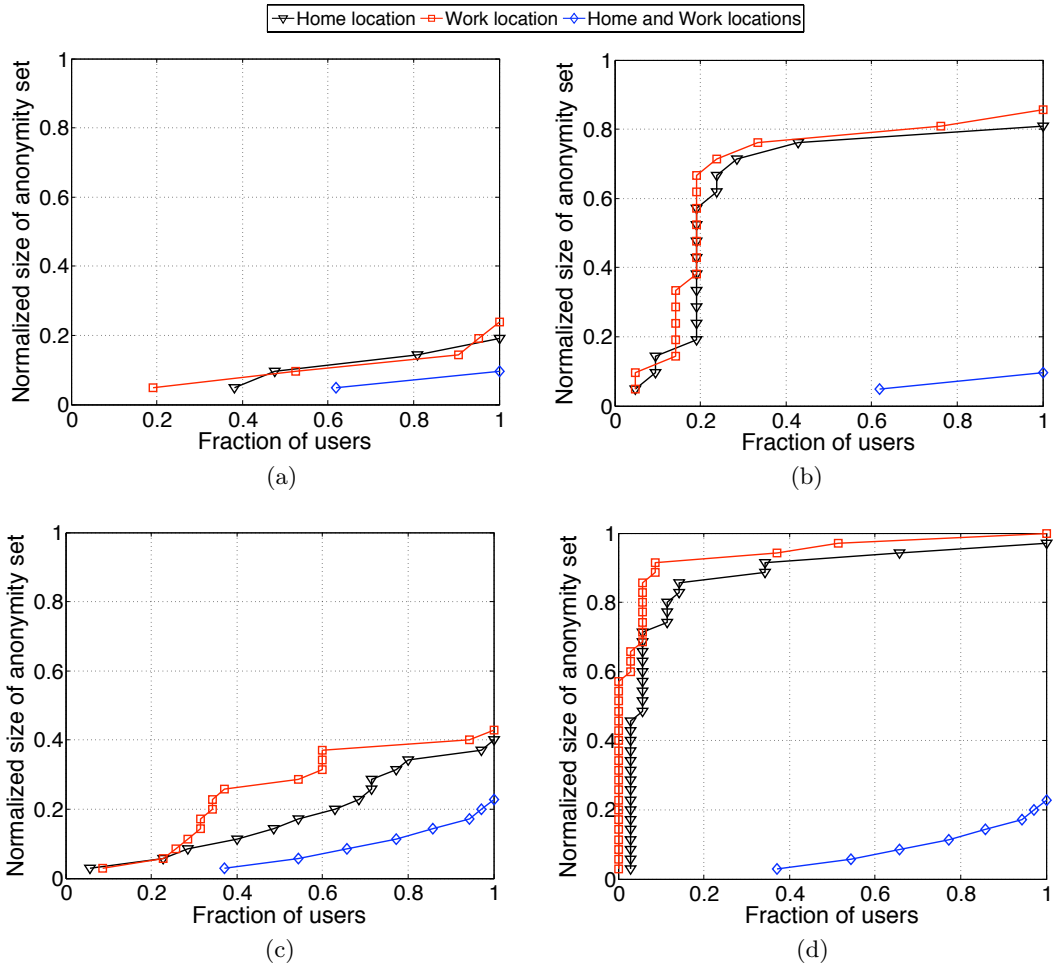


Figure 1.6: Normalized size of anonymity set A^i_{home}, A^i_{work} and $A^i_{homeWork}$. Borlange with (a) $R_A = 1$km and (b) $R_A = 5$km. Lausanne with (c) $R_A = 1$km and (d) $R_A = 5$km.

1.5.4 Results

Unless otherwise stated, we consider that users share their location with the LBS with a 10 meters precision (i.e., GPS), that the clustering radius in the spatial clustering algorithm is 100 meters and that the adversary has a tolerable error margin of 50 meters to correctly guess a home/work/uPOI locations.

Size of Anonymity Set

The graphs in Fig. 1.6 detail the size of the anonymity set for home locations, work locations, or both normalized with the number of users in the data set. On the x-axis, the graphs show the fraction of users that has an anonymity set of less than a given normalized size on the y-axis. We consider two radius $R_A = 1$km and $R_A = 5$km. As predicted in [104], the anonymity set size is low especially when a small radius is used and revealing the home and work locations is much more identifying than only revealing one of them. In general, we observe that more users share a common work place than home. In the Lausanne data

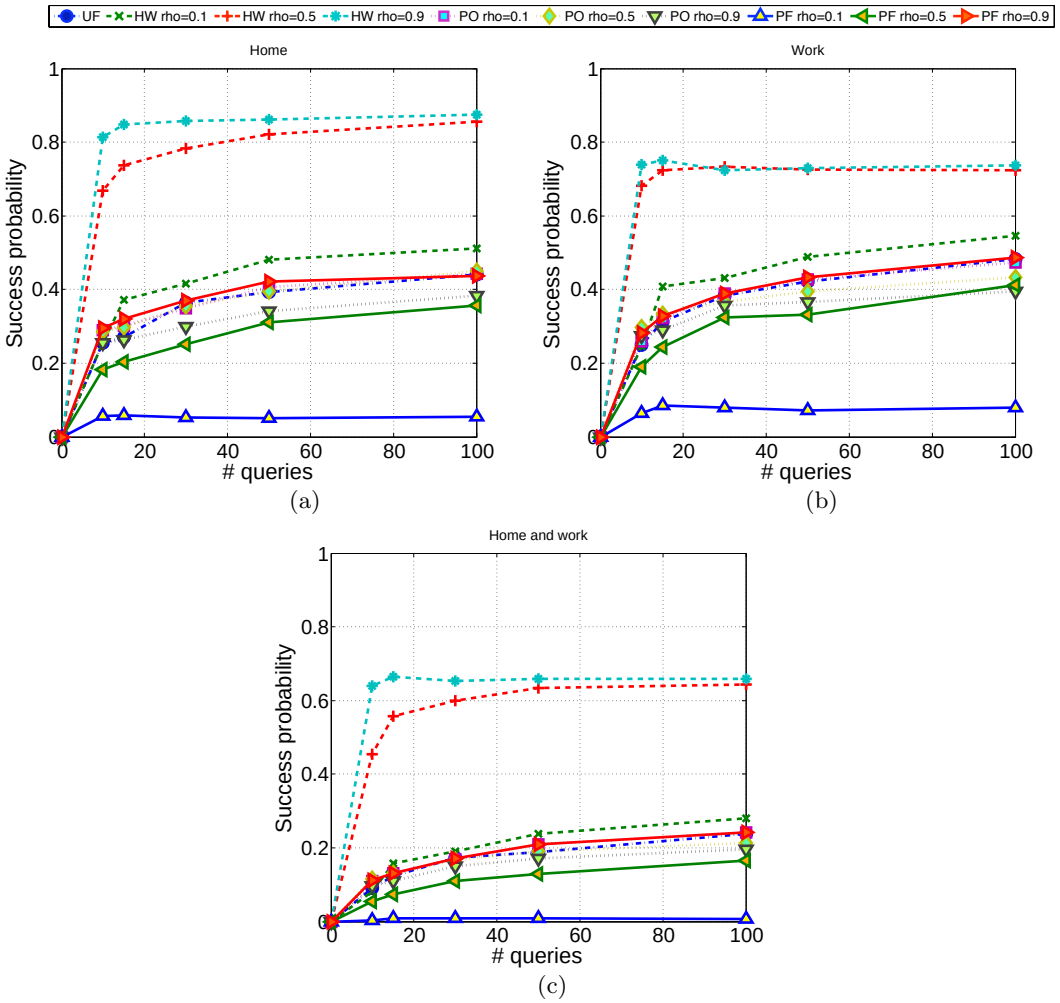


Figure 1.7: Privacy erosion in Borlange with varying selection probability ρ and number of queries λ. (a) Home identification. (b) Work identification. (c) Home and work identification (P_s).

set, many users have a larger anonymity set than in the Borlange data set due to the larger number of users.

Privacy Erosion

We evaluate the privacy erosion of users from the Borlange and Lausanne data sets in multiple scenarios. We measure the probability that an LBS successfully identifies the home location, the work location, or both. In the case of a successful home and work identification, the LBS successfully identifies its users. We consider different data collection scenarios as described earlier (UF, HW, PO and PF) with three selection probabilities: $\rho = 0.1$, $\rho = 0.5$ and $\rho = 0.9$. We also vary λ, the amount of information shared with LBSs.

In Fig. 1.7 and Fig. 1.8, we show the erosion of privacy in Borlange and Lausanne for various ρ, λ and selection strategies. We observe that with HW selection, the probability of

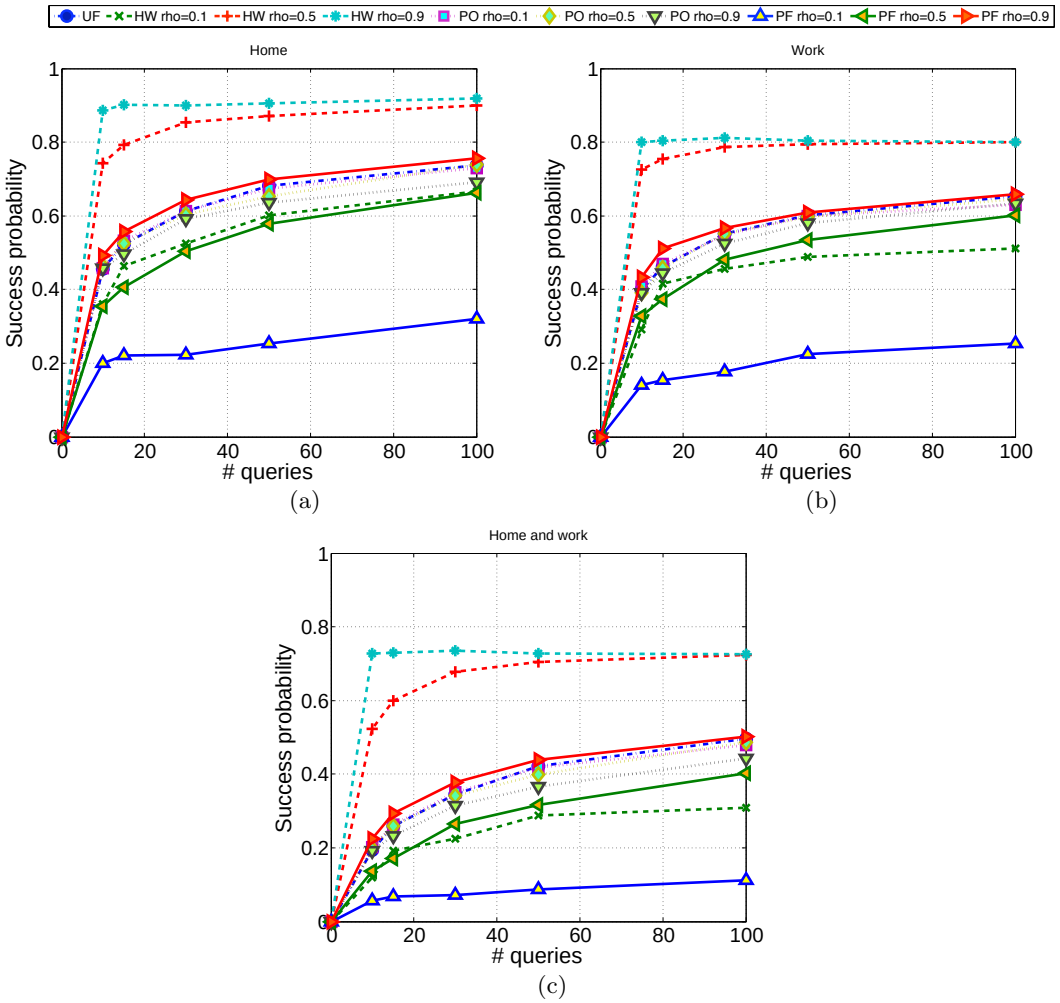
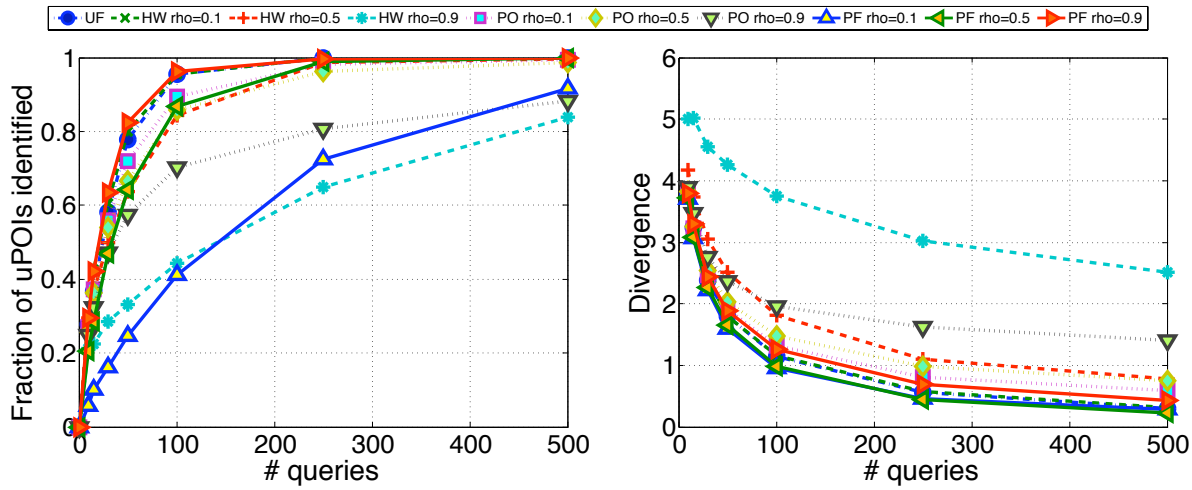


Figure 1.8: Privacy erosion in Lausanne with varying selection probability ρ and number of queries λ. (a) Home identification. (b) Work identification. (c) Home and work identification (P_s).

Data Set	**uPOIs**									
	1	2	3	4	5	6	7	8	9	10
Borlange	0.357	0.209	0.112	0.078	0.052	0.031	0.021	0.017	0.015	0.012
Lausanne	0.401	0.14	0.092	0.063	0.045	0.035	0.028	0.023	0.019	0.016

Table 1.1: Average probability $E[P_v^i]$ of visiting a uPOI.

identification of a home, work, or home/work pair increases the fastest with respect to the number of sent queries indicating that LBSs uniquely identify users with few locations: in Borlange, if $\rho = 0.9$, 20 queries are sufficient to identify 65% of users. We observe that as ρ increases, so does the identification success. In Lausanne, the identification success is slightly higher but still leads to the same conclusions. We observe that PF selection with $\rho = 0.1$ makes de-anonymization particularly difficult. In this case, users share their location only in unfamiliar areas and it is thus difficult for LBSs to infer users' identity. For other selection



(a) (b)

Figure 1.9: Inferring the top ten uPOIs in Borlange data set. (a) Average fraction of uPOIs identified $E[P^i_{uPOI}]$. (b) Average divergence $E[D_{KL}(P^i_v||Q^i_v)]$.

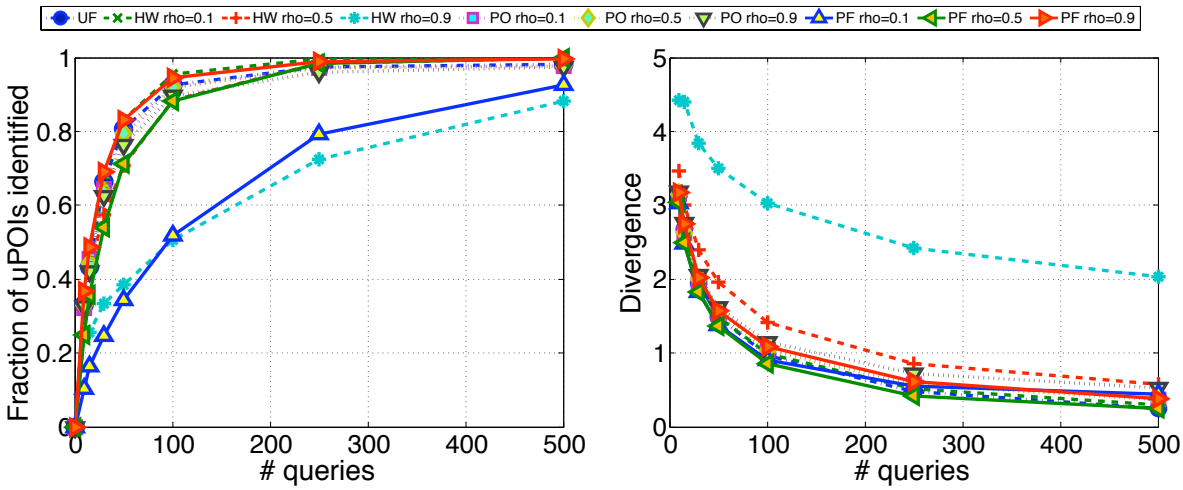
strategies, the identification success saturates around 20 to 40% and increases slowly with the number of queries.

Inferred User Points of Interest

Table 1.1 shows the average fraction of visits to uPOIs. Each uPOI identifies a region of 200 meters radius frequently visited by each user. In both data sets, the distribution is long tail showing that few uPOIs are frequently visited.

In Figure 1.9 and 1.10, we show the ability of LBSs to infer the top ten uPOIs of each user: we compute the average fraction of uPOIs identified $E[P^i_{uPOI}]$ within a 100 meters error margin and evaluate the average divergence $E[D_{KL}(P^i_v||Q^i_v]$. We observe that the adversary can infer a large number of uPOIs with a small number of samples: with 30 samples, it can learn up to 65% of uPOIs in the case of PF $\rho = 0.9$. The best selection strategies are PF $\rho = 0.9$, HW $\rho = 0.1$ and UF. Intuitively, revealing preferred visited locations reveals clusters, similarly, uniform across visited locations will have high probability to sample from frequently visited location. On the contrary, with PO $\rho = 0.9$, HW $\rho = 0.9$ or PF $\rho = 0.1$, the attack works less efficiently. Hence, even with a few location samples, the adversary is also able to infer most uPOIs.

In terms of divergence, a divergence of zero indicates a perfect match. We observe that the divergence decreases fast indicating that the adversary obtains a probability distribution similar to the true one and identifies the most probable uPOIs. We observe a similar behavior with the Lausanne data set. Note that the ability to infer uPOIs is at odds with the ability to infer users' identity: with HW $\rho = 0.9$, it is harder to identify uPOIs and easier to identify users.



(a) (b)

Figure 1.10: Inferring the top ten uPOIs in Lausanne data set. (a) Average fraction of uPOI identified $E[P^i_{uPOI}]$. (b) Average divergence $E[D_{KL}(P^i_v||Q^i_v)]$.

1.6 Conclusion

We have considered the problem of privacy erosion when using location-based services. We identify the quantity and type of location information that statistically helps LBSs find users' real identity and points of interest. In contrast with previous work (mostly showing that de-anonymization based on location information is possible), we push the understanding of the threat further by showing how de-anonymization depends on the collected data. We experiment with two real data sets of mobility traces, model the collection of traces by LBSs and implement various attacks. Our results show that in many scenarios a small amount of information shared with LBSs may enable to uniquely identify users. These results stem from the fact that the spatio-temporal correlation of location traces tends to be unique to individuals and persistent. We also show that in some scenarios, users have high privacy without using privacy-preserving mechanisms.

The results of this work can help prevent the false sense of anonymity that users of LBSs might have by increasing the awareness of location privacy threats. In particular, it may encourage users to stop revealing sensitive information to third-parties, such as their home and work locations, and adopt privacy-preserving mechanisms. These results notably question the ability of privacy-preserving mechanisms to obfuscate highly correlated information such as users' whereabouts. They can also help design more efficient privacy-preserving mechanisms [94, 95, 99] and may encourage the use of distributed solutions in which users store maps and the related information directly on their mobile devices.

Publication: [100]

Part II

Location Privacy in Peer-to-Peer Wireless Networks

Chapter 2

Mix Zones for Location Privacy

Tout esprit profond avance masqué.
René Descartes

2.1 Introduction

In virtually all deployed wireless networks, mobile devices communicate through a wired infrastructure, typically through a cellular base station or a WiFi access point. However, the growing popularity of Bluetooth, WiFi in ad hoc mode [11] and other similar techniques is likely to fuel the adoption of peer-to-peer wireless communications. Corporations are developing wireless peer-to-peer technologies such as Nokia Instant Community [32] and Qualcomm FlashLinQ [144]. In addition to classic infrastructure-based communications, mobile devices can communicate directly with each other in an ad hoc wireless fashion. Such communications dramatically increase mobile devices' *awareness* of their environment, enabling a new breed of context-aware applications. For example, cell phones could exchange messages with other nearby cell phones to provide a geographic extension of online social network [3, 4, 10, 86]; likewise, vehicles could communicate with each other, thereby increasing road safety and optimizing traffic [113, 222].

The integration of peer-to-peer wireless communications into mobile devices brings new security challenges, due to their mobile and ad hoc nature. Unlike wired communications, wireless communications are inherently dependent on geographic proximity: mobile devices detect each other's presence by periodically broadcasting beacon messages. These messages include *pseudonyms* such as MAC/IP addresses and public keys in order to identify communicating parties, route communications and secure communications. Security of these messages is essential in order to authenticate mobile nodes [45], revoke misbehaving nodes [184] and ensure communication integrity.

Much to the detriment of privacy, external parties can monitor pseudonyms in broadcasted messages in order to track the locations of mobile devices, thus compromising *location privacy*. Even if pseudonyms are not directly related to the real identity of mobile device owners, previous work [27, 119, 138] shows that, by using the *spatial* and *temporal* correlation between successive pseudonymous locations of mobile devices, an external party can infer the real identity of mobile devices' owners.

Hence, we must avoid revealing privacy-sensitive information, such as network identifiers and authentication credentials. There are multiple solutions to anonymously authenticate mobile devices. One possible solution is the *multiple pseudonym* approach [58] suggested in the context of Internet communications: it assigns to every device a list of public keys that are used alternatively to achieve *pseudonymous authentication*. Network identifiers must also be anonymized in order to avoid traceability at the networking layer. In line with the multiple pseudonym approach, long-term identifiers should be replaced by short-term identifiers. For example, MAC addresses could serve solely for short-term communications [108]. Both industry [7, 8, 158] and academia [28, 43, 122, 145, 195] have adopted the multiple pseudonym approach at the network layer in order to achieve location privacy. In particular, GSM-based cellular networks are the most prominent example of privacy protection based on short-term identifiers [19, 116].

A pseudonym changed by an isolated device in a wireless network can be trivially guessed by an external party observing beacon messages. Hence, a change of pseudonym should be spatially and temporally coordinated among mobile devices [28]. More specifically, a device cannot free-ride on the pseudonym change of others to achieve location privacy as its pseudonym can still be tracked. Hence, in peer-to-peer wireless networks, location privacy requires a collective effort by neighboring mobile devices.

The coordination of pseudonym changes has become a central topic of research with various approaches proposed. One solution [43] consists in changing pseudonyms periodically, at a pre-determined frequency. The mechanism works if at least two mobile devices change their pseudonyms in proximity, a condition that is rarely met (as the probability of a synchronous change is low). Base stations can be used as coordinators to synchronize pseudonym changes [122, 75], but this solution requires help from the infrastructure. The approach in [108] enables mobile devices to change their pseudonyms at specific time instances (e.g., before associating with wireless access points). However, this solution achieves location privacy only with respect to the infrastructure of access points. Another approach [28, 97, 99] coordinates pseudonym changes by forcing mobile devices to change their pseudonyms within pre-determined regions called *mix zones*. The effectiveness of a mix zone, in terms of the location privacy it provides, depends on the adversary's ability to relate mobile devices that enter and exit the mix zone [27]. Hence, mix zones should be placed in locations with high device density and unpredictable mobility [28, 122]. This approach however may lack flexibility because the locations of mix zones are fixed by a central authority and must be learned by mobile devices prior to entering the network. Several researchers advocated the use of a distributed solution [121, 122, 145], where mobile devices coordinate pseudonym changes to dynamically obtain mix zones. To do this, a mobile device broadcasts a pseudonym change request to its neighbors.

The multiple pseudonym approach has drawbacks that affect performance of current solutions. First, a pseudonym change causes considerable overhead, thus reducing networking performance: for example, routing algorithms must update their routing tables [198]. Second, given the cost of asymmetric key generation and management by the central authority, mobile devices are usually assigned a limited number of pseudonyms that can quickly become a scarce resource if changed frequently. Pseudonyms used in pseudonymous authentication may thus be costly to acquire and use. Third, mix zones have a cost because they impose limits on the services available to mobile users: in order to protect against spatial correlation of location traces, mix zones can conceal the trajectory of mobile devices by not allowing devices in the mix zone to communicate [121]. Hence, the number of mix zones traversed by mobile

devices must be kept small. Finally, even if the distributed solution synchronizes pseudonym changes, it does not align incentives between mobile devices: because the achieved location privacy depends on both the device density and the unpredictability of device movements in mix zones [28], a selfish mobile device might decide not to change its pseudonym in settings offering low location privacy guarantees.

In this Chapter, we introduce a privacy architecture for peer-to-peer wireless networks based on the mixing of network identifiers. We describe multiple techniques to create mix zones, show the criteria for a successful pseudonym change and introduce measures of mixing effectiveness. Then, in the following Chapters, we tackle one of the main issues that has hindered the use of multiple pseudonym schemes in mobile networks. We propose and evaluate novel approaches to deploy mix zones. First, we consider a trusted central authority that uses a *centralized algorithm* for deploying mix zones in a given area and investigate several deployment strategies. Second, we propose a distributed solution for deploying mix zones that takes rationality into account. *Rational* mobile devices locally decide whether to change their pseudonyms based on their degree of privacy and cost. Third, we provide a framework for *analytically* evaluating privacy obtained with mix zones. We focus on the time period over which a pseudonym is used, i.e., the age of a pseudonym, and provide *critical conditions* for the emergence of location privacy in mobile networks.

2.2 Related Work

Research on location privacy has recently gained tremendous momentum with the apparition of location-based services. Given privacy implications of such services, previous work investigated privacy threats and proposed several privacy-preserving mechanisms.

Location Privacy Threats

Previous work [27, 119, 138] shows that the adversary can implicitly obtain the true identity of the owner of a mobile device from the analysis of its visited locations. For example, using pseudonymous location traces collected in an office environment, Beresford and Stajano [27] correctly identify all participants by simply examining where the participants spent most of their time. Similarly, using pseudonymous GPS traces from vehicles, two studies by Hoh *et al.* [119] and Krumm [138] find the identities of most drivers. Hence, pseudonyms are not sufficient to protect the location privacy of mobile nodes and should be changed to avoid such attacks. But even if location traces of mobile nodes do not contain any pseudonyms (i.e., they are anonymous), Hoh and Gruteser [117] reconstruct the tracks of mobile nodes using a multiple target tracking (MTT) algorithm. Hence, location traces should also be altered *spatially* to reduce the granularity of location traces. In other words, spatial and temporal correlations between successive locations of mobile nodes must be carefully eliminated to prevent external parties from compromising their location privacy. In this work, location privacy is achieved by changing pseudonyms in regions called *mix zones* [27], thus eliminating temporal and spatial correlations of pseudonymous location traces.

Previous work also shows that it is possible to identify devices relying on their distinctive characteristics (i.e., fingerprints) at the physical, link and application layer. At the physical layer, the wireless transceiver has a wireless fingerprint that can identify wireless devices in the long term by using modulation-based [40], transient-based [71] or amplitude-based techniques [214] or a combination of features [110, 181]. However, these techniques are only

evaluated with specific technologies and countermeasures could be developed. Recent results by Danev *et al.* [72] show that wireless fingerprints can be *impersonated*, i.e., a device can copy the wireless fingerprint of another device. Ability to impersonate a wireless fingerprint is directly related to the ability of an adversary to recognize a fingerprint. In other words, if it is easy for an adversary to fingerprint wireless interfaces, it is also easy for a mobile device to change its wireless fingerprint. At the link layer, it is possible to distinguish between a number of devices and drivers [90, 170]. At the application layer, devices can be identified based on clock skews [136]. However, such techniques require an active adversary and can be countered by ignoring requests sent by the adversary. Similarly, a reduction of the differences between drivers would limit the effectiveness of such attacks. In this thesis, we do not consider fingerprinting threats. Note that, independently from the presence of fingerprinting attacks, higher-layer privacy mechanisms such as mix zones remain useful. Some applications may, for example, require keeping location traces (e.g., for congestion analysis).

Privacy-Preserving Mechanisms

The multiple pseudonym approach considered in this dissertation is a popular mechanism to preserve location privacy: it is used in cellular networks and Bluetooth [85] to provide location privacy. For example, in cellular networks, cell phones periodically register their location to the cellular network operator in order to enable the routing of calls. A unique identifier sent in clear, the International Mobile Subscriber Identity (IMSI) identifies users in these registration messages. In order to avoid traceability, the IMSI is most of the time replaced by a temporary identifier, the Temporary Mobile Subscriber Identity (TMSI), that changes over time. The TMSI is changed when a user moves from one group of adjacent cells to another (called location areas). Cellular networks thus make use of *network-issued pseudonyms* to protect location privacy. In contrast, in Bluetooth, devices can independently change their MAC address over time: the MAC address is composed of a random number concatenated with the encryption of the same random number. The random number is encrypted with a shared secret previously derived from a pairing operation, thus allowing paired devices to recognize each other. In this thesis, we push the multiple pseudonym approach further by investigating its use in distributed wireless networks.

There are other techniques, besides the multiple pseudonyms approach, for achieving location privacy [139, 19]. In location-based services, mobile nodes can intentionally add noise to their location [107] by for example reporting their location as a region instead of a point [210]. Mobile devices can also send dummy messages containing dummy locations [96]. However, in mobile wireless networks, even if messages do not contain location information, the peer-to-peer communications between mobile nodes implicitly reveal their locations. Hence, obfuscating the location data contained in messages is insufficient to protect the location privacy of mobile nodes. Another possibility [106] is to encrypt communications and include the MAC address in the encrypted part of packets. This way, the link layer protocol is identifier-free and packets are completely anonymized. This works well in the case of WiFi access points [106] but would not scale to peer-to-peer wireless communications because of the ad hoc nature of communications.

2.3 System Model

We study a network where mobile devices/nodes are autonomous entities equipped with WiFi-enabled devices that communicate upon coming in range. In other words, we describe a pervasive communication system (a mobile ad hoc network) such as a vehicular network [113], a delay tolerant network [86], or a network of directly communicating hand-held devices [3, 4, 10] where mobile nodes in proximity automatically exchange information.

As commonly assumed in such networks [171], we consider an offline Certification Authority (CA) run by an independent trusted third party that helps manage, among other things, security and privacy of the network. In line with the multiple pseudonym approach, we assume that, prior to joining the network, every mobile node i registers with the CA that preloads a finite set of pseudonyms and required cryptographic material for anonymous authentication (e.g., a set of asymmetric keys for pseudonymous authentication or a set of anonymous credentials). Upon changing pseudonyms, we consider for simplicity that the old pseudonym expires and is removed from the node's memory. Once a mobile node has used all its pseudonyms, it contacts the CA to obtain a new set or generates a new set by itself.

We consider a discrete time system with initial time $t = 0$. At each time step t, mobile nodes move in the network. We assume that mobile nodes automatically exchange information (unbeknownst to their users) as soon as they are in communication range. Note that our evaluation is independent from the communication protocol. Still, we align our communication model with common assumptions of pervasive communication systems: mobile devices advertise their presence by periodically broadcasting proximity beacons containing their identity and the current time. We consider a beaconing mode similar to that in Ad Hoc mode of IEEE 802.11. Due to the broadcast nature of wireless communications, beacons enable mobile devices to discover their neighbors and can be used for time synchronization. When node i receives a beacon from node j, node i can start interacting with node j by using data messages. Any node can control the validity of data messages by requesting the certificate of the public key from the sender and verifying the signatures on data messages. Subsequently, if confidentiality is required, a security association can be established (e.g., with Diffie-Hellman). Note that there is ongoing work in the literature [46, 53] to reduce the cryptographic overhead induced by the processing of all messages.

2.4 Threat Model

We assume that an adversary $\mathcal{A}$ aims to track the whereabouts of mobile nodes using inference attacks. We consider that $\mathcal{A}$ can have the same credentials as mobile nodes (e.g., insider attack) and can eavesdrop communications. In practice, $\mathcal{A}$ can be a rogue individual, a set of malicious mobile nodes, or may even deploy its own infrastructure by placing eavesdropping devices in the network. In the worst case, $\mathcal{A}$ obtains complete coverage and tracks nodes throughout the entire network. We characterize the latter type of adversary as *global*.

The nature of wireless communications makes eavesdropping particularly easy. $\mathcal{A}$ can collect identifying information (i.e., pseudonyms) from the entire network and obtain *location traces*. Finally, we assume that the key-pair generation and distribution process cannot be altered or controlled by the adversary.

In addition to eavesdropping abilities, the knowledge of the adversary depends on other information it has, e.g., background information about users' mobility and points of interest.

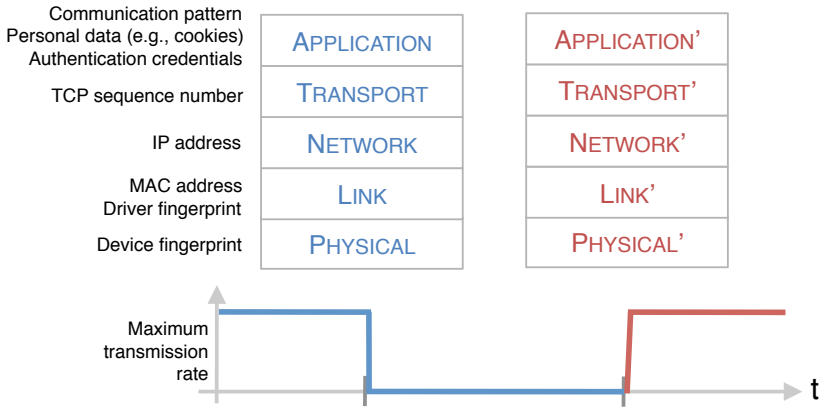


Figure 2.1: Illustration of a pseudonym change. All identifiers and quasi-identifiers of the networking stack must change. In this example, the maximum transmission rate is momentarily set to zero to stop all wireless communications of the device during the pseudonym change.

2.5 Privacy Architecture

Peer-to-peer wireless networks have to meet security requirements such as confidentiality, integrity, availability, authentication and real-time delivery in order to enable successful deployment of higher-layer services. They also need to protect the privacy of mobile users. In this section, we discuss multiple primitives to protect privacy.

2.5.1 Atomicity of Pseudonym Change

In wireless networks, a pseudonym change is successful if and only if a mobile device, before a pseudonym change, cannot be identified after a pseudonym change. A pseudonym change must be *atomic*: even if a single identifier remains unchanged, the entire pseudonym change fails (Fig. 2.1).

Mere removal of explicit identifiers, such as network identifiers, may be insufficient because individuals can be tracked by linking so-called *quasi-identifiers* [70, 192, 193], i.e., distinctive attributes that facilitate the indirect re-identification of individuals.

Identifiers

Network protocols use explicit identifiers to route packets from sources to destinations (e.g., MAC and IP addresses) and to link multiple user interactions (HTTP cookies or usernames). Similarly, authentication protocols may use explicit identifiers to link the identity of a user to a certificate.

A pseudonym change must guarantee that these explicit identifiers change. In practice, it may be difficult to alter network-assigned identifiers such as HTTP cookies or usernames. Thus, such information should be transmitted over an encrypted channel to thwart eavesdropping.

Quasi-Identifiers

In addition to explicit identifiers, implicit identification based on quasi-identifiers may be used to track mobile devices. Quasi-identifiers characterize a set of attributes of a mobile device that in combination may be linked to identity users. For example, the communication pattern, the TCP sequence number, or device fingerprints may be unique to a mobile device (Fig. 2.1). Even if a pseudonym change occurs, an adversary could track a mobile device based on its communication pattern.

To overcome the threat of quasi-identifiers, a pseudonym change must also guarantee that all fingerprints of a device disappear or change to make a device indistinguishable from others.

2.5.2 Authentication

The promised ad hoc sharing of information might turn into a pervasive nightmare if undesired communications cannot be filtered out: for example, if mobile nodes cannot verify the source of information, they are susceptible to mobile spam. To thwart rogue devices from polluting the network, nodes should authenticate each other: the existence of an authentication feature (and the implied procedure to obtain the appropriate credentials) makes it more difficult for attackers to join the network in the first place and thus increases the cost of misbehavior. Hence, by verifying the authenticity of their interlocutor before exchanging information, mobile nodes reduce the amount of undesired data. Authentication also enables the creation of security associations in order to encrypt communication and protect data privacy.

Misbehavior by insiders is still possible: for example, an authenticated mobile node can engage in spamming attacks. However, because mobile devices are authenticated, the CA can then exclude misbehaving nodes by revoking their keying material. To do so, keys can be blacklisted using traditional revocation algorithms [226].

If authentication is done without appropriate precautions, it would break the atomicity property of pseudonym changes, thus rendering the privacy problem particularly challenging [194]. Multiple techniques can be used to *anonymously authenticate* mobile users.

Pseudonymous Authentication

In pseudonymous authentication, a set of asymmetric keys is preloaded into mobile devices by an off-line certification authority [198] or directly generated by mobile nodes [46, 142, 155, 225]. At each pseudonym change, the material used for authentication is changed. Typically, the validity period of a public key certificate is short and overlaps in time with another certificate to provide flexibility with the timing of pseudonym changes.

This approach enables fast authentication of users and effective revocation. It is thus favored for real-time systems, such as vehicular networks. In case of preloaded pseudonyms, pseudonymous authentication may require intensive key management from the certification authority to maintain availability of the system in place. Recent results for directly generated pseudonyms demonstrate their feasibility for real-time systems such as vehicular networks [46].

Group Signatures

Another type of pseudonymous authentication consists in using a group identifier instead of a user identifier: Group signatures [60] allow a group member to sign on behalf of a group without revealing the identity of the signer. Highly efficient group signatures schemes

exist with constant size signatures and efficient signing and verification even when nodes are revoked [18, 35, 51]. The size of the group determines the achievable privacy of its member.

Group signatures require a group manager to add and revoke group members thus making the flexibility of groups dependent on the availability and computational capacity of the group manager. One important drawback is that if several groups coexist in a mobile network, mobile users may be traceable based on their group membership (see the work on secret handshakes described in Chapter 6).

Like group signatures, ring signatures [188] allow nodes to sign on behalf of an ad hoc group of nodes. Unlike group signatures, they do this without the help of a central coordinator. The location privacy provided by ring signatures is investigated in [98]. The main drawbacks of ring signatures are their significant communication overhead and the impossibility to trace the origin of the signer even in case of misbehavior. Hence, group signatures seem more appropriate for many peer-to-peer wireless applications.

Anonymous Credentials

Usually based on group signatures, anonymous credentials [47, 49, 50, 59, 152] allow mobile nodes to anonymously prove ownership of credentials to third parties, with the help of an online credential issuer. They can thus be used for anonymous authentication as well. However, these schemes require the online availability of a credential issuer, which is often not possible in wireless networks, and do not prevent misbehavior: a rogue device can undetectably provide misleading data in bulk. To circumvent the issue, techniques based on unclonable identifiers, such as e-tokens [48], allow nodes to anonymously authenticate themselves a given number of times per period. This limits the amount of false information a rogue device can provide per time period. Nevertheless, anonymous credentials impose a high overhead due to costly credential verification (i.e., zero knowledge proofs) and costly revocation procedure.

In this thesis, we consider that any of these mechanisms can be used to provide anonymous authentication.

2.5.3 Darwinian Privacy

As privacy threats evolve, strategies to counter the threats must also be adaptive [191]: it is important to craft efficient security strategies based on the potential of *threats* and the *cost* of privacy mechanisms. The presence of a passive adversary is uncertain in mobile networks as it requires much effort to detect eavesdropping attacks. This makes it difficult to identify the potential of the threat and properly allocate resources to privacy-preserving mechanisms. For these reasons, in this thesis, we consider a worst-case adversary (i.e., global) and model the cost as an abstract parameter of pseudonym change protocols. Then, we predict the effect of cost on the privacy-preserving strategy.

The multiple pseudonym approach incurs various types of costs on mobile devices and network operators. First, mix zones impose limits on services available to mobile users: in order to protect against spatial correlation of location traces, mix zones can conceal the trajectory of mobile devices by not allowing devices in the mix zone to communicate [121]. Hence, the number of mix zones traversed by mobile devices must be kept small. Second, anonymous authentication is costly. Any anonymous authentication technique can be used to protect location. Consider the pseudonymous authentication approach: given the cost

of asymmetric key generation and management by the central authority, mobile devices are usually assigned a limited number of pseudonyms that can quickly become a scarce resource if changed frequently. Pseudonyms may be costly to acquire and use. Note that it is still unclear whether nodes could reuse their previous pseudonyms. It could facilitate tracking by an adversary and make it difficult to prevent Sybil attacks [80]. It could also reduce costs and provide the ability to recognize the same device without identifying it. Third, a pseudonym change causes considerable overhead, thus reducing networking performance. For example, routing algorithms must update their routing tables [198]. Similarly, if a mobile device changes its pseudonyms very frequently, it makes it more difficult for another device in vicinity to initiate a communication session. Finally, the multiple pseudonym approach does not align incentives between mobile devices. Because the achieved location privacy depends on both device density and unpredictability of device movements in mix zones [28], a selfish device might decide not to change its pseudonym in settings that offer low location privacy guarantees.

Consequently, the multiple pseudonym approach has to be carefully crafted into the design of the network in order to provide efficient privacy protection at tolerable cost. In Chapters 3 and 4, we propose multiple designs and evaluate their trade-off between privacy and cost. The analysis of such trade-off is essential to understand how privacy-preserving mechanisms affect communications.

Note that the design of privacy-preserving mechanisms could also take into account the cost of attacks [124]. Passive attacks typically require a phase of data collection and a phase of data analysis. The first phase involves the deployment of sniffing stations in order to collect wireless messages in the area under attack. The data analysis phase requires processing power and background information to mine collected data. We consider a worst-case adversary that can afford to implement global passive attacks.

2.6 Privacy Analysis

We describe how to create mix zones and their effectiveness in achieving location privacy.

2.6.1 Mix Zone Definition

In order to protect against the temporal and spatial correlation of location traces, mobile devices can change their pseudonyms in mix zones. Mix zones can conceal the trajectory of mobile nodes to the external adversary by using: (i) Silent/encrypted periods [44, 97, 121, 145], (ii) a mobile proxy [195], or (iii) regions where the adversary has no coverage [43]. Note that, if the beaconing mechanisms is probabilistic (e.g., using a backoff window), trajectories are concealed without requiring special techniques because devices will not beacon at regular intervals.

Consider a group of $n(t)$ mobile nodes at time t that are in proximity. They conceal their trajectories using one of the methods described above. At the end of the silent period, it appears to an external adversary as if all pseudonym changes occurred simultaneously.

A mix zone $i \in \mathcal{Z}$ is defined by a pair of coordinates (x_i, y_i), where $\mathcal{Z}$ is the set of all mix zones in the considered area. The x_i and y_i coordinates are the center of the mix zone i and determine the location of the mix zone in the network.

If a centralized algorithm is used to deploy mix zones, the central authority can predetermine the size and shape of mix zones, forcing all nodes to conceal their trajectory for the

same time period. In this thesis, we consider for simplicity that every mix zone i will have the shape of a circle with a certain radius R_i. We assume constant mix zones sizes, $R_i = R$, for all centralized algorithms. Recent work considered other shapes for mix zones and show that it can improve the achievable privacy [168]. If a distributed algorithm is used to deploy mix zones, each mix zone will have a different shape and size depending on how long nodes conceal their trajectory and on their trajectory.

Finally, there is a mixing attempt if and only if at least two nodes change pseudonyms in a mix zone.

2.6.2 Mix Zone Effectiveness

Mix zones are *effective* in anonymizing the trajectory of mobile nodes if the adversary is unable to predict (with certainty) the relation between mobile nodes entering and exiting mix zones. In particular, a mix zone becomes a *confusion point* for the adversary if the mixing attempt achieves high location privacy.

Adversary $\mathcal{A}$ observes a set of $n(T)$ nodes changing pseudonyms, where T is the time at which the pseudonym change occurs. Assuming that $\mathcal{A}$ knows the *mobility profile* of the nodes within each mix zone, the adversary can attempt to map entering to exiting events.

In the following, we discuss various metrics to measure location privacy achieved with mix zones. In general, we show that mix zones should be placed in locations with high node density and unpredictable mobility [28, 122].

As presented by Beresford and Stajano [27] for mobile networks and by Diaz *et al.* [77] and Serjantov and Danezis [199] for mix networks, the uncertainty of the adversary (i.e. entropy) is a measure of the location privacy/anonymity achieved by a mixing attempt.

Event-based Metric

The goal of the event-based metric is to measure untraceability of all users in a mix zone. It measures the probability that the adversary finds the assignment of all entering events to all exiting events.

Consider a sequence of entering/exiting nodes traversing a mix zone i over a period of T time steps, the uncertainty of the adversary is:

$$H_T(i) = -\sum_{v}^{I} p_v \log_2(p_v) \tag{2.1}$$

where p_v is the probability of different assignments of entering nodes to exiting nodes and I is the total number of such hypothesized assignments. Each value p_v depends on the entering/exiting nodes and the mobility profile. In other words, the anonymity provided by mix zones mostly depends on factors beyond control of the nodes. It is thus interesting to compute the average location privacy provided by a mix zone to evaluate its *mixing effectiveness*. The entropy measure is bound to the set of events happening in an interval of T time steps and does not capture the average mixing of a mix zone. The average mixing effectiveness of a mix zone i can be computed by taking the average entropy over n successive periods of T time steps: $E[H(i)] = \frac{1}{n}\sum_{v=1}^{n} H_{T_v}(i)$.

User-centric Metric

The goal of the user-centric metric is to measure the untraceability of a particular user traversing a mix zone, instead of the mix zone in general. Let us define the set B of pseudonyms before the change with the set R of pseudonyms after the change. Let $p_{r|b} = Pr(\text{“Pseudonym } r \in R \text{ corresponds to } b \in B\text{”})$, which is the probability that a new pseudonym $r \in R$ corresponds to an old pseudonym $b \in B$. The untraceability of node i using pseudonym b is defined as:

$$A_i(T) = -\sum_{r=1}^{n(T)} p_{r|b} \log_2(p_{r|b}) \tag{2.2}$$

The achievable location privacy depends on both the number of nodes $n(T)$ and mobility of nodes $p_{r|b}$ in the mix zone. If node i changes its pseudonym alone, then the adversary can track it, and we get $A_i(T) = 0$. Entropy is maximal for a uniform probability distribution $p_{r|b}$ and the achievable location privacy after a coordinated pseudonym change at time T is upper-bounded by $\log_2(n(T))$. If at least two mobile nodes (including i) change their pseudonyms, then the pseudonym change is successful and generates a confusion point. We denote T_i^{ℓ} as the time of the *last* successful pseudonym change of node i.

2.7 Application Scenario: Vehicular Networks

Vehicular Networks (VNs) consist of vehicles and Road-Side Units (RSUs) equipped with radios and a collection of backbone servers accessible via the RSUs. Using Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communications, vehicles share safety-related information and access location-based services. Initiatives in Europe [66] and the US [81] are evaluating VNs promises of safer driving conditions and efficient traffic management. Envisioned safety-related applications require vehicles to periodically broadcast their current position, speed and acceleration in authenticated *safety messages* (Fig. 2.2).

Due to their relevance to life-critical applications, VNs have to satisfy several strict security requirements, namely *sender and data authenticity*, *availability*, *liability*, and *real-time delivery*. Similarly to the work in [172, 173, 183, 186], we assume that a public key infrastructure is available and that the messages are properly signed to ensure liability. A certificate is attached to each message to enable other vehicles to verify the sender's authenticity. Vehicles are equipped with *Tamper-Proof Devices* (TPDs) that guarantee correct execution of cryptographic operations and non-disclosure of private keying material.

Although not designed with that purpose in mind, VNs facilitate tracking of vehicles. In fact, the cost of tracking vehicles by radio eavesdroppers is reduced compared to that of tracking vehicles with cameras. Similarly, tracking granularity is higher because an eavesdropper obtains identifiers, location and other information from safety messages. Moreover, unlike mobile phones and laptop wireless adapters, vehicle transceivers cannot be switched off [174]. Consequently, vehicles' whereabouts can be monitored at all times. All an adversary needs to do is deploy its devices across the area of the network that it wishes to monitor. In this application scenario, we are concerned with achieving location privacy against such an adversary.

To thwart tracking by an adversary, vehicles can use mix zones. Prior to entering the network, each vehicle i has to register with a Certification Authority (CA) and preloads a large set of *pseudonyms* $P_{i,k}$, with $k = 1, ..., \mathcal{F}$, where $\mathcal{F}$ is the size of the pseudonym set.

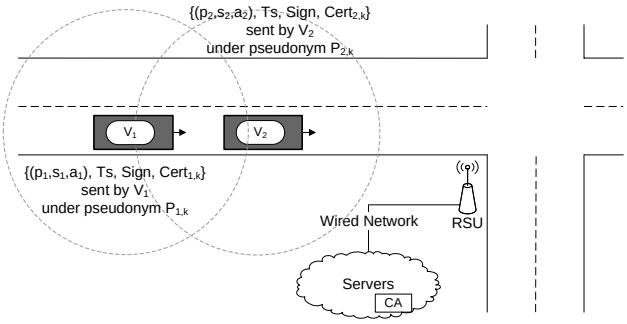


Figure 2.2: Example of vehicular network. Safety messages are emitted periodically (typically, every 100 ms to 300 ms) [81]. The Certification Authority (CA) is accessible through the RSUs. p_i, s_i and a_i are the vehicle i position, speed, and acceleration.

CAs are fully *trusted* and interoperable entities, operated by governmental organizations; they conform to privacy policies and keep the relation of the pseudonyms to the driver's real identity secret. In case of liability issues, this relation can be learned by law enforcement. For each pseudonym $P_{i,k}$ the corresponding CA generates a unique public/private key pair $(K_{i,k}, K_{i,k}^{-1})$ and a corresponding certificate $Cert_{i,k}(K_{i,k})$. Each vehicle sequentially updates its pseudonym at regular time intervals independently of other vehicles. Pseudonyms have a short validity period and cannot be reused.

As previously described, mix zones prevent the adversary from accessing the content of (safety) messages, including vehicle's signatures that are trivially linkable to the corresponding pseudonym. Silent mix zones or mobile proxies are difficult to implement in vehicular networks because they challenge the effectiveness of safety messages in preventing collisions. In particular, with silent mix zones, vehicles could not detect each others' presence in many situations and would fail at avoiding collisions. Recent work suggested to use them when vehicles move slowly [44]. We propose a protocol to create cryptographic mix zones. This solution thwarts computationally-bounded eavesdroppers while preserving the functionality of safety messages and offering more flexibility that silent mix zones.

2.7.1 The CMIX Protocol

We introduce the *CMIX Protocol* for creating Cryptographic mix zones (CMIXes): all legitimate vehicles within a mix zone obtain a symmetric key from the road-side unit (RSU) of the mix zone, and utilize this key to encrypt all their messages while within the zone. The symmetric key is obtained through a key establishment phase. To ensure functionality of safety messages, this mix zone key can be obtained by nodes approaching the mix zone with the help of a key forwarding mechanism, and finally, the RSU can switch to a new key through a key update mechanism.

CMIX Key Establishment

Vehicles rely on the presence of RSUs at mix zone locations (e.g., at road intersections) to initiate a *Key Establishment* mechanism and obtain a symmetric key. RSUs advertise their presence by periodically broadcasting beacons. As soon as v_i enters the transmission range of an RSU, R_{Beacon}, it initiates the key establishment protocol described in Table 2.1. As the

$v_i \rightarrow$ RSU: $\text{Request}, \text{Ts}_i, Sign_i(\text{Request}, \text{Ts}_i), Cert_{i,k}$
RSU $\rightarrow v_i$: $E_{K_{i,k}}(v_i, SK, \text{Ts}_{RSU}, Sign_{RSU}(v_i, SK, \text{Ts}_{RSU})), Cert_{RSU}$
$v_i \rightarrow$ RSU: $\text{Ack}, \text{Ts}_i, Sign_i(\text{Ack}, \text{Ts}_i), Cert_{i,k}$

Table 2.1: Key Establishment protocol. Ts is a time stamp, $Sign(\dot{})$ is the signature of the message, $Cert$ is the certificate of the message sender.

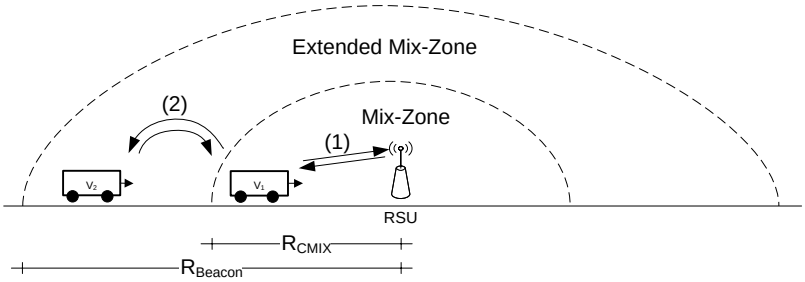


Figure 2.3: Extended mix zones. (1) v_1 uses the Key Establishment to learn the symmetric key. (2) v_2 uses the Key Forwarding protocol.

vehicle knows its own location and the location of the RSU (announced in the beacon), it can estimate whether it is within the mix zone, defined by a transmission range $R_{CMIX} < R_{Beacon}$. If so, v_i broadcasts one or (if needed) several key request messages (first message in Table 2.1). The RSU replies with the symmetric key SK encrypted with the public key of vehicle v_i and a signature. Vehicle v_i receives and decrypts the message. If the message is validated, v_i acknowledges it and uses SK to encrypt all subsequent safety messages until it leaves the mix zone. In case RSUs are co-located (i.e., their mix zones overlap), vehicles are aware of all CMIX keys so that they can decrypt all messages. Note that GSM uses the same technique. Alternatively, co-located RSUs could use the same CMIX key.

CMIX Key Forwarding

Vehicles in the *extended mix zone*, that is, at a distance d from the RSU where $R_{CMIX} < d < R_{Beacon}$ are unable to obtain directly the key from the RSU as they are beyond their transceiver's range for bidirectional communication and thus cannot decrypt safety messages coming out of the CMIX. As vehicles know they are within an RSU transmission range, when they receive encrypted safety messages, they issue one or, if needed, several key requests to obtain the SK key from vehicles already in the mix zone.

Consider the example of Figure 2.3: vehicle v_1 already knows the CMIX key and can forward it to v_2. When v_2 enters the extended mix zone, as soon as it receives an encrypted (intelligible) message, it initiates the broadcast of one or, if needed, several key requests. v_1 eventually receives a key request from v_2, and forwards it the symmetric key:

$$E_{K_{2,k}}(v_2, v_1, SK, \text{Ts}_{RSU}, Sign_{RSU}(v_1, SK, \text{Ts}_{RSU}))$$

The signature from the RSU, along with the time stamp, allows a receiver to validate the transmitted symmetric key. Note that vehicles in the extended region do not encrypt their safety messages with the CMIX key before entering the mix zone (R_{CMIX}).

CMIX Key Update

We propose a *Key Update* mechanism to renew or revoke CMIX symmetric keys. The RSU is responsible for such key updates and determines when to initiate the process. Key updates occur only when the mix zone is empty and vehicles obtain new keys via the key transport and key forward protocols. The CA obtains the new symmetric key from the RSU over a secure channel, to satisfy the liability requirements (i.e., possibly, decrypt safety messages in the future). The robustness provided by the system is increased, if key updates are asynchronous across different base stations.

2.7.2 Analysis of the CMIX Protocol

CMIX requires exchange of messages in the key establishment phase: a key request is sent until either an RSU or a vehicle receives it and sends back a key reply. To avoid reply flooding, vehicles sending key requests must acknowledge the acquisition of the key. Cryptographic overhead can also be easily sustained with relatively low delay [182].

In terms of security, the adversary is computationally bounded and cannot launch brute-force cryptanalytic attacks on the mix zone encrypted messages. Because messages are authenticated, an external adversary cannot impersonate vehicles, and replay attacks would not be successful either, thanks to time stamps. Similarly, the adversary cannot forge RSU messages or impersonate an RSU; as a result, the adversary cannot create fictitious mix zones.

Recent work by Dahl *et al.* [69] provided a formal analysis of the CMIX protocol. The authors identify an attack in which the CMIX protocol can inadvertently leak information. An adversary can obtain the pseudonym of a vehicle in a mix zone by triggering a key establishment while the vehicle is still in a mix zone. The authors propose a fix of the CMIX protocol that encrypts key establishment operations and impedes such attacks.

2.8 Summary

In this Chapter, we have introduced the location privacy problem in peer-to-peer wireless networks. We described multiple solutions and focused on the description of the multiple pseudonym approach. We introduced the concept of mix zone used to achieve untraceability in wireless networks with the multiple pseudonym approach. We described existing mechanisms to create mix zones and to measure their effectiveness. Because mix zones introduce a cost on both mobile devices and network operators, we argue that privacy-preserving mechanisms must take into account the overhead they introduce in order to, first, predict the effect of cost on the privacy-preserving strategy and, second, understand the effect of privacy-preserving mechanisms on communication networks. Finally, we provided a detailed example of the application of mix zones in vehicular networks. Our proposal suggests to authenticate vehicles in local areas and encrypt their wireless communications for short periods of time in order to provide location privacy without altering potential features of vehicular networks.

Publication: [97]

Chapter 3

Centralized Coordination of Pseudonym Changes

> The appalling present, the awful reality - but sublime, but significant, but desperately important precisely because of the imminence of that which made them so fearful.
>
> Brave New World - Aldous Huxley

3.1 Introduction

While traversing a given area, mobile nodes may go through a *sequence of mix zones*. Between mix zones, the adversary can trivially track the locations of mobile devices. However, at each mix zone, the adversary has to infer the most likely assignment of entering and exiting devices. Considering the confusion created by each mix zone, the adversary may not be able to track trajectories of nodes. In particular, a sequence of traversed mix zones may create an exponential number of possible trajectories making it difficult to track devices. Hence, by traversing a sequence of mix zones, mobile devices "accumulate" untraceability [43, 123].

Unlike wired mix networks, such as Tor [79], where packets can be freely routed, the sequence of mix zones traversed by mobile nodes depends on the mobility of each node. In other words, the path of mobile nodes cannot be controlled to increase the number of traversed mix zones. Instead, we propose to control the *placement* of mix zones. Mix zones can be deployed to maximize both the number of traversed mix zones and their mixing effectiveness.

However, similar to the delay introduced by mix nodes on packets, mix zones induce a cost for mobile nodes. With silent mix zones, mobile nodes cannot communicate while they are in the mix zone, and with a mobile proxy, all messages have to transit through the same mobile node. Hence, the number of deployed mix zones over a given area must be kept small.

In this Chapter, we evaluate centralized algorithms for deploying mix zones. We consider that a trusted central authority (responsible for the establishment of security and privacy in the network) deploys a limited number of mix zones in a given area. We introduce a novel metric, based on mobility profiles, that helps the CA evaluate the effectiveness of possible mix zone locations prior to network operation. Then, we analyze optimal placement of mix

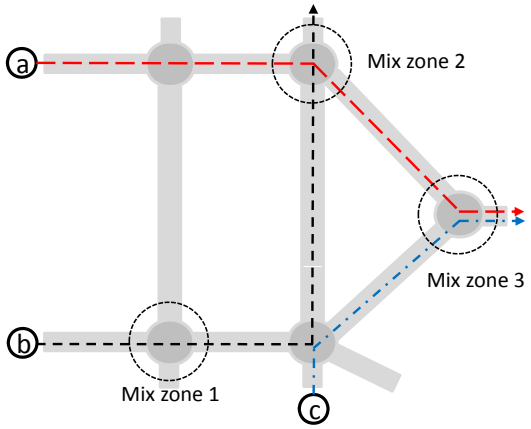


Figure 3.1: Example of flow-based mobility model. Nodes move on plane (x, y) according to trajectories defined by flows a, b, and c. To achieve location privacy, nodes change pseudonyms in mix zones.

zones in a constrained mobility environment with combinatorial optimization techniques. We propose an algorithm to find optimal placement of mix zones by maximizing mixing effectiveness of the system at an acceptable cost for mobile devices. The algorithm offers minimum location privacy guarantees by enforcing a maximum distance between traversed mix zones. Finally, we compare optimal deployment of mix zones to other deployments by using a realistic mobility simulator [5]. Simulations show that the placement recommended by our algorithm significantly reduces tracking success of the adversary.

3.2 Related Work

Huang *et al.* suggest in [123] the use of cascading mix zones. Mix zones are created by repeatedly turning off the transceivers of mobile nodes. The authors evaluate the quality of service implications on real-time applications of users traversing several mix zones, but do not evaluate strategies of mix zones deployments. In [43], Buttyan *et al.* evaluate the performance of sequences of mix zones for vehicular networks. The locations of mix zones correspond to regions where the adversary has no coverage. In their system, the adversary is highly successful in tracking mobile devices because of insufficient mixing of vehicles. In this work, we provide an alternative solution that optimizes placement of mix zones in a considered area.

In wired anonymous networks, multiple mixing strategies were investigated [34]. The mixing network can form mix cascades and force all packets to go through a predefined sequences of mixes. Nodes can also freely select their routing strategy over fully connected mix network [29] or over a restricted mix network [73] (each mix only communicating with a few neighbors). Recent work also evaluated the effectiveness of different mix network topologies to thwart active timing attacks in low latency anonymous communications [76, 131]. In this work, we study an equivalent problem for mobile networks considering optimal positioning of mix zones and its effect on achievable location privacy.

3.3 System Model

As introduced in the system and threat model of Chapter 2, we consider mobile devices equipped with peer-to-peer wireless interfaces and a global passive adversary. In order to measure the mixing provided by different mobile environments, we rely on a generic mobility model based on statistical mobility information. In particular, we use the notion of mobility flows to model the probability to move from one location to another.

3.3.1 Flow-based Mobility Model

As shown by Gonzalez, Hidalgo and Barabasi [105], mobile users tend to regularly return to certain locations (e.g., home and workplace), indicating that, despite diversity of their travel locations, humans follow simple patterns. Hence, we consider a *flow-based* mobility model (Fig. 3.1). Based on real trajectories of mobile nodes (e.g., pedestrian or vehicular), we construct $f \in F$ flows of nodes between few highly frequented locations, where F is the set of all flows. In practice, such real trajectories could be provided, for example, by city authorities in charge of road traffic optimization. Thus, each flow f defines a trajectory shared by several nodes during a period of time. For example, in Fig. 3.1, each node is assigned to one of the three flows a, b, or c and follows the trajectory defined by the flow. In stationary regime, a flow is characterized by its average number of nodes, λ. Note that, during the course of the day, flows usually vary. For simplicity, we consider one of the possible stationary regimes of the system. Flows are defined over the road segments in the considered area. Nodes' mobility is thus bound to road segments.

3.4 Mixing Effectiveness and Flows

Each mix zone i is traversed by flows $f_j \in F_i \subseteq F$. Mobile nodes traversing a mix zone create entering and exiting *events*. Each node in a flow takes a certain amount of time, called *sojourn time*, to traverse the mix zone. Sojourn time models the speed diversity of mobile nodes traversing mix zones. Speed differences can be caused, for example, by a higher density of nodes on specific flows or by traffic lights. Each mix zone i has a set of entry/exit points L_i typically corresponding to the road network. Consider the example in Fig. 3.1: mix zone 3 has three entry/exit points all traversed by some flows. Based on flows traversing a mix zone, we can evaluate different *trajectories* of nodes in each mix zone. The *mobility profile* of a mix zone captures the typical behavior of mobile nodes traversing the mix zone (i.e., their sojourn time and trajectory). In practice, city authorities in charge of traffic lights could provide measured sojourn time distributions as well as typical trajectories over the course of the day as done in [124]. For example in Fig. 3.1, three mix zones have been established, each encompassing an entire intersection.

In order to efficiently place mix zones, we need to know - prior to deployment - their mixing effectiveness. As the previously proposed entropy metric [27] depends on entering/exiting events of mix zones (after deployment), we propose a new metric based exclusively on the mobility profile of mix zones (before deployment).

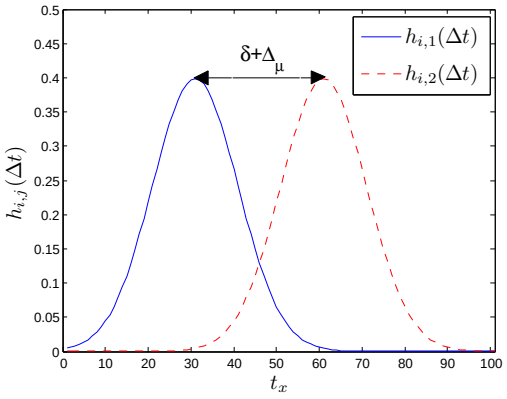


Figure 3.2: Example of exiting time distribution of two flows with $h_{i,j}(\Delta t) \sim \mathcal{N}(\mu_j, \sigma_j)$, $j = 1, 2$. In this example, $(\mu_1, \sigma_1) = (2, 1)$, $(\mu_2, \sigma_2) = (4, 1)$, $\Delta_\mu = \mu_2 - \mu_1$, and δ is the arrival time difference between events of two flows (i.e., the first node arrives at time $t = 0$, and the second one arrives at time δ).

Flow-Based Metric

We propose a new method to analytically evaluate the mixing effectiveness of mix zones. The proposed metric relies on the statistics of the mix zone, i.e., mobility flows and mobility profile, to compute the mixing effectiveness of the mix zone. The advantage of the proposed metric is that the mixing effectiveness can be computed prior to the operation of the mobile network as it does not rely on a particular set of events.

The metric is generic and independent of the nature of traffic. However, to simplify the treatment, we model each flow f_j as a homogeneous Poisson process with intensity λ_j. The distribution $Pois(b; \lambda_j)$ denotes the probability that b nodes enter the flow f_j during a time step t_s. Each flow f_j that traverses a mix zone i is subject to a sojourn time distribution $h_{i,j}(\Delta t)$, where Δt is the time spent in the mix zone. Observing the exit of a mix zone i, the adversary is confronted to a classical *decision-theory problem*: $\mathcal{A}$ must classify each exit event $x \in X$ happening at time t_x as coming from one of the F_i possible entering flows.

Let $m = |F_i|$ be the number of flows in mix zone i. Assume that $m = 2$ flows $\{f_1, f_2\}$ converge to the same mix zone exit l. The probability that the adversary misclassifies x depends on the number of nodes that can potentially correspond to it. This is related to the time spent in the mix zone and the inter-arrival time. We focus on a simple scenario where one mobile node from each flow enters the mix zone. Without loss of generality, we assume that the first mobile node arrives at time $t = 0$ from f_1 and that the second node arrives with a time difference δ from f_2. Figure 3.2 shows the exiting time probability distribution time for a given δ. We first compute the error probability with a fixed value of δ and then generalize our model by considering different values of δ.

To compute the location privacy generated by a mix zone, we are interested in computing the probability that an adversary misclassifies an event. In other words, for one exit l, a successful mixing occurs whenever the adversary makes an error, i.e., assigns an exit event to the wrong flow. It is well known that the decision rule that minimizes the probability of error is the Bayes decision rule (i.e., choosing the hypothesis with the largest a posteriori probability). According to Bayes' theorem, the *a posteriori* probability that an observed

event x belongs to flow f_j is

$$p(f_j|x) = \frac{p_j(x)\pi_j}{\sum_v p_v(x)\pi_v}, j = 1, 2 \tag{3.1}$$

where $p_j(x) = p(x|f_j)$ is the conditional probability of observing x knowing that x belongs to f_j and $\pi_j = p(f_j)$ is the *a priori* probability that an observed exit event belongs to flow f_j. The Bayes probability of error [115] is then given by:

$$p_e(p_1, p_2) = \sum_{x \in X} \min(p_1(x)\pi_1, p_2(x)\pi_2) \tag{3.2}$$

The a priori probabilities depend on the intensity of the flows and are equal to: $\pi_j = \lambda_j/(\sum_{v:f_v \in F_i} \lambda_v)$. The conditional probabilities $p_1(x)$, $p_2(x)$ are equal to the probability that f_j generates an exit event at time t_x: $p_1(x) = \int_{t_x}^{t_x+t_s} h_{i,1}(t)dt$ and $p_2(x) = \int_{t_x}^{t_x+t_s} h_{i,2}(t-\delta)dt$.

A large body of research has focused on minimizing the probability of error. For example, the MTT algorithm minimizes the probability of error when tracking multiple moving objects. In the location privacy context, it is used to measure the effectiveness of path perturbation techniques by Hoh and Gruteser [117]. In our case, we evaluate the probability of error in order to find mix zones with high mixing effectiveness, i.e., that maximize the probability of error. Because computing the probability of error is most of the time impractical [132] (when $m > 2$), we consider the *distance* between the two probability distributions p_1, p_2 to compute bounds on the error probability. Intuitively, the further apart these two distributions are, the smaller the probability of mistaking one for the other should be. The *Jensen-Shannon divergence* [149] (JS) is an information-theoretic distance measure that is particularly suitable for the study of decision problems as the one considered here. It provides both a lower and an upper bound for the Bayes probability of error.

$$JS_\pi(p_1, p_2) = H(\pi_1 p_1(x) + \pi_2 p_2(x)) - \pi_1 H(p_1(x)) - \pi_2 H(p_2(x)) \tag{3.3}$$

where H is the entropy.

The JS divergence (3.3) provides a simple way to estimate the misclassification error of the adversary over a mix zone. The Bayes probability of error is lower/upper bounded as follows [149]:

$$\frac{1}{4}(H(\pi_1, \pi_2) - JS_\pi(p_1, p_2))^2 \leq p_e(p_1, p_2) \leq \frac{1}{2}(H(\pi_1, \pi_2) - JS_\pi(p_1, p_2)) \tag{3.4}$$

where $H(\pi_1, \pi_2)$ is the entropy of the a priori probabilities. The JS divergence is thus particularly useful in order to select mix zones with a high mixing effectiveness. In addition, the JS divergence can be extended to a larger number of flows [149]:

$$JS_\pi(p_1, ..., p_m) = H(\sum_{i=1}^{m} \pi_i p_i(x)) - \sum_{i=1}^{m} \pi_i H(p_i(x)) \tag{3.5}$$

Consider the following example: Two flows f_1, f_2 with equal input Poisson intensities $\lambda_j = 0.2$ share an exit l of mix zone i. The sojourn times are distributed according to a Normal distribution $h_{i,j}(\Delta t) = \mathcal{N}(\mu_j = 2, \sigma_j = 0.5)$, $j = 1, 2$, and $\delta = 0$. Figure 3.3 shows how the lower and upper bounds on the probability of error are influenced by a difference Δ_μ

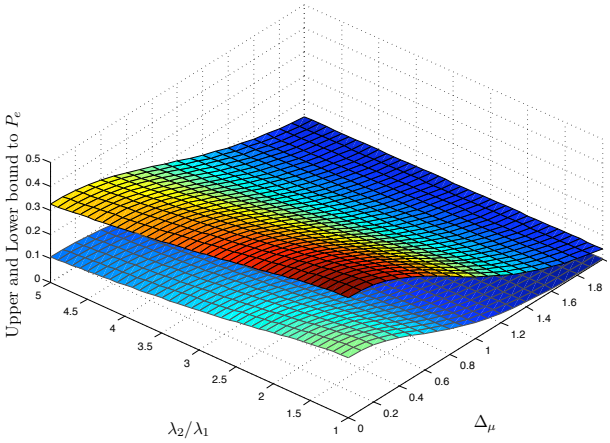


Figure 3.3: Lower and upper bounds of the probability of error with $Pois(b;\lambda_j)$, $h_{i,j}(\Delta t) = \mathcal{N}(\mu_j, \sigma_j = 0.5)$, $j = 1, 2$, $\lambda_1 = 0.2$, $\lambda_2 \in [0.2, 2]$, $\mu_1 = 2s$, and $\mu_2 \in [2, 4]s$. As λ_2/λ_1 increases, the difference between the two probability functions increases as well and it becomes easier to classify the events (p_e becomes smaller). The decrease in p_e is faster if Δ_μ increases as well.

of the sojourn time distributions ($\Delta_\mu = \mu_2 - \mu_1$) and by the ratio λ_2/λ_1 of flows' intensities. We observe that if Δ_μ increases and $\lambda_2/\lambda_1 = 1$, p_e decreases, showing that, with a fixed δ, a difference in the sojourn time distributions alone helps distinguish between the two distributions. We also observe that if λ_2/λ_1 increases and $\Delta_\mu = 0$, the probability of error decreases. The intuition is that as the difference between the flows' intensities increases, the flow with higher intensity dominates the exit of the considered mix zone. In addition, we observe that if both λ_2/λ_1 and Δ_μ increase, p_e decreases faster. The mixing effectiveness is maximal when both flows have the same intensity and sojourn time distribution.

Until now, we focused on scenarios with one mobile node entering from each flow, and a fixed δ. We generalize our model by considering the average difference in arrival time of nodes in flows. More specifically, based on the average arrival rate λ_j, we compute the average difference in arrival time between flows and the average number of nodes that can potentially correspond to an exit event x. The average difference in arrival time between any two flows depends on the flow intensities. The average number of nodes that can be confused with an event x depends on the maximum sojourn time window $\omega_{i,l} = \max_{f_j \in F_{i,l}}(\Delta t_{f_j})$, where Δt_{f_j} is the time spent in the mix zone by nodes in flow f_j and $F_{i,l}$ is the set of flows in F_i that exit at l. For each flow $f_j \in F_{i,l}$, there is a set of possible entering events with average arrival time differences in a time window $\omega_{i,l}$ with respect to beginning of the window: $\zeta^i_{j,l} = \{\delta_{j,v}\colon\ v/\lambda_j \leq \omega_{i,l}, v \in \mathbb{N}\}$, where $\delta_{j,v} = v/\lambda_j$. We compute the probability of error of the adversary at exit l as follows:

$$p^i_{e,l} = \frac{\sum_{f_j \in F_{i,l}} p_e\Big(p_j(x,0), p_{\kappa_1}(x, \delta_{\kappa_1,v_1}), p_{\kappa_1}(x, \delta_{\kappa_1,v_2}), ..., p_{\kappa_2}(x, \delta_{\kappa_2,v_1}), ...\Big)}{|F_{i,l}|} \tag{3.6}$$

where $p_j(x,0)$ is the conditional probability $p_j(x)$ with $\delta = 0$, $p_{\kappa_1}(x, \delta_{\kappa_1,v_1})$ corresponds to the conditional probability $p_{\kappa_1}(x)$ with $\delta_{\kappa_1,v_1} \in \zeta^i_{\kappa_1,l}$, and κ_1, κ_2, ..., κ_{m-1} are not equal to j. In other words, we evaluate the confusion of the adversary for each flow with respect to

other flows. Finally, we compute the average probability of error caused by a mix zone i by considering the error created by each exit $l \in L_i$ of mix zone i:

$$\bar{p}_e^i = \frac{\sum_{L_i} p_{e,l}^i}{|L_i|} \tag{3.7}$$

With this model, we consider the average arrival rate of the nodes and can thus compute the mixing effectiveness prior to network operation. Note that we assumed for simplicity that the sojourn time distribution is independent of the flows' intensity. The model can be extended to capture the interactions between nodes in the mix zone and their effect on the sojourn time distributions [103].

3.5 Pseudonym Change Coordination

In principle, mix zones can be placed anywhere in the considered area. Their placement determines the accumulated location privacy provided by each mix zone. Thus, the optimal solution consists in placing mix zones on the entire surface of the considered area. However, mix zones have a cost because they impose limits on services available to mobile users and require a pseudonym change. Hence, the total number of mix zones deployed in the network should be limited to minimize disruptions on mobile nodes. We assume that a central authority is responsible for organizing mix zones in the network. Users must trust the central authority protects their privacy. We propose a solution based on combinatorial optimization techniques that relies on the divergence metric to select appropriate mix zones. By making a possible algorithm public, our work increases trustworthiness of the authority and provides a basis for comparison.

3.5.1 Mix Zones Placement

After Chaum's seminal work on *mixes* [58], there have been multiple proposals on the way mixes should be connected and organized to maximize anonymity [34]. This led to a classification of different organization concepts. For example, the choice of the sequence of mixes is either distributed (i.e., *mix networks*) or centrally controlled (i.e., *mix cascades*).

The system considered in this work, mix zones deployed over a considered area, presents three different characteristics: (i) Organization of mixes depends on the placement of mix zones in the area, (ii) nodes move in the considered area according to flows constrained by the underlying road network, and (iii) the road network is a connected network with a restricted number of routes. Hence, we must characterize mix zone placements that maximize location privacy.

In order to evaluate location privacy provided by mix zones deployed over a mobile network, one solution is to compute the uncertainty accumulated by the adversary with the joint entropy [199]. However, complexity of the formulation increases with the number of mix zones, making it hard to evaluate. Instead, to compute the overall location privacy, we maximize the total probability of error of the adversary by considering the sum of error probabilities over each deployed mix zone and we guarantee that the distance over which the adversary can successfully track mobile nodes is upper-bounded.

Distance-to-Confusion

The average *distance-to-confusion* (dtc) is defined as the average distance over which the adversary can successfully track mobile nodes before entering a confusion point. A mix zone is a *confusion point* if the error probability of the adversary is larger than a given threshold Tr [118]. As shown in Chapter 1, few location samples can reveal significant information about users. Hence, distance-to-confusion must be minimized. Yet, mix zones impose a cost on mobile nodes that must be taken into account in the mix zone deployment phase. The cost associated to each mix zone depends on the considered application. For example, with silent periods, the cost is typically directly proportional to the duration of the imposed silent period (i.e., the size of the mix zone). Similarly, the cost also depends on the number of used pseudonyms. Pseudonyms are costly to use because they are a limited resource that requires contacting the CA for refill.

By considering both the distance over which the adversary can track mobile nodes, and the probability of error of the adversary at confusion points, we maximize the overall location privacy in the considered area.

3.5.2 Placement Optimization

We model mix zones placement as an optimization problem. Formally, consider a finite set $\mathcal{Z}$ of all possible mix zones' locations, a set F of mobility flows in the system, and a mobility profile for each potential mix zone in the considered area. The goal is to optimize placement of mix zones to maximize the overall probability of error tracking nodes in the considered area, while respecting cost and distance-to-confusion constraints. We select a subset $\widehat{\mathcal{Z}} \subseteq \mathcal{Z}$ of *active* mix zones, which is a solution of the following combinatorial optimization problem:

$$\max_{\widehat{\mathcal{Z}}} \sum_{i \in \mathcal{Z}} \bar{p}_e^i \cdot 1_{\{i \text{ is active}\}} \tag{3.8}$$

$$\text{subject to } \sum_{i \in f_j} w_i \cdot 1_{\{i \text{ is active}\}} \leq W_{max}, \forall f_j \tag{3.9}$$

$$E[dtc(f_j, \widehat{\mathcal{Z}})] \leq C_{max}, \forall f_j \tag{3.10}$$

where $1_{\{i \text{ is active}\}}$ returns one if mix zone i is active and zero otherwise, $\widehat{\mathcal{Z}}$ is the set of active mix zones, $\bar{p}_e^i$ captures the error introduced by mix zone i, w_i is the cost associated with mix zone i, W_{max} is the maximum tolerable cost, $E[dtc(f_j, \widehat{\mathcal{Z}})]$ is the average distance-to-confusion of flow f_j with the set of active mix zones $\widehat{\mathcal{Z}}$, and C_{max} is the maximum tolerable distance-to-confusion. We compute the probability of error $\bar{p}_e^i$ by using the lower bound obtained with the Jensen-Shannon divergence in the previous section. The first constraint limits the number of mix zones that can be deployed per flow by taking into account the cost associated with each mix zone. The second constraint ensures that the average distance-to-confusion is upper bounded, i.e., C_{max} defines a maximal distance over which mobile nodes can be tracked on average.

3.6 Application Example

To test the relevance of our approach, we implemented a simulator in Java that evaluates tracking efficiency.[1] The simulator takes as input a mobility trace and a set of mix zone locations. It first computes the mobility profile of mix zones and then attempts to predict trajectories of mobile nodes using inference attacks.

3.6.1 Simulation Setup

We simulate mobility traces with Sumo [5], an urban mobility simulator, over a cropped map [6] of Manhattan of 6 km^2. Sumo features creation of routes using mobility flows. Each flow is defined by a source, a destination and a traffic intensity. Each mobile node belongs to a single flow and is routed from source to destination over the shortest path. Roads have one lane in each direction, and intersections are modeled with yields. Some roads (e.g., highways) have higher priority and do not have to yield.

In this application example, constraints of the optimization algorithm are defined as follows. The cost of mix zones w_i is proportional to the cost of a pseudonym change γ. We assume that the cost of a pseudonym change is fixed for all nodes, $\gamma = 1$. We set $W_{max} = 3$, meaning that each node can traverse a maximum of three mix zones. Similarly, we set $C_{max} = 2000$m, i.e., the adversary cannot track nodes over more than two kilometers. A total of 40 flows were deployed over the area, generating 1210 nodes in a fluid scenario ($\lambda_j \sim 0.02$) and 2000 nodes in a congested scenario ($\lambda_j \sim 0.04$). The radius of mix zones is a constant $R = 100$m. We simulate a mobile network for 20 minutes with nodes moving at a maximum speed of 50km/h and with an average trip time of 6 minutes. Finally, a mix zone is considered as a confusion point if the introduced error is larger than zero, i.e., $\mathrm{Tr} = 0$.

Mobility Profiles

We consider a powerful (worst-case) adversary that can construct a mobility profile of each mix zone i by measuring the time when nodes enter/exit mix zones. We denote with Q the measuring precision of the adversary, and assume $Q = 1$ second. $\mathcal{A}$ knows for each mix zone: (i) distribution of nodes' trajectories, and (ii) sojourn time distributions. The distribution of nodes' trajectories is captured in a matrix of directions D_i. For each entering/exiting points (k, l), the matrix contains the probability of the trajectory: $D_i^{k,l} = Pr(\text{“ Enter at } k \text{ and exit at } l \text{ ”})$. Sojourn time distribution is captured in a matrix of sojourn times J_i: For each entering/exiting points (k, l), the matrix contains the probability distribution of the sojourn time: $J_i^{k,l}(\Delta t) = Pr(\text{“Enter at } k \text{ and spend } \Delta t \text{ before exiting at } l\text{”})$. Note that a significant amount of memory is required to store mobility profiles. In the simulation scenario considered here, mobility profiles require approximately 2 Gigabytes.

Attack

Based on the mobility profiles, the adversary $\mathcal{A}$ predicts the most probable assignment of entering/exiting mobile nodes for each mix zone. To do so, $\mathcal{A}$ can model entering/exiting events with a weighted bipartite graph, as suggested by Beresford in [26]. Each edge is weighted according to the a priori probability of linking an exiting event at l to an entering

[1] Code is available at: http://mobivacy.sourceforge.net .

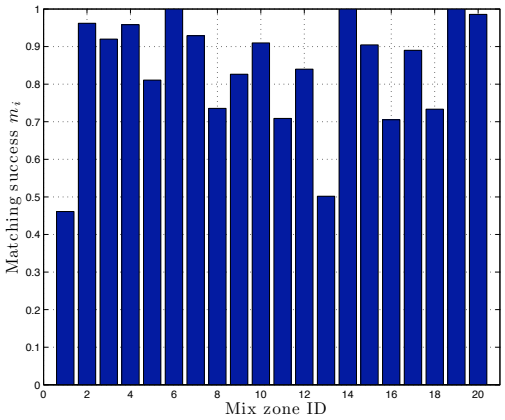


Figure 3.4: Matching success m_i of the 20 potential mix zone locations.

event at k: $D_i^{k,l} \cdot J_i^{k,l}(\Delta t)$. Then, the maximum weight matching of the bipartite graph corresponds to the optimal guess of the adversary. As discussed in [211], a more elaborate attack consists in computing all perfect matchings of the bipartite graph to weight edges, according to the a posteriori probability of linking entering/exiting events. However, this attack has a large complexity, increasing exponentially with the number of entering/exiting pairs and its scalability remains an open problem.

Metrics

Assume that $\mathcal{Z}_s$ is the set of mix zones traversed by node s and let $G_s \subseteq \mathcal{Z}_s$ be the set of mix zones successfully matched by the adversary. $\mathcal{A}$ is *successful* in tracking the location of node s in a mix zone if the real trajectory of node s is correctly guessed. For example, $G_s = \{3, 5, 10\}$ means that node s was successfully tracked in three mix zones.

For each mix zone i, the mixing effectiveness is $m_i = \frac{u_i}{N_i}$ where u_i is the number of successful matches and N_i is the total number of nodes that entered over the course of the simulation. The tracking success of the adversary is defined as the percentage of nodes that can be tracked over k consecutive mix zones: $ts(k) = \frac{N_{suc}(k)}{N(k)}$, where $N_{suc}(k)$ is the number of nodes successfully tracked over k consecutive mix zones, and $N(k)$ is the total number of nodes traversing k consecutive mix zones. This metric reflects the distance over which nodes can be tracked before confusing the adversary.

3.6.2 Results

Mix Zone Performance

Figure 3.4 shows the histogram of mixing effectiveness for 20 potential mix zone locations. We observe that the mixing effectiveness can vary significantly across mix zones and some nodes might experience a poor mixing while traversing a given mix zone. This affects optimal deployment, because mix zones with low mixing effectiveness are sometimes chosen to fulfill the distance-to-confusion constraint. Other than that, the optimization algorithm will tend to choose mix zones that offer lowest tracking success. We observe that mix zones 1 and 13 are particularly effective, whereas mix zones 6, 14 and 19 do not provide any mixing.

# of traversed mix zones	0	1	2	3	4	5	6	7	8	avg	Tracked (%)
Bad (6 mix zones)	68	20	7	5	0	0	0	0	0	0.48	98
Random (10 mix zones)	14	43	24	10	9	0	0	0	0	1.56	78
Optimal (6 mix zones)	14	33	37	16	0	0	0	0	0	1.55	53
Full (20 mix zones)	0	8	24	24	16	14	8	4	2	3.56	48

Table 3.1: Percentage of mobile nodes traversing a certain number of mix zones for various mix zone deployments. The avg column gives the average number of traversed mix zones. The last column gives the percentage of nodes that were successfully tracked over all mix zones in the considered area.

Mix Zone Placement

We consider a total of 20 possible mix zone locations and test 4 deployments of mix zones: (i) *optimal* deployment resulting in 6 deployed mix zones, (ii) *random* deployment of 10 mix zones, (iii) *bad* deployment of 6 mix zones with poor mixing effectiveness, and (iv) *full* deployment where 20 mix zones are in use. We observe in Table 3.1 that in the optimal deployment, the majority of the nodes traverses at least one mix zone and none exceeds the tolerable cost of three mix zones. The random and optimal deployment perform relatively close in terms of the number of traversed mix zones, but with the optimal deployment, fewer nodes are tracked (53%) approaching the performance of full deployment (48%). As expected, bad deployment performs the worst.

The average number of traversed mix zones in Table 3.1 also reflects the total cost. We observe that optimal deployment has a higher cost than bad deployment for the same number of deployed mix zones. However, compared to full deployment, optimal deployment achieves tolerable cost and approaches the same mixing effectiveness.

Tracking Success

We compare the tracking success for optimal, random, bad and full deployments. We observe in Fig. 3.5 (a) that, in general, the probability of success of the adversary decreases as mobile nodes traverse more mix zones. Optimal deployment of mix zones is more effective at anonymizing flows than other deployments and complies with the cost constraint. In particular, optimal deployment is superior to full deployment, since it avoids bad placements.

Note that, in the case of full deployment, traversing more mix zones does not necessarily increase (and actually decreases) location privacy. The reason is that the majority of the flows traversing more than five mix zones actually go through a sequence of ineffective mix zones. Hence, not all flows are equal in terms of the achievable location privacy.

In Fig. 3.5 (b), we observe the effect of an increase in flow intensity λ_j (leading to a congested scenario). Optimal deployment is not affected by the change of intensity because it places mix zones in regions with high traffic density. Random deployment significantly improves mixing effectiveness and approaches the performance of optimal deployment.

In Fig. 3.5 (c), we observe that as the tracking precision Q of the adversary diminishes, so does its ability to track nodes. Reduction of tracking precision of the adversary reflects scenarios where the knowledge of the adversary about mobility profiles is noisy. For example, the adversary may be unable to obtain precise mobility profiles.

In Fig. 3.5 (d), we observe that increasing mix zone radius R from 50 to 100 does not

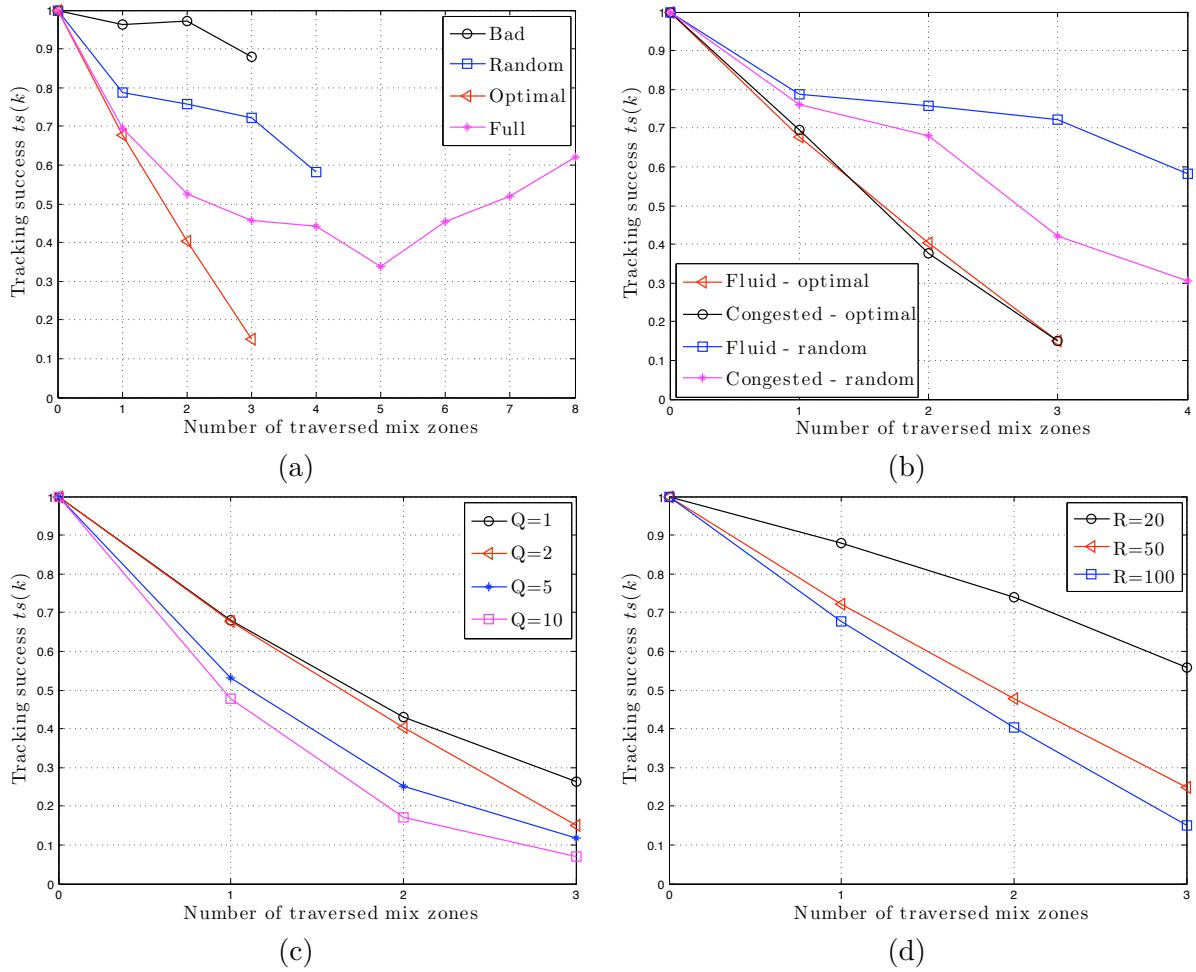


Figure 3.5: Tracking success of adversary $ts(k)$, i.e., the fraction of nodes that can be tracked over k consecutive mix zones. (a) For various mix zones' deployments. (b) In a fluid and congested scenario. (c) With various adversary's precision. (d) For various sizes of mix zone.

increase the mixing effectiveness, whereas, a small radius $R = 20$ dramatically reduces the achieved location privacy. One reason is that changes in speed and direction occur mostly at the center of mix zones. Another reason is that, with $R = 20$, the size of mix zones tends to be smaller than the size of crossroads of the considered map. On one hand, it is important to choose mix zones that are not too small. On the other hand, large mix zones are inappropriate because they do not significantly increase location privacy and incur a high cost.

We also vary the parameters of the optimization problem. Cost w_i, associated with mix zones, changes optimal placement of mix zones. As we increase the cost, fewer mix zones are deployed and achievable location privacy decreases compared to full deployment. Instead, if the tolerable cost increases, optimal deployment performs closer to full deployment in terms of achieved location privacy. Finally, if tolerable distance-to-confusion is lowered, the optimization problem might not have a solution. If there is a solution, it will require more mix zones and will increase the cost per node.

Discussion

Our results show some limitations of mix zones and exhibit the importance of optimizing their placement. An interesting result is that, traversing more mix zones is not necessarily an advantage. Another interesting observation is the high tracking success of the adversary. It may be partly due to our application example. First, we consider a worst-case adversary with global coverage and access to precise mobility profiles ($Q = 1$). Second, we consider a relatively small map with a simple road intersection model. From a general perspective, it is interesting to question our goal of hiding the regular commute of users. After all, this information is mostly publicly available. Home locations are usually public information and can be found phone directories. Similarly, work locations are increasingly shared online and can be found on personal websites. Users also tend to move along the shortest route between these two locations on a regular basis. In Part I of this dissertation, we even showed that users' whereabouts are highly correlated to users' identities and that little information may be sufficient to infer most habits of users. In addition to the attack considered in this work, statistical disclosure attacks [74] could be used to further increase the success of the adversary. Hence, privacy-preserving mechanisms for location privacy should perhaps aim at protecting deviant behavior (users not following their typical routine) instead of the routine itself. In this regard, our results, by indicating the limits of tracking highly correlated information, show that non-standard behavior may be hard to track for an adversary.

3.7 Summary

We considered the problem of constructing a network of mix zones in a mobile network. We first showed how to evaluate the mixing effectiveness of mix zones prior to network operation by using the Jensen-Shannon divergence measure. The proposed metric relies on statistical information about mobility of nodes in mix zones. Then, we modeled the placement of mix zones as an optimization problem by taking into account distance-to-confusion and cost incurred by mix zones. This approach assumes that users trust the central authority responsible for the establishment of security and privacy in the system. Our work, by making a possible algorithm public, contributes to the trustworthiness of the authority as it provides a basis for comparison.

By means of simulations, we investigated the importance of the mix zone deployment strategy and observed that the optimal algorithm prevents bad placement of mix zones and reduces total cost. In addition, we measured the benefit brought by the optimal placement of mix zones, i.e., a 30% increase of location privacy, compared to a random deployment of mix zones. We also noticed that the optimal mix zone placement performs comparatively well to the full deployment scenario, but at a lower cost. As we simulated users' whereabouts with mobility flows modeling humans that move according to reproducible patterns, our results measured the ability of privacy-preserving mechanisms to obfuscate highly correlated spatio-temporal traces. We observed that a global passive adversary could track a large fraction of the nodes and thus infer most of the regular commute of users. The mixing provided by mix zones may thus be insufficient to obfuscate highly correlated location traces. Our results may appear negative at first; yet, the threat of a global passive adversary is unlikely due to the high cost to put in place such attack. In practice, an adversary may obtain lower coverage and noisy mobility profiles. We show that such adversary is considerably less effective at tracking mobile devices. In addition, if a global passive adversary equipped with precise

mobility profiles is unable to track mobile devices following highly reproducible patterns, this indicates that an adversary will not track users deviating from their regular pattern.

Publication: [99]

Chapter 4

Distributed Coordination of Pseudonym Changes

> Il est de toute première instance que nous façonnions nos idées comme s'il s'agissait d'objets manufacturés. Je suis prêt à vous procurer les moules. Mais...
>
> La Solitude - Léo Ferré

4.1 Introduction

Centrally deploying mix zones is not always possible or even sufficient. In some cases, mobility statistics required to estimate the mixing effectiveness of potential mix zones may not be readily available or precise enough. In other cases, the central authority responsible for deployment may not exist or may not be willing to provide such a service. Also, it may be impractical for users to share in real-time optimal locations of mix zones for different mobile environments. Finally, our results of Chapter 3 show some limits in the achievable location privacy of centrally deploying mix zones.

Several researchers advocated the use of a distributed approach [121, 122, 145] whereby mobile nodes coordinate pseudonym changes to dynamically obtain mix zones. To do this, a mobile node simply broadcasts a pseudonym change request to its neighbors. Nearby users independently decide whether to change pseudonyms. The decision depends on the potential location privacy gain and associated cost.

As previously discussed in Chapter 2, pseudonym changes are costly. Pseudonym can quickly become a scarce resource if changed frequently. Hence, existing distributed coordinations of pseudonym changes do not align incentives between mobile nodes. Because the achieved location privacy depends on node density and unpredictability of nodes' movements, rational utility-optimizing agents might not change pseudonyms in settings offering low location privacy.

In this Chapter, we investigate novel distributed algorithms for deploying mix zones. In contrast with other distributed approaches, we consider *rational* mobile devices that locally decide whether to change their pseudonyms. Although selfish behavior can reduce the cost of location privacy, it can also jeopardize the welfare achieved with a location privacy scheme.

We investigate whether the multiple pseudonym approach achieves location privacy in non-cooperative scenarios. We propose a *user-centric location privacy model* that captures the evolution of the location privacy level of mobile users over time and helps them determine when to change pseudonyms. We then define a game-theoretic model - the *pseudonym change game* - that models decisions of mobile nodes in a mix zone.

We first analyze the game with *complete information* (i.e., every node knows the user-centric location privacy level of other nodes) and obtain both pure and mixed Nash equilibria [162]. We show that nodes should coordinate their strategies. Nodes should either cooperate when there is a sufficient number of neighbors with low privacy, or defect. Then, because mobile nodes will, in general, not have good knowledge about payoffs of other nodes, we study, using a Bayesian approach [112], the *incomplete information* scenario. We evaluate (both analytically and numerically) the game model, and derive Bayesian Nash equilibria for a class of threshold strategies where nodes decide whether to change their pseudonyms based on a comparison of their privacy level to a threshold value. We find a symmetric equilibrium, where all nodes cooperate with the same probability, as determined with respect to a distribution over privacy levels. We compare our game-theoretic approach with random and socially-optimal strategies and show that, by using the Bayesian Nash equilibrium, players can reduce their consumption of pseudonyms. Nevertheless, if uncertainty is high, the achievable location privacy is reduced. We then analyze a dynamic version of the game and show that it copes better with uncertainty. Finally, based on the results of the games, we design the PseudoGame protocol that implements the pseudonym change game and evaluate it with simulations.

4.2 Related Work

Game theory is a fundamental tool for evaluating multiplayers decision making: it captures the rational considerations of individual parties and *predicts* the strategies that perform best under different conditions. As security protocols require the active participation of several (rational) parties, game theory can help analyze their strategic behavior. A game-theoretic analysis of security protocols considers the strategies of the parties involved and the nature of their utility. Based on this, the analysis predicts the resulting strategies of rational utility-optimizing entities (i.e., Nash equilibra). If the resulting equilibria are not desirable, game theory can also help us redesign protocols to improve their efficiency.

For these reasons, game theory has been used to evaluate the strategic behavior of mobile nodes in security protocols [14, 45] such as revocation strategies in ephemeral networks [184]. There is also a recent trend of blending game theory with cryptographic mechanisms when rational parties are involved [111, 126, 134, 166]. Game theory has also been used to study privacy. Acquisti [9] explores the reasons why decentralized anonymity infrastructures are still not in wide use today. Varian [216] depicts the role of privacy in economic transactions, showing that because of the advantages of price discrimination consumers may be less inclined to protect their privacy. In this thesis, we study a new aspect of privacy by evaluating how privacy can be achieved among non-cooperative nodes.

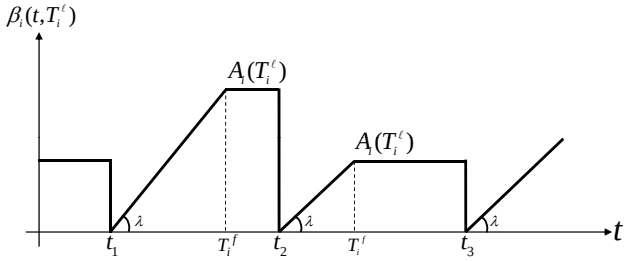


Figure 4.1: Location privacy loss function $\beta_i(t, T_i^\ell)$. At t_1, node i changes pseudonym and updates its time of last successful pseudonym change: $T_i^\ell := t_1$. The function $\beta(t, T_i^\ell)$ increases according to the user sensitivity λ and estimates the time at which mobile i becomes unsatisfied with its location privacy (T_i^f). At t_2, node i changes pseudonym again and updates $T_i^\ell := t_2$.

4.3 System Model

As introduced in the system and threat model of Chapter 2, we consider mobile devices equipped with peer-to-peer wireless interfaces and a global passive adversary. In order to measure the level of location privacy of every mobile device over time, we introduce a user-centric model of location privacy.

4.3.1 User-Centric Location Privacy

The entropy metric evaluates the location privacy achieved in mix zones of the network. However, the location privacy needs of individual users vary depending on time and location. It is thus desirable to protect the location privacy in a user-centric manner, such that each user can decide when and where to protect its location privacy. Hence, we consider a user-centric model of location privacy. *User-centric location privacy* [118, 120, 145] is a distributed approach where each mobile node locally monitors its location privacy level over time. The user centric approach is easily scalable and permits a more fine-grained approach to maintaining location privacy. Each mobile node can evaluate the distance over which it is potentially tracked by an adversary (i.e., the *distance-to-confusion* [118]) and can act upon it by deciding whether and when to change its pseudonym. Whereas, a network wide metric measures average location privacy and might ignore that some nodes have a low location privacy level and are traceable for long distances.

With a user-centric model, mobile nodes can request a pseudonym change from other nodes in proximity when their local location privacy level is lower than a desired level. Nodes in proximity will then choose to cooperate when their location privacy level is low as well. The drawback of the user-centric model is that nodes may have misaligned incentives (i.e., different privacy levels) and this can lead to failed attempts to achieve location privacy.

In this work, we formalize this problem and introduce a *user-centric location privacy model* to capture the evolution of user-centric location privacy level over time. The user-centric location privacy level of each mobile node i is modeled via a *location privacy loss function* $\beta_i(t, T_i^\ell) : (\mathbb{R}^+, \mathbb{R}^+) \rightarrow \mathbb{R}^+$ where t is the current time and $T_i^\ell \leq t$ is the time of the last successful pseudonym change of mobile i. The maximum value of $\beta_i(t, T_i^\ell)$ equals the level of location privacy achieved at the last pseudonym change. The privacy loss is initially

zero and increases with time according to a sensitivity parameter, $0 < \lambda_i < 1$, which models the belief of node i about the tracking power of the adversary. The higher the value of λ_i, the faster the rate of privacy loss increase. For simplicity, we consider that $\lambda_i = \lambda$, $\forall i$. For a given T_i^ℓ, we write:

$$\beta_i(t, T_i^\ell) = \begin{cases} \lambda \cdot (t - T_i^\ell) & \text{for } T_i^\ell \leq t < T_i^f \\ A_i(T_i^\ell) & \text{for } T_i^f \leq t \end{cases} \tag{4.1}$$

where $T_i^f = \frac{A_i(T_i^\ell)}{\lambda} + T_i^\ell$ is the time when the function reaches the maximal privacy loss (i.e., the user-centric location privacy is null). Figure 4.1 illustrates how the function evolves with time. Given this location privacy loss function, the user-centric location privacy of node i at time t is:

$$A_i(t) = A_i(T_i^\ell) - \beta_i(t, T_i^\ell),\ t \geq T_i^\ell \tag{4.2}$$

Time T_i^f is the time at which node i's location privacy will be zero unless it is successful in changing its pseudonym at a new confusion point. Based on the time of the last successful pseudonym change T_i^ℓ, mobile nodes rationally estimate when next to change pseudonyms.[1] Note that, in practice, nodes cannot compute $A_i(T_i^\ell)$ precisely. Hence, we consider that nodes use an approximation such as the upperbound $\log_2(n)$.

In our model, a node's location privacy does not accumulate over time. Rather, it depends only on the number of nodes that cooperate in the last successful pseudonym change. Moreover, mobile nodes are given the ability to control the length of path that is revealed to an adversary before the next pseudonym change. If a mix zone is a strong confusion point (i.e., $A_i(T_i^\ell)$ is large), then a node can choose to reveal a longer distance before changing pseudonym again. If a mix zone is a weak confusion point, a node can attempt another pseudonym change as soon as possible. In doing so, a node has autonomy to control the period of time over which its location can be tracked. Because the achievable location privacy defined by Eq. (2.2) is logarithmic and the location privacy loss function is linear, the user-centric location privacy level will decrease quickly. In our future work, we plan to analyze the effect of other loss functions (e.g., super-linear functions).

4.4 Pseudonym Change Games

In this section, we present the game-theoretic aspects of achieving location privacy with multiple pseudonyms in a selfish environment. We introduce a game-theoretic model that we refer to as the *pseudonym change game G*.

A protocol to change pseudonyms is in general composed of two parts [145]: a pseudonym change *initiation* phase, in which nodes issue request to others asking for pseudonym changes, and a pseudonym change *decision* phase, in which nodes decide upon receiving a request whether to change pseudonyms or not. Pseudonym change games model the latter with game theory. The key point of the game-theoretic analysis is to consider costs and the potential location privacy gain when making a pseudonym change decision.

On one hand, pseudonyms are costly to acquire and use because they are owned in a limited number and require contacting a central authority for refill. Similarly, routing [198]

[1] In a user-centric model, users are actually not involved: the devices make decisions on their behalf.

becomes more difficult and requires frequent updates of routing tables. In addition, while traversing silent mix zones, mobile nodes cannot communicate and thus momentarily lose access to services. We take into account the various costs involved in changing pseudonym in a parameter γ that can be expressed as: $\gamma = \gamma_{acq} + \gamma_{rte} + \gamma_{sil}$, where γ_{acq} is the cost of acquiring new pseudonyms, γ_{rte} is the cost of updating routing tables, and γ_{sil} is the cost of remaining silent while traversing a mix zone. The cost is expressed in privacy units (e.g., bits), causing a decrease in the achieved privacy. Thus, rational mobile nodes might refuse to change pseudonym in order to reduce their costs. Moreover, selfish behavior might jeopardize the achievable location privacy.

On the other hand, the available location privacy gain (upperbounded by the density of nodes and their locations unpredictability) and the user-centric location privacy level might encourage selfish mobile nodes to change pseudonym and obtain a satisfactory location privacy level.

Hence, nodes may delay their decision to change pseudonyms in order to try to find the optimal conditions to maximize the effectiveness of pseudonym changes. Therefore, using a game-theoretic analysis, we investigate whether location privacy can emerge in a non-cooperative system despite the cost incurred by a node in changing its pseudonym, differentiated privacy levels, and the need for coordinated pseudonym changes to achieve a confusion point. We consider rational mobile nodes that maximize their payoff function, which depends on the current location privacy and the associated pseudonym management cost.

4.4.1 Game Model

Game theory allows for modeling situations of conflict and for predicting the behavior of participants. In our *pseudonym change game* G, nodes must decide upon meeting in the network whether to change pseudonym or not. We model the pseudonym change game both as a *static* and *dynamic* game depending on the constraints on the pseudonym change protocol. The static version of the game captures protocols in which nodes are unable to sense their wider environment when deciding whether or not to change its pseudonym, i.e., all nodes stop/start transmitting at the same time. The dynamic version of the game models protocols in which nodes do not start/stop transmitting at the same time and may thus observe each others messages before making their decision.

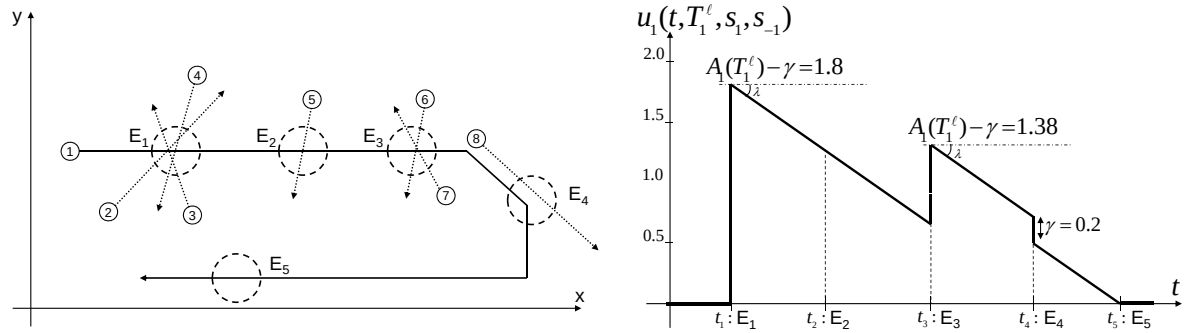
The game G is defined as a triplet $(\mathcal{P}, \mathcal{S}, \mathcal{U})$, where $\mathcal{P}$ is the set of players, $\mathcal{S}$ is the set of strategies and $\mathcal{U}$ is the set of payoff functions. At any time t, several games are played in parallel (but nodes participate in a single game at a time).

Players

The set of players $\mathcal{P} = \{P_i\}_{i=1}^{n(t)}$ corresponds to the set of mobile nodes in transmission range of each other at time t. For a valid game we require $n(t) > 1$. We assume that each node knows the number of other nodes in the mix zone. To achieve a consensus on this number, each node can adopt a neighbor discovery protocol [217].

Strategy

Each player has two moves s_i: *Cooperate* (C) or *Defect* (D). By cooperating, a mobile node changes its pseudonym. The set of strategies of node i is thus $S_i = \{C, D\}$ and the set of strategies in the game is $\mathcal{S} = \{S_i\}_{i=1}^{n(t)}$.



(a) (b)

Figure 4.2: Example of pseudonym change. (a) 7 nodes move on the plane (x, y). (b) Evolution of the payoff of node 1 over time. At t_1 (event E_1 in (a)), nodes 2, 3, and 4 meet in a mix zone and cooperate with node 1. Their payoff u_i and the time of the last successful pseudonym change are updated: $u_i = A_i(T_i^\ell) - \gamma = \log_2(4) - \gamma = 1.8$, and $T_i^\ell := t_1$, $i \in \{1, 2, 3, 4\}$. The payoff of node 1 then decreases according to β_1 with slope λ. At t_2 (event E_2), node 1 defects. At t_3 (event E_3), node 1 cooperates with nodes 6 and 7. Consequently, the 3 nodes update their payoff and the time of the last successful pseudonym change. At t_4, (event E_4) node 1 cooperates but node 8 does not. Hence, the payoff of node 1 decreases by γ. Finally, at $T_1^f = t_5$, the payoff of node 1 reaches 0 (event E_5).

Payoff Function

We model the *payoff* function of every node i as $u_i(t) = b_i(t) - c_i(t)$, where the benefit $b_i(t)$ depends on the level of location privacy of node i at time t, whereas the cost $c_i(t)$ depends on the privacy loss function and the cost of changing pseudonym at time t. If at least two nodes change pseudonyms, then each participating node improves its location privacy for the cost of a pseudonym change γ. If a node is alone in changing its pseudonym, then it still pays the cost γ and, in addition, its location privacy continues to decrease according to the location privacy loss function. If a node defects, its location privacy continues to decrease according to its location privacy loss function. Formally, we have:
If $(s_i = C) \wedge (n_C(s_{-i}) > 0)$,

$$T_i^\ell := t \tag{4.3}$$

$$\alpha_i(t, T_i^\ell) := 0 \tag{4.4}$$

$$u_i(t, T_i^\ell, C, s_i) := \max\left(A_i(T_i^\ell) - \gamma, u_i^- - \gamma\right) \tag{4.5}$$

If $(s_i = C) \wedge (n_C(s_{-i}) = 0)$,

$$u_i(t, T_i^\ell, C, s_i) := \max\left(0, u_i^- - \gamma\right) \tag{4.6}$$

$$\alpha_i(t, T_i^\ell) := \alpha_i(t, T_i^\ell) + 1 \tag{4.7}$$

If $(s_i = D)$,

$$u_i(t, T_i^\ell, D, s_i) := \max\left(0, u_i^-\right) \tag{4.8}$$

where $u_i^- = A_i(T_i^\ell) - \gamma - \beta_i(t, T_i^\ell) - \gamma\alpha_i(t, T_i^\ell)$ is the payoff function at time t^-, which is the time immediately prior to t. s_{-i} is the strategy of the other players, and $n_C(s_{-i})$ is the number of cooperating nodes besides i, and $\alpha_i(t, T_i^\ell)$ is the number of pseudonyms wasted by node i since its last successful pseudonym change T_i^ℓ. (Note that in contrast with the equality sign $=$, the sign $:=$ refers to the assignment of a new value to a variable.)

Figure 4.2 (a) shows seven users moving in a network and playing a total of four pseudonym change games. Figure 4.2 (b) illustrates the evolution of the payoff of node 1 playing the four games. Because we analyze only a single strategic interaction between players, we simplify notation and write in the following $n = n(t)$, $\beta_i = \beta_i(t, T_i^\ell)$, $\alpha_i = \alpha_i(t, T_i^\ell)$, and $u_i(s_i, s_{-i}) = u_i(t, T_i^\ell, s_i, s_{-i})$.

Type

Upon meeting other players, the strategy of a player depends on its knowledge of its opponent payoff function. As both the time of the last pseudonym change and the corresponding location privacy gain are unknown to other players, each player has *incomplete information* about its opponents payoffs. To solve the problem, Harsanyi [102] suggests the introduction of a new player named *Nature* that turns an incomplete information game into an *imperfect information game*. To do so, Nature assigns a type θ_i to every player i according to a *probability density function* $f(\theta_i)$ known to all players, where θ_i belongs to space of types Θ. The type of the players captures the private information of the player, $\theta_i = u_i^-$, where u_i^- is the payoff to player i at time t^- just prior to the current opportunity to change pseudonym. Because γ is common and known to all nodes, this completely defines the payoff of the node.

4.4.2 Equilibrium Concepts

In this section, we introduce the game-theoretic concepts that will help us get an insight into the strategic behavior of mobile nodes. In a complete information game, a pure-strategy for player i is $s_i \in S_i$, where $S_i = \{C, D\}$ is the pure-strategy space. A strategy profile $s = \{s_i\}_{i=1}^n$ defines the set of strategies of the players. Let us write $br_i(s_{-i})$, the best response of player i to the opponent's strategy s_{-i}.

Definition 1. *The best response $br_i(s_{-i})$ of player i to the profile of strategies s_{-i} is a strategy s_i such that:*

$$br_i(s_{-i}) = \arg\max_{s_i} u_i(s_i, s_{-i}) \tag{4.9}$$

If two strategies are mutual best responses to each other, then no player has the motivation to deviate from the given strategy profile. This leads us to the concept of Nash Equilibrium [162].

Definition 2. *A strategy profile s^* is a* Nash equilibrium *(NE) if, for each player i:*

$$u_i(s_i^*, s_{-i}^*) \geq u_i(s_i, s_{-i}^*), \forall s_i \in S_i \tag{4.10}$$

In other words, in a NE, none of the players can unilaterally change his strategy to increase his payoff. A player can also play each of his pure strategies with some probability using *mixed strategies*. A *mixed strategy* x_i of player i is a probability distribution defined over the pure strategies s_i.

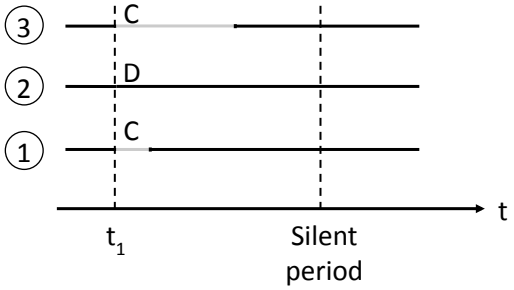


Figure 4.3: Static Pseudonym Change Game. Players 1 and 3 cooperate and remain silent for some time before transmitting again. Player 2 defects and keeps broadcasting as usual.

In an incomplete information game, a pure-strategy for player i is a function $\underline{s}_i : \theta_i \rightarrow S_i$ where $S_i = \{C, D\}$. The pure-strategy space is denoted S_i^{Θ}. A strategy profile $\underline{s} = \{\underline{s}_i\}_{i=1}^n$ is the set of strategies of the players. In incomplete information games, the NE concept does not apply as such because players are unaware of the payoff of their opponents. Instead, we adopt the concept of Bayesian Nash equilibrium [102, 112]. Consider that Nature assigns a type to every player according to a common probability distribution $f(\theta_i)$. Because the type of a player determines its payoff, every player computes its best move based on its belief about the type (and thus the strategy) of its opponents.

Definition 3. *A strategy profile* $\underline{s}^* = \{\underline{s}_i^*\}_{i=1}^n$ *is a pure-strategy* Bayesian Nash equilibrium *(BNE) if, for each player* i*:*

$$\underline{s}_i^*(\theta_i) \in \arg\max_{s_i \in S_i} \sum_{\theta_{-i}} f(\theta_{-i}) \cdot u_i(s_i, \underline{s}_{-i}^*(\theta_{-i})), \forall \theta_i \tag{4.11}$$

4.5 Analysis

In this section, we study several types of pseudonym change games with complete or incomplete information, and two type of strategies static or dynamic.

4.5.1 Static Game with Complete Information

We begin the analysis with a complete information model called the $\mathcal{C}$-game ($\mathcal{C}$ stands for complete information). We assume that there exists only one time step, which means that the players have only one move as a strategy. In game-theoretic terms, this is called a single-stage or static game. This is a realistic assumption because in mix zones, nodes are unable to sense their environment. Hence, each player with common knowledge about the type of all players chooses a strategy simultaneously (Fig. 4.3). We compute the NE for the static 2-player game, and generalize the results for static n-player $\mathcal{C}$-games. We consider that upon a pseudonym change, every node achieves the same level of privacy and thus we consider the upperbound $A_i = \log_2(k)$, where $k \leq n$ is the number of cooperating nodes.

2-player $\mathcal{C}$-game

The strategic representation of the two player $\mathcal{C}$-game is shown in Table 4.1. Two players P_1 and P_2, meeting in a mix zone at time t, take part in a pseudonym change game. Each mobile node decides independently whether to change its pseudonym without knowing the decision of its opponent. The game is played once and the two players make their moves simultaneously. The value in the cells represents the payoff of each player. As usual, the players want to maximize their payoff. We assume here that $u_i^- > \gamma$ for both players, so that $u_i^- - \gamma > 0$. Since u_i^- is itself bounded from above by $\log_2(2) - \gamma = 1 - \gamma$ in a 2-player game, we require $\gamma < 1/2$, so that the cost is bounded.

Table 4.1: 2-player strategic form $\mathcal{C}$-game.

$P_1 \backslash P_2$	C	D
C	$(1-\gamma, 1-\gamma)$	$(u_1^- - \gamma, u_2^-)$
D	$(u_1^-, u_2^- - \gamma)$	(u_1^-, u_2^-)

Each player knows u_{-i}^-, i.e. the payoff of the other player immediately before the game, which is sufficient to define its payoff for different strategy profiles because the cost γ is common knowledge. Theorem 1 identifies the potential equilibrium strategies for the players.

Theorem 1. *The 2-player pseudonym change $\mathcal{C}$-game has two pure-strategy Nash equilibria (C, C) and (D, D) and one mixed-strategy Nash equilibrium (x_1, x_2) where $x_i = \frac{\gamma}{1-u_{-i}^-}$ is the probability of cooperation of P_i.*

Proof. We first prove the existence of the pure-strategy NE. (C, C) is a NE since $1 - \gamma > u_i^-$ for $i = 1, 2$. Similarly (D, D) is a NE because $u_i^- > u_i^- - \gamma$ for $i = 1, 2$. For the mixed strategy NE, let x_i denote the probability of cooperation of u_i. The average payoff of player 1 is:

$$\begin{aligned} u_1(x_1, x_2) &= x_1 x_2(1-\gamma) + x_1(1-x_2)(u_1^- - \gamma) \\ &\quad + (1-x_1)x_2 u_1^- + (1-x_1)(1-x_2)u_1^- \\ &= x_1 x_2(1-u_1^-) - \gamma x_1 + u_1^- \end{aligned}$$

The payoff is maximized for:

$$\frac{\partial}{\partial x_1} u_1(x_1, x_2) = x_2(1-u_1^-) - \gamma = 0$$

which gives $x_2 = \frac{\gamma}{1-u_1^-}$ and by symmetry $x_1 = \frac{\gamma}{1-u_2^-}$. □

We observe that the pseudonym change game is a *coordination game* [67] because $log_2(2) - \gamma > u_i^- > u_i^- - \gamma$. Coordination games model situations in which all parties can realize mutual gains, but only by making mutually consistent decisions. Coordination games always have three NE as obtained with Theorem 1. (C, C) is the Pareto-optimal strategy and thus the preferred equilibrium. If the probability of cooperation x_i of each player equals 1, then the mixed equilibrium equals (C, C). Figure 4.4 illustrates the best response correspondence of the two players. For example, if both players have a low u_i^- (meaning a high propensity to cooperate), the mixed-strategy equilibrium approaches $(0, 0)$. In such a scenario, the basin

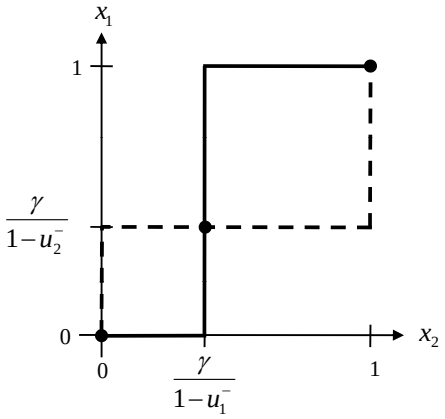


Figure 4.4: Best response correspondence for the 2×2 pseudonym change $\mathcal{C}$-game. The best response function of player P_1 is represented by the dashed line; that of player P_2 is represented by the solid one. The NE are where the two players' best responses cross.

of attraction of the (C, C) NE (i.e., the surface of the rectangle between the mixed NE and the (C, C) NE) is larger than that of the (D, D) NE. In other words, (C, C) would be the most likely NE in settings where players find their best response with an adaptive behavior. The complete information pseudonym change game is *asymmetric* because the payoff of each player depends on its private type. For example, the mixing probability is different for each node (i.e., $x_1 \neq x_2$).

n-player $\mathcal{C}$-game

We extend the 2-player $\mathcal{C}$-game by considering a set of $n \leq N$ players meeting in a mix zone at time t. Each player has complete information and knows the payoff function u_i^- of its $n-1$ opponents. Let C^k and D^{n-k} denote the sets of k cooperating players and $n-k$ defecting players, respectively. Lemma 1 identifies the existence of an All Defection NE.

Lemma 1. *The All Defection strategy profile is a pure-strategy Nash equilibrium for the n-player pseudonym change $\mathcal{C}$-game.*

Proof. All Defection is a NE, because if any player P_i unilaterally deviates from D and cooperates, then its payoff is equal to $u_i^- - \gamma$, which is always smaller than its payoff of defection u_i^-. □

Lemma 2 identifies the existence of NE with cooperation.

Lemma 2. *There is at least one cooperative pure-strategy Nash equilibrium (i.e., at least two players cooperate) for the n-player pseudonym change $\mathcal{C}$-game if there exists a set of cooperating nodes C^{k^*} s.t. $\forall P_i \in C^{k^*}$, $\log_2(|C^{k^*}|) - \gamma > u_i^-$. The strategy profile is then $s^* = \{s_i^* | s_i^* = C \; if \; P_i \in C^{k^*}, s_i^* = D \; if \; P_i \in D^{n-k^*}\}$.*

Proof. If any $P_i \in C^{k^*}$ unilaterally deviates from cooperation to defect, then its payoff u_i^- is smaller than $\log_2(|C^{k^*}|) - \gamma$. Now let D^{n-k^*} be the set of all nodes except those in C^{k^*}. As C^{k^*} is the largest group of nodes where $\log_2(|C^{k^*}|) - \gamma > u_i^-$, no mobile node in D^{n-k^*} can

increase its payoff by joining the set of nodes in C^{k^*}. Hence, none of the nodes can unilaterally change its strategy to increase its payoff and s^* is a NE when $|C^{k^*}| > 1$.

□

Lemma 3. *There are at most $\lfloor \frac{n}{2} \rfloor$ cooperative pure-strategy Nash equilibria for the n-player pseudonym change $\mathcal{C}$-game.*

Proof. Assume that the minimal set of cooperating nodes is $C^{k_1^*}$ s.t. $\forall P_i \in C^{k_1^*}$, $\log_2(|C^{k_1^*}|) - \gamma > u_i^-$. This is the pure-strategy Nash equilibrium with the least number of cooperative players.

We show by contradiction that if another set of cooperating nodes $C^{k_2^*}$ exists, then it must be a superset of $C^{k_1^*}$. Consider $C^{k_1^*}$ and $C^{k_2^*}$ such that $C^{k_1^*} \cap C^{k_2^*} = \emptyset$ and $\forall P_i \in C^{k_j^*}$, $\log_2(|C^{k_j^*}|) - \gamma > u_i^-$ for $j = 1, 2$. There always exists a $C^{k^*} = C^{k_1^*} \cup C^{k_2^*}$ such that $\forall P_i \in C^{k^*}$, $\log_2(|C^{k_1^*}| + |C^{k_2^*}|) - \gamma > u_i^-$ because $\log_2(|C^{k_1^*}| + |C^{k_2^*}|) > \log_2(|C^{k_j^*}|)$ for $j = 1, 2$ and users will merge into the larger group C^{k^*} and create a new cooperative equilibrium. Thus if $C^{k_2^*}$ exists, it must be a superset of $C^{k_1^*}$.

Another set of cooperating players $C^{k_2^*}$ exists if $C^{k_1^*} \subset C^{k_2^*}$ and $\forall P_i \in C^{k_2^*} \setminus C^{k_1^*}$, $\log_2(|C^{k_2^*}|) - \gamma > u_i^- \geq \log_2(|C^{k_1^*}|) - \gamma$. Indeed, with such condition, none of the players in $C^{k_2^*} \setminus C^{k_1^*}$ can deviate from cooperation to unilaterally improve its strategy. Thus, a superset of $C^{k_1^*}$ can be another cooperative NE.

Finally, we observe that $|C^{k_2^*}| - |C^{k_1^*}| \geq 2$ meaning that at least two players must change their strategy to obtain a new NE. Otherwise, one player could unilaterally deviate to improve its strategy. Hence, the *maximum* number of cooperative NE will depend on the number of pairs of players that can exist, i.e., $\lfloor \frac{n}{2} \rfloor$.

□

Considering Lemma 1, 2 and 3, and as there are not any NE in which only one player cooperates, we immediately have the following theorem.

Theorem 2. *The n-player pseudonym change $\mathcal{C}$-game has at least one and at most $\lfloor \frac{n}{2} \rfloor + 1$ pure-strategy Nash equilibria.*

To illustrate the above results, we consider the set of all possible strategy profiles in a 3-player $\mathcal{C}$-game. Assume that $N = 10$, the payoff of each P_i before playing the game is in the interval $[0, \log_2(10) - \gamma]$, depending on the number of nodes that have cooperated with P_i in the past (at T_i^ℓ) as well as the number of failed attempts and the rate of privacy loss. The set of all strategy profiles of this 3-player $\mathcal{C}$-game is: $s = \{(s_1, s_2, s_3) | s_i \in \{C, D\}\}$.

Lemma 1 proves that (D, D, D) is always a NE. From Lemma 2, (C, D, D), (D, D, C), and (D, C, D) are not NE, because $|C^{k^*}|$ must be strictly larger than 1 to satisfy $\log_2(|C^{k^*}|) - \gamma > u_i^-$. Among the remaining strategy profiles, there might be $\lfloor 3/2 \rfloor = 1$ cooperative NE as defined by Lemma 3. The existence of this equilibrium depends on the payoff of each player. Assume that P_3 cooperated with 6 nodes at T_3^ℓ and its payoff is $\log_2(7) - \gamma - \beta_3 - \gamma\alpha_3$ that is bigger than $\log_2(2) - \gamma$ before playing the game. Consider that the payoff of P_1 and P_2 is less than $\log_2(2) - \gamma$ before playing the game. Then, the only cooperative NE strategy profile is (C, C, D), corresponding to $|C^{k^*}| = 2$.

If multiple NE may exist (including cooperate NE and All Defection NE), players have to converge to one of these equilibria. In $\mathcal{C}$-games, a simple technique is to start from the All Cooperation strategy profile and then consider whether certain players have incentive to

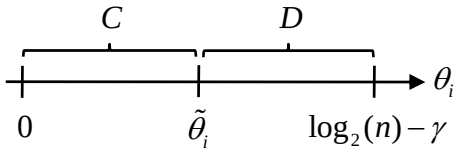


Figure 4.5: Description of the threshold equilibrium in the 2-player $\mathcal{I}$-game. There is a threshold $\tilde{\theta}_i$ that determines the best response of player i.

defect until a NE is found. This apporach also guarantees that players choose the equilibrium with the largest number of cooperative players.

In summary, in $\mathcal{C}$-games, each mobile node tries to reduce its consumption of pseudonyms by changing pseudonyms: (i) only when necessary (i.e., low user-centric location privacy level) and (ii) when the other encountered nodes are willing to cooperate as well. In the 2-player $\mathcal{C}$-game, we prove the existence of two pure and one mixed NE. The payoff to both players in (C,C) is higher than in all other outcomes of the game and thus (C,C) is pareto-optimal. Because the payoffs in the n-player scenario are more asymmetric than those of the 2-player game (i.e., with a larger difference across players), NE with cooperation do not always exist. Still, the All Defection equilibrium always exists because one player cannot gain by cooperating alone. Moreover, we obtain that several NE with cooperation may exist in some cases.

4.5.2 Static Game with Incomplete Information

In this section, we consider games of incomplete information, which we call $\mathcal{I}$-games ($\mathcal{I}$ stands for incomplete information): the players do not know the payoff type of their opponents. The incomplete information assumption better models the knowledge of mobile nodes.

Threshold Equilibrium

In an $\mathcal{I}$-game, players decide their move based on their belief about their opponent's type. Recall that a player's type is defined as $\theta_i = A_i - \beta_i - \gamma\alpha_i - \gamma$; this defines the payoff immediately before the game. We establish an equilibrium in which each player adopts a strategy based on a threshold: if the type of a player is above a *threshold* $\tilde{\theta}_i$, it defects, otherwise it cooperates. Hence, the space of types is divided into two regions (Figure 4.5). A player that has $0 \leq \theta_i \leq \tilde{\theta}_i$ always cooperates, whereas a player with $\tilde{\theta}_i < \theta_i \leq \log_2(n) - \gamma$ always defects. With this *threshold equilibrium*, we define the probability of cooperation of node i as:

$$F(\tilde{\theta}_i) = Pr(\theta_i \leq \tilde{\theta}_i) = \int_0^{\tilde{\theta}_i} f(\theta_i) d\theta_i \tag{4.12}$$

and $1 - F(\tilde{\theta}_i)$ is the probability of defection. The equilibrium strategy at BNE of player i, denoted by $\underline{s}^* = (\tilde{\theta}_1^*; ...; \tilde{\theta}_n^*)$, depends only on the thresholds. In the next section, we obtain the threshold equilibrium for the 2-player $\mathcal{I}$-game.

Remark: In identifying a symmetric BNE with threshold strategies we do not constrain the game so that these are the only strategies available. Rather, we show that a node's best-response is a threshold strategy across all strategies when every other node plays a threshold strategy; i.e., it continues to be a best response even if a node can play a non-threshold strategy, such as playing C for some range of θ_i, then D, then C, and then D again.

2-player $\mathcal{I}$-Game

Each player predicts the type of its opponent based on the probability distribution $f(\theta_i)$. To determine the threshold values that define a BNE, fix a threshold strategy $\underline{s}_2$ associated with threshold $\tilde{\theta}_2$ for player 2, and define the average payoff to player 1 for C and D, given type θ_1, as:

$$\begin{aligned} E[u_1(C, \underline{s}_2)|\theta_1] &= F(\tilde{\theta}_2)(1-\gamma) + (1 - F(\tilde{\theta}_2)) \cdot \\ &\quad \max(0, (\theta_1 - \gamma)) \qquad (4.13) \\ E[u_1(D, \underline{s}_2)|\theta_1] &= \theta_1, \qquad (4.14) \end{aligned}$$

and similarly for player 2. A necessary condition for a threshold equilibrium is that when a player's type is its threshold type it is indifferent between C and D. This is by continuity of payoffs.

So, we can consider the effect of requiring that $E[u_i(C, \underline{s}_{-i})|\tilde{\theta}_i] = E[u_i(D, \underline{s}_{-i})|\tilde{\theta}_i]$ for each player $i \in \{1, 2\}$, directly imposing this condition on the threshold types. This yields a system of two non-linear equations on the two variables $\tilde{\theta}_1$ and $\tilde{\theta}_2$. The following lemma establishes that this is also sufficient: solving for thresholds with this property defines a BNE for the 2-player $\mathcal{I}$-game.

Lemma 4. *The threshold strategy profile $\underline{s}^* = (\tilde{\theta}_1^*, \tilde{\theta}_2^*)$ is a pure-strategy Bayesian Nash equilibrium of the 2-player, incomplete information pseudonym change $\mathcal{I}$-game if:*

$$\begin{cases} E[u_1(C, \underline{s}_2^*)|\tilde{\theta}_1^*] = E[u_1(D, \underline{s}_2^*)|\tilde{\theta}_1^*] \\ E[u_2(C, \underline{s}_1^*)|\tilde{\theta}_2^*] = E[u_2(D, \underline{s}_1^*)|\tilde{\theta}_2^*] \end{cases} \qquad (4.15)$$

Proof. Fix player 2's strategy to threshold $\tilde{\theta}_2^*$ and consider player 1 with type $\theta_1 < \tilde{\theta}_1^*$. We have $E[u_1(C, \underline{s}_2^*)|\tilde{\theta}_1^*] = E[u_1(D, \underline{s}_2^*)|\tilde{\theta}_1^*]$. Now, $E[u_1(D, \underline{s}_2^*)|\tilde{\theta}_1^*] - E[u_1(D, \underline{s}_2^*)|\theta_1] = \tilde{\theta}_1^* - \theta_1 \geq (1 - F(\tilde{\theta}_2^*))(\tilde{\theta}_1^* - \theta_1) \geq E[u_1(C, \underline{s}_2^*)|\tilde{\theta}_1^*] - E[u_1(C, \underline{s}_2^*)|\theta_1]$, where the first inequality follows because $F(\tilde{\theta}_2^*) \geq 0$. Therefore, the drop in payoff from D relative to with type $\tilde{\theta}_1^*$ is at least that from C and a best-response for the player is to play C. Now consider player 1 with type $\theta_1 > \tilde{\theta}_1^*$. By a similar argument, we have $E[u_1(D, \underline{s}_2^*)|\theta_1] - E[u_1(D, \underline{s}_2^*)|\tilde{\theta}_1^*] = \theta_1 - \tilde{\theta}_1^* \geq (1 - F(\tilde{\theta}_2^*))(\theta_1 - \tilde{\theta}_1^*) \geq E[u_1(C, \underline{s}_2^*)|\theta_1] - E[u_1(C, \underline{s}_2^*)|\tilde{\theta}_1^*]$, and the increase in payoff for D is greater than the increase in utility for C and the player's best response is to play D. □

Theorem 3 guarantees the existence and symmetry of the 2-player $\mathcal{I}$-game BNE. As before, we continue to require $\gamma < 1/2$ to make the 2 player game interesting (so that a player retains non-zero privacy value for more than one period after a successful pseudonym change.) For stating the result we assume continuous type distributions, so that probability density $f(\theta_i) > 0$ for all $\theta_i \in [0, 1 - \gamma]$.

Theorem 3. *The 2-player pseudonym change $\mathcal{I}$-game has All Cooperate and All Defect pure-strategy Bayesian-Nash equilibrium, and every threshold equilibrium $\underline{s}^* = (\tilde{\theta}_1^*, \tilde{\theta}_2^*)$ is symmetric for continuous type distributions.*

Proof. To see that *All Defection* is a BNE with thresholds $\tilde{\theta}_1^* = \tilde{\theta}_2^* = 0$, simply note that $E[u_1(C, \underline{s}_2^*)|\tilde{\theta}_1^* = 0] = 0 = E[u_1(D, \underline{s}_2^*)|\tilde{\theta}_1^* = 0]$ and appeal to Lemma 4. Similarly, to see that *All Cooperation* is a BNE consider thresholds $\tilde{\theta}_1^* = \tilde{\theta}_2^* = 1 - \gamma$, for which $F(\tilde{\theta}_1^*) = F(\tilde{\theta}_2^*) = 1$ since $\theta_i \in [0, 1 - \gamma]$. With this, we have $E[u_1(C, \underline{s}_2^*)|\tilde{\theta}_1^* = 1 - \gamma] = 1 - \gamma = E[u_1(D, \underline{s}_2^*)|\tilde{\theta}_1^* = 1 - \gamma]$.

Second, we prove by contradiction the symmetry of any threshold equilibrium. Assume without loss of generality that there exists an asymmetric equilibrium $\underline{s}_2^* = (\tilde{\theta}_1; \tilde{\theta}_2)$, such that $\tilde{\theta}_1 = \tilde{\theta}_2 + \epsilon$, where ϵ is a strictly positive number. Adopt short hand F for $F(\tilde{\theta}_2^*)$ and F_ϵ for $F(\tilde{\theta}_2^* + \epsilon)$. Then, for this to be a BNE we require by Eq. (4.15) that

$$F \cdot (1 - \gamma) + (1 - F) \max(0, \tilde{\theta}_2^* + \epsilon - \gamma) - \tilde{\theta}_2^* - \epsilon = 0 \tag{4.16}$$

$$F_\epsilon \cdot (1 - \gamma) + (1 - F_\epsilon) \max(0, \tilde{\theta}_2^* - \gamma) - \tilde{\theta}_2^* = 0 \tag{4.17}$$

Three cases can be identified considering the values of $\tilde{\theta}_2$, ϵ, and γ.

(Case 1) $\tilde{\theta}_2^* \leq \gamma - \epsilon$. By equating Eq. (4.16) and (4.17) and simplification, we have

$$F(1 - \gamma) - \epsilon = F_\epsilon \cdot (1 - \gamma) \tag{4.18}$$

$$\Rightarrow \ \epsilon = F \cdot (1 - \gamma) - F_\epsilon \cdot (1 - \gamma) < 0, \tag{4.19}$$

since $F_\epsilon > F$ because the type distribution is continuous with $f(\theta_i) > 0$ everywhere. This is a contradiction.

(Case 2) $\gamma - \epsilon < \tilde{\theta}_2^* < \gamma$. By equating Eq. (4.16) and (4.17) and simplification, we have

$$F \cdot (1 - \tilde{\theta}_2^*) + \tilde{\theta}_2^* - \gamma - F\epsilon = F_\epsilon \cdot (1 - \gamma) \tag{4.20}$$

$$\Rightarrow \epsilon = \frac{F \cdot (1 - \tilde{\theta}_2^*) - F_\epsilon \cdot (1 - \gamma) - (\gamma - \tilde{\theta}_2^*)}{F} \tag{4.21}$$

Now, we have $F \cdot (1 - \tilde{\theta}_2^*) - F_\epsilon \cdot (1 - \gamma) < F \cdot (1 - \tilde{\theta}_2^*) - F \cdot (1 - \gamma) = F \cdot (\gamma - \tilde{\theta}_2^*) < \gamma - \tilde{\theta}_2^*$, where the first inequality follows because $F_\epsilon > F$ and the second inequality because $\tilde{\theta}_2^* < \gamma$, by assumption of this case. From this it follows that $\epsilon < 0$ since $F > 0$, and a contradiction.

(Case 3) $\gamma \leq \tilde{\theta}_2^*$. By equating Eq. (4.16) and (4.17) and simplification, we have

$$F \cdot (1 - \tilde{\theta}_2^*) - F\epsilon = F_\epsilon \cdot (1 - \tilde{\theta}_2^*) \tag{4.22}$$

$$\Rightarrow \ \epsilon = \frac{F \cdot (1 - \tilde{\theta}_2^*) - F_\epsilon \cdot (1 - \tilde{\theta}_2^*)}{F} < 0, \tag{4.23}$$

where the inequality holds because $F < F_\epsilon$. This is a contradiction.

□

In simulations we find an intermediate, symmetric threshold equilibrium in almost all cases, where players don't simply always cooperate or always defect.[2]

To illustrate the results of the theorem we consider the following example. Consider that the distribution on types is uniform, with $\theta_i \sim U(0, 1 - \gamma)$, and cumulative probability function $F(\theta_i) = \theta_i/(1 - \gamma)$. Looking for an equilibrium with a threshold, $\tilde{\theta}_i^* \geq \gamma$, so that

[2]Note that previous works [62, 187] obtain similar results showing the existence and symmetry of the BNE for this type of games (infinite games of incomplete information).

the $\max(0, \cdot)$ term in defining the payoff of the cooperation action can be dropped, we can simplify Eq. (4.15) and obtain the system of equations:

$$\tilde{\theta}_i^* \triangleq 1 - \frac{\gamma}{F(\tilde{\theta}_{-i}^*)}, i = 1, 2 \tag{4.24}$$

Imposing symmetry and solving, we obtain $(\tilde{\theta}_i^*)^2 - \tilde{\theta}_i^* + \gamma(1-\gamma) = 0$ for $i \in \{1, 2\}$, which leads to the following solutions:

$$\tilde{\theta}_i^* \in \{\gamma, 1-\gamma\} \tag{4.25}$$

Recall that we assume $\gamma < 1/2$, so that $\gamma < 1-\gamma$. The solution $\tilde{\theta}_i^* = 1-\gamma$ corresponds to an *All Cooperation* BNE because $\theta_i \leq 1-\gamma$ in a two player game. Looking at the intermediate equilibrium when $\tilde{\theta}_i^* = \gamma$, we see that $E[u_1(C, \underline{s}_2^*)|\theta_1] = F(\tilde{\theta}_2^*)(1-\gamma) + (1-F(\tilde{\theta}_2^*)) \cdot 0 = \tilde{\theta}_2^* = \tilde{\theta}_1^*$ while $E[u_1(D, \underline{s}_2^*)|\theta_1) = \theta_1$, and can confirm that C is the best response for $\theta_1 < \tilde{\theta}_1^*$ and D is the best response for $\theta_1 > \tilde{\theta}_1^*$. By further analysis of Eq. (4.15) for the case of $\tilde{\theta}_j^* < \gamma$, there are a multiplicity of symmetric threshold equilibrium in this problem, for *any* $\theta_1^* = \tilde{\theta}_2^* < \gamma$, including $(\underline{s}_1^*, \underline{s}_2^*) = (0, 0)$ which is the *All Defection* BNE. These results are in line with Theorem 3.

We numerically solve Eq. (4.15) to find symmetric threshold equilibrium for three different probability distributions (using *fsolve()* in Matlab). We consider the beta distribution $\mathcal{B}(a, b)$, a family of continuous probability distributions defined on the interval $[0, 1]$ and parameterized by two positive shape parameters a and b. If $\theta \sim \mathcal{B}(2, 5)$, nodes have a small θ with a high probability (i.e., long-tail distribution), whereas with $\theta \sim \mathcal{B}(5, 2)$, nodes have a large θ with a high probability. If $\theta \sim \mathcal{B}(2, 2)$, θ is symmetric and centralized around 0.5. Figure 4.6 shows the BNE $\tilde{\theta}_i^*$ and the related probability of cooperation $F(\tilde{\theta}_i^*)$ as a function of the cost γ. For each distribution of type, we obtain three BNE: $\tilde{\theta}_{i,1}^*$ is an All Defection equilibrium, $\tilde{\theta}_{i,2}^*$ is an intermediate equilibrium, and $\tilde{\theta}_{i,3}^*$ is an All Cooperation equilibrium. With the BNE $\tilde{\theta}_{i,1}^*$ and $\tilde{\theta}_{i,3}^*$, nodes always play the same strategy. With $\tilde{\theta}_{i,2}^*$, we observe that as γ increases, the probability of cooperation $F(\tilde{\theta}_{i,2}^*)$ increases as well, indicating that players should cooperate more when the cost of changing pseudonyms increases. In other words, with a high γ, users care more about the coordination success with others. If γ is small, then the cooperation success becomes less important and nodes become selfish.

The probability of cooperation also depends on the type of Beta distribution. With a lower type distributions $\mathcal{B}(2, 5)$, the probability of cooperation at equilibrium is smaller than other distribution types. In other words, selfish nodes cooperate less because whenever they must change pseudonym, they know that the majority of their neighbors also needs to change pseudonym. On the contrary, for $\mathcal{B}(5, 2)$, selfish nodes cooperate more to maintain high privacy.

In considering the welfare achieved in the pseudonym change game, we focus on the performance under the intermediate BNE $\tilde{\theta}_{i,2}^*$. This is more interesting to study than the *All Cooperation* or *All Defection* equilibrium. We simulate the 2-player $\mathcal{I}$-game in Matlab. The results are averaged over 1000 simulations. We consider three metrics: (i) the welfare of the system defined as the average achieved utility, $E[u_i]$ of the nodes; (ii) the fraction of interactions in which a pseudonym is changed FC; and (iii) the fraction of successful coordination between nodes, CS (i.e., nodes play the same action). We compare the BNE

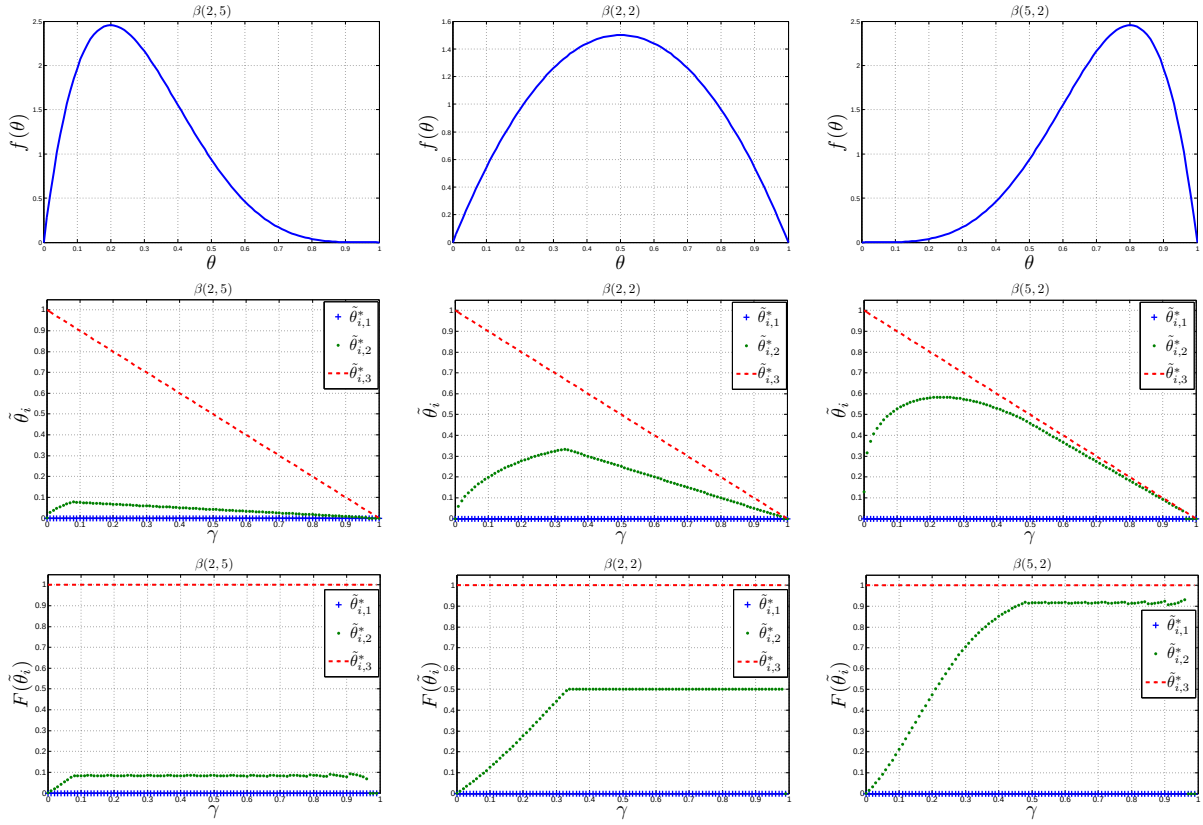


Figure 4.6: Probability distribution of user types $f(\theta)$, threshold $\tilde{\theta}_i^*$, and probability of cooperation $F(\tilde{\theta}_i^*)$ at the equilibrium as a function of γ for different distributions of type: $\beta(2,5)$, $\beta(2,2)$, and $\beta(5,2)$. For each type distribution, there are three BNE: $\tilde{\theta}_{i,1}^*$ corresponds to All Defection, $\tilde{\theta}_{i,3}^*$ to All Cooperation, and $\tilde{\theta}_{i,2}^*$ is an intermediate equilibrium. As the cost γ of changing pseudonyms increases, $\tilde{\theta}_2^*$ approaches $\tilde{\theta}_1^*$, meaning that the probability of cooperation increases.

performance with a *random* strategy, in which all nodes choose their threshold randomly, and to the *socially-optimal* strategy, which is *All Cooperation*.

We observe that the welfare achieved in the BNE is less than with the socially-optimal strategy and in general similar to that of the random strategy. The difference with the random strategy is particularly large for $\mathcal{B}(5,2)$ because the probability of cooperation is then larger than that of the random strategy. It is informative to consider the ratio of welfare in the BNE with that at the socially-optimal, by analogy to the price of anarchy (which considers the performance of the worst-case NE [137]). This ratio provides a measure of the cost of non-cooperative behavior. For example in Table 4.2 for $\mathcal{B}(5,2)$ and $\gamma = 0.3$, we have $0.56/0.70 = 0.80$ meaning that the system performance is degraded by 20%. We notice that the system performance is only degraded by 7% in the case of $\gamma = 0.7$, showing that nodes are less selfish when the cost of a pseudonym change is large. The cost FC in Table 4.2 shows the fraction of interactions in which a pseudonym is changed. We observe that in general less pseudonyms are changed with $\tilde{\theta}_{i,2}^*$ (20% decrease with respect to the random strategy when

Table 4.2: Welfare of system $E[u_i]$, fraction of interactions in which a pseudonym is changed (FC), and fraction of successful coordinations (CS).

Strategy	$E[u_i]$ | FC | CS		
	$\mathcal{B}(2,5)$	$\mathcal{B}(2,2)$	$\mathcal{B}(5,2)$
$\tilde{\theta}^*_{i,2}$, $\gamma = 0.3$	0.20|0.08|0.84	0.39|0.44|0.50	0.56|0.70|0.58
$\tilde{\theta}^*_{i,2}$, $\gamma = 0.5$	0.15|0.09|0.85	0.29|0.49|0.50	0.46|0.91|0.85
$\tilde{\theta}^*_{i,2}$, $\gamma = 0.7$	0.09|0.08|0.85	0.17|0.49|0.49	0.28|0.91|0.85
Random	$(1-\gamma)/2$ | 0.5 | 0.5		
Socially Opt.	$1-\gamma$ | 1 | 1		

$\gamma = 0.3$) showing that less pseudonyms are needed.

n-player $\mathcal{I}$-Game

Assume $n \leq N$ players meet at time t and take part in a pseudonym change $\mathcal{I}$-game. Let $Pr(K = k)$ be the probability that k nodes cooperate. We can again obtain the thresholds that define a BNE in the n-player game by comparing the average payoff of cooperation with that of defection, now defined as:

$$\begin{aligned} E[u_i(C, \underline{s}_{-i})] &= \sum_{k=0}^{n-1} Pr(K = k) u_i(C, \underline{s}_{-i}) \\ E[u_i(D, \underline{s}_{-i})] &= u_i^- \end{aligned}$$

By a similar argument to that for the 2-player $\mathcal{I}$-game (Lemma 4), a BNE $\underline{s}^* = (\tilde{\theta}^*_1; \cdots; \tilde{\theta}^*_n)$ can be obtained as the solution to the following system of n non-linear equations for the n variables $\tilde{\theta}_i$:

$$\sum_{k=0}^{n-1} Pr(K = k) u_i(C, \underline{s}_{-i}) = u_i^-, \quad i = 1, 2, \cdots, n \tag{4.26}$$

We denote the probability of cooperation $q_i = F(\tilde{\theta}_i)$. Assume that the thresholds $\tilde{\theta}^*_i$ are all equal: We obtain $q_i = q$ and thus have a symmetric equilibrium. Consequently, the probability that k nodes cooperate is $Pr(K = k) = \binom{n}{k} q^k (1-q)^{n-k}$. For example, consider the limit values of q:

- If $q \to 0$, then $\tilde{\theta}^*_i = 0$, $Pr(K > 0) = 0$ and $Pr(K = 0) = 1$. Thus, the All Defection equilibrium exists.
- If $q \to 1$, then $\tilde{\theta}^*_i = 1$, $Pr(K < n-1) = 0$ and $Pr(K = n-1) = 1$. Thus, the All Cooperation equilibrium occurs when $\log_2(n) - \gamma > u_i^-$ for all nodes i.

For intermediate values of q, we numerically derive the thresholds $\tilde{\theta}^*_i$ by solving Eq. (4.26) with Matlab (Figure 4.7). For $\gamma = 0.3$, we observe that with a higher density of nodes n, $\tilde{\theta}^*_{i,2}$ decreases, which means that players cooperate with a lower probability. Similarly, $\tilde{\theta}^*_{i,3}$ disappears for large values of n, which means that Always Cooperation is not a BNE anymore. Yet in the case of $\beta(5,2)$, the All Cooperation equilibrium $\tilde{\theta}^*_{i,4}$ persists. The reason is that with

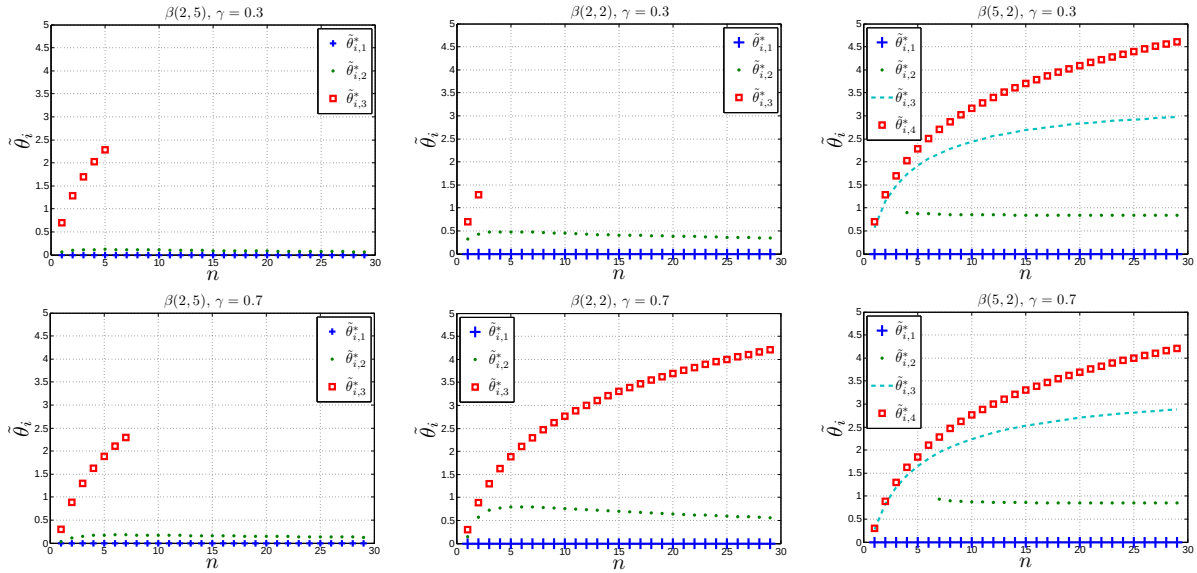


Figure 4.7: Threshold $\tilde{\theta}_i^*$ at the equilibrium as a function of n for different values of γ and distributions of type: $\beta(2,5)$, $\beta(2,2)$, and $\beta(5,2)$. For each type distribution, the number of BNE changes depending on the cost γ.

such a distribution of types, selfish nodes need to cooperate more. For a larger value $\gamma = 0.7$, we observe a similar behavior. Note that with $\beta(5,2)$ an additional threshold equilibrium, denoted by $\tilde{\theta}_{i,3}^*$, appears in which nodes cooperate more when n increases. Moreover, All Cooperation equilibrium survives longer when γ increases.

In summary, we first analytically prove the existence and symmetry of BNE in 2-player games and then obtain numerically three BNE for each possible distribution of users' types. We observe that the intermediate BNE $\tilde{\theta}_{i,2}^*$ reduces the number of pseudonyms used (FC in Table 4.2) and achieves a high level of privacy. However, non-cooperative behavior affects the achievable location privacy. In particular, we notice that, a larger n encourages selfish nodes to not cooperate (Figure 4.7). In contrast, in the 2-player game, when the cost γ of changing pseudonym is high, we observe that selfish nodes cooperate more, which means that a high cost of pseudonyms provides an incentive to cooperate. In summary, even with incomplete information, it is possible to find an equilibrium that achieves high location privacy, and reduces the number of used pseudonyms.

4.5.3 Dynamic Game with Complete Information

Until now, we assumed that the players made their moves simultaneously in mix zones without knowing what the other players do. This is a reasonable assumption because in mix zones, nodes are unable to sense their environment. Yet, nodes could exchange messages in mix zones to advertise their decision. In this case, players have several moves as a strategy and can have sequential interactions: the move of one player can be conditioned by the move of other players (i.e., the second player knows the move of the first player before making his decision). These games are called dynamic games, and we refer to dynamic pseudonym change games with complete information as *dynamic $\mathcal{C}$-games*. We can represent dynamic games by

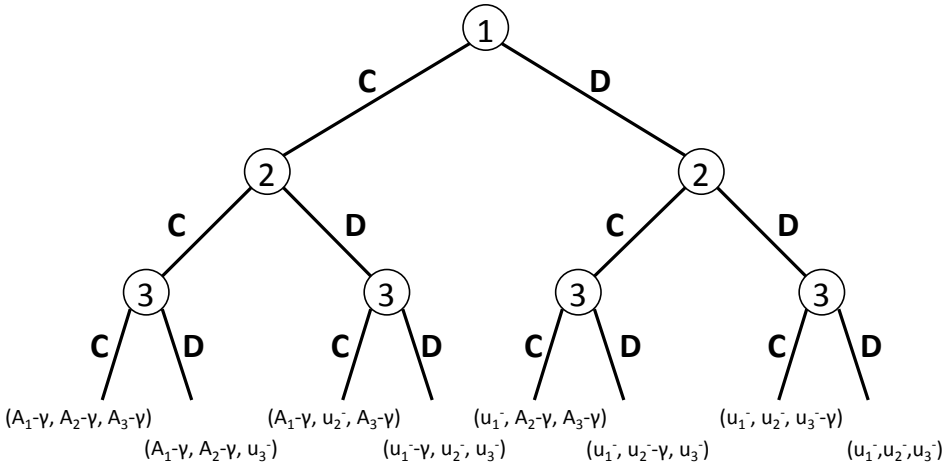


Figure 4.8: Extensive form of the Pseudonym Change Game. The game is represented by a tree and node 1 plays the first action. The game has three stages corresponding to the moves of the three players. The actions (cooperate C and defect D) are represented on each branch of the tree. The leaves of the tree represent the payoff of the game for all players.

their extensive form (Fig. 4.8), similar to a tree where branches represent the strategies for a given player. Each level of the tree represents a stage of the game.

For such dynamic scenarios to exist, nodes must be able to observe the action of other nodes. There are several solutions to do so in practice. A simple solution is that players broadcast their decision to cooperate in a sequential manner [145]. Nonetheless, this increases the communication overhead. Another solution is that players observe the messages of other nodes exiting a mix zone. For example, if a node decides to defect, then it continues broadcasting messages that can be observed by other nodes in the mix zone. In other words, nodes participating in a mix zone can use defection as a signal. Any of these solutions can be used, but we consider the latter because it requires less network resources (Fig. 4.9).

Backward Induction

To predict the outcome of the extensive-form games, one can use the well-known concept of Nash equilibrium. Unfortunately, the Nash equilibrium concept sometimes predicts outcomes that are not credible for some players (i.e., these outcomes are unreachable because the players will not play, out of self-interest, according to the incredible Nash equilibrium path). Hence, we use the stronger concept of subgame-perfect equilibrium. The strategy profile s is a subgame-perfect equilibrium of a finite extensive-form game G if it is a Nash equilibrium of any subgame G' of the original game G [102].

One can check the existence of subgame-perfect equilibria by applying the one-deviation property. This property requires that there exists no single stage in the game, in which a player i can gain by deviating from her subgame-perfect equilibrium strategy while conforming to it in other stages. Hence, we can state that strategy profile s is a subgame-perfect equilibrium of a finite extensive-form game G if the one-deviation property holds. We will check for the existence of subgame-perfect equilibria by backward induction [102].

Backward induction works by eliminating sub-optimal actions, beginning at the leaves of

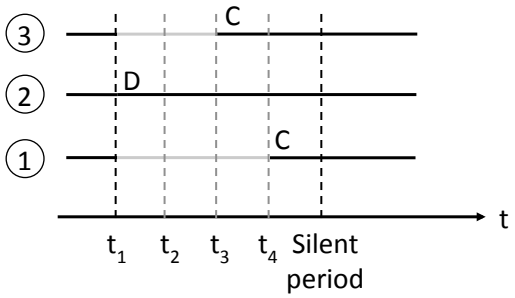


Figure 4.9: Dynamic Pseudonym Change Game. Player 2 defects and keeps broadcasting as usual. Players 1 and 3 notice player 2 defection, but decide to cooperate and remain silent for some time before sending new messages.

the extensive-form tree. The obtained path (sequence of actions) in the game tree defines the backward induction solution and any strategy profile that realizes this solution is a subgame-perfect equilibrium.

n-player Dynamic $\mathcal{C}$-Game

In the complete information scenario, each player can predict deterministically the decisions of other nodes, and its own best response. Hence, for any order of players, the subgame-perfect Nash equilibrium can be derived by all nodes with the following theorem.

Theorem 4. *Let C^{k^*} be a maximal set of cooperating nodes s.t. $\forall P_i \in C^{k^*}, \log_2(|C^{k^*}|) - \gamma > u_i^-$. If there exist such a C^{k^*}, then in the n-player dynamic pseudonym change $\mathcal{C}$-game, there is a strategy that results in a single subgame-perfect equilibrium:*

$$s_i^* = \begin{cases} C & \text{if } P_i \in C^{k^*} \\ D & \text{else} \end{cases} \tag{4.27}$$

If there does not exist such a C^{k^*}, then the subgame perfect equilibrium is all defection.

Proof. Similar to the proof of Lemma 2, no player $P_i \in C^{k^*}$ has an incentive to unilaterally deviate from cooperation to defection as its payoff u_i^- would be smaller than $\log_2(|C^{k^*}|) - \gamma$. The same is true for players that defect, i.e., that are not in C^{k^*}. Hence, none of the nodes can unilaterally change its strategy to increase its payoff and s^* is an subgame-perfect equilibrium when $|C^{k^*}| > 1$. If C^{k^*} is empty, then the subgame-perfect equilibrium corresponds to an All Defection strategy. Because the actions of the players are dynamic, a single subgame-perfect equilibrium will be selected. □

We observe that the All Defection equilibrium does not systematically exist anymore as there is only one subgame-perfect equilibrium. Indeed, one advantage of the dynamic moves is that the All Defection equilibrium will often be an incredible threat. Similarly, among possible cooperative equilibria, the equilibrium with the largest number of cooperating devices is selected. In other words, coordination is simpler in dynamic games than in static games.

4.5.4 Dynamic Game with Incomplete Information

In this section, we consider dynamic games of incomplete information, which we call dynamic $\mathcal{I}$-games. The concept of subgame-perfect Nash equilibrium introduced in the previous section cannot be used to solve games of incomplete information. Even if players observe one another's actions, the problem is that players do not know the others' types and cannot predict each others' strategy.

Dynamic games of incomplete information can be solved using the concept of *perfect Bayesian equilibrium* (PBE). This solution concept results from the idea of combining subgame perfection, Bayesian equilibrium and Bayesian inference. Strategies are required to yield a Bayesian equilibrium in every subgame given the a posteriori beliefs of the players about each others' types. To do so, players update their beliefs about their opponents' types based on others' actions using Bayes' rule. The resulting game is called a dynamic Bayesian game where "dynamic" means that the game is sequential and "Bayesian" refers to the probabilistic nature of the game. For further details, we refer the interested reader to [102].

n-player Dynamic $\mathcal{I}$-Game

Consider that a pseudonym change game starts at time t_0. Every player can decide to cooperate or defect at each stage of the game. Hence, players can delay their decision and enter the game at any time $t \geq t_0$. The actions of players at time t is denoted $a^t = (a_1^t, ..., a_n^t)$ and can be cooperate C, defect D or wait. The history of actions of the game is $h^t = (a^0, ..., a^{t-1})$. The following theorem provides a strategy that leads to the perfect Bayesian equilibrium.

Theorem 5. *In the n-player dynamic pseudonym change $\mathcal{I}$-game, the following strategy results in a unique perfect Bayesian equilibrium:*

$$\underline{s}_i^* = \begin{cases} C & \text{if } (n_D(t) = 0) \wedge (u_i^- < \log_2(n_r)) \\ D & \text{else} \end{cases} \tag{4.28}$$

where $n_r < n$ is the number of nodes remaining in the game (i.e., that did not defect) and $n_D(t)$ is the number of nodes that defect at time t.

Proof. The strategy of players depends on their belief about other players' types. We define $\mu_i\left(\theta_j|h^t\right)$ as the belief of a player i about the type of another player j given a history of actions h^t. In order to obtain a perfect Bayesian equilibrium, Bayes' rule is used to update beliefs from $\mu_i\left(\theta_j|h^t\right)$ to $\mu_i\left(\theta_j|h^{t+1}\right)$. Formally, for all i, j, h^t and a_j, we have:

$$\mu_i\left(\theta_j|\left(h^t, a^t\right)\right) = \frac{\mu_i\left(\theta_j|h^t\right)\sigma_j\left(a_j^t|h^t, \theta_j\right)}{\sum_{\tilde{\theta}_j}\mu_i\left(\tilde{\theta}_j|h^t\right)\sigma_j\left(a_j^t|h^t, \tilde{\theta}_j\right)} \tag{4.29}$$

where σ_j is the probability that a user j plays a certain action a_j. Assume that the number of remaining nodes in the game is n_r (i.e., the number of nodes that did not defect) and that the initial belief function is: $\mu_i(\theta_j) = f(\theta_j)$. If at time $t_1 > t$ player j defects, it indicates that the type of player j is above the current threshold $\tilde{\theta} = \log_2(n_r)$. Hence, the behavior strategy $\sigma_j(a_j^{t_1}|h^{t_1}, \theta_j)$ returns 0 if $\theta_j \leq \tilde{\theta}$ and 1 otherwise. The denominator computes the belief about all possible types of player j and thus normalizes $\mu_i(\theta_j|h^{t_1})$ according the current threshold $\tilde{\theta}$. Other players that observe the action of player j can thus update their belief

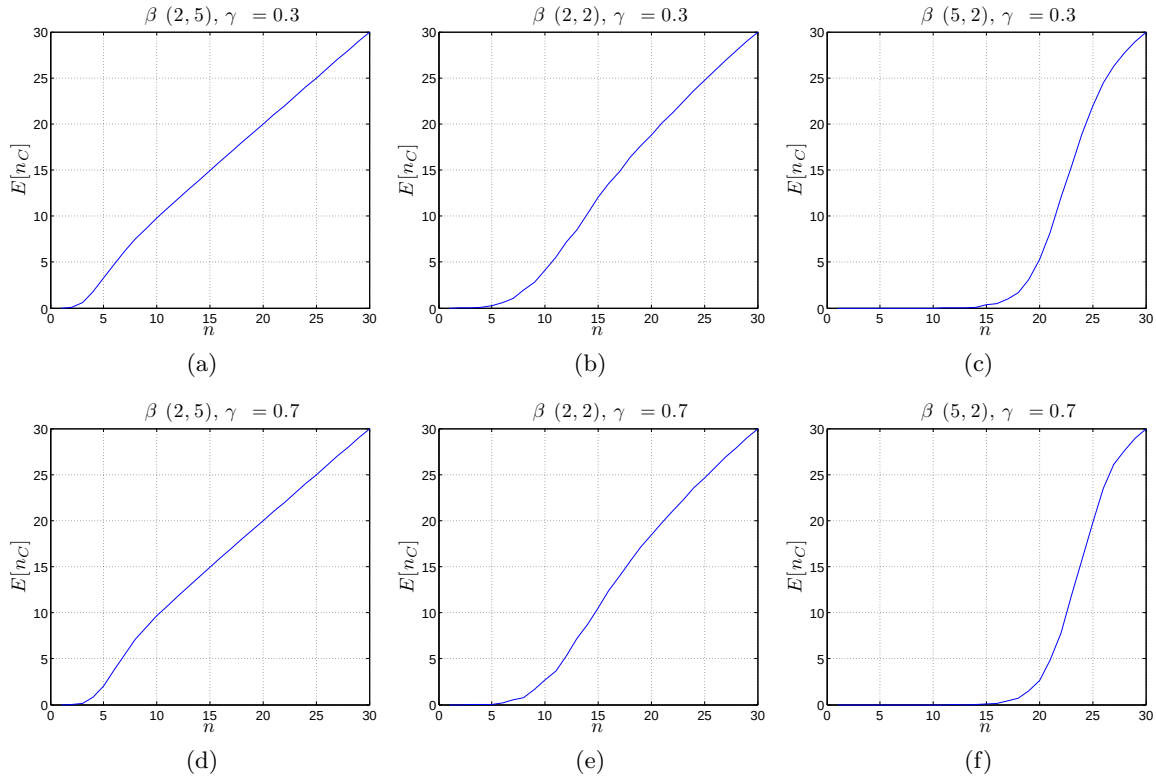


Figure 4.10: Average number of nodes that cooperate in a game with respect to the number of nodes participating in the game for different values of γ and distributions of types: $\beta(2,5), \beta(2,2), \beta(5,2)$.

about the type of player j and obtain: $\mu_i(\theta_j > \tilde{\theta}|h^{t_1}, a^{t_1}) = 1$, i.e., they know that player j had a type above the current threshold. If at some time $t_2 > t_1$ no nodes defect ($n_D(t_2) = 0$), it indicates that with probability one all remaining players have types below the current threshold: $\mu_i(\theta_j \leq \tilde{\theta}|h^{t_2}, a^{t_2}) = 1$. Hence, all these players will cooperate and $\tilde{\theta}^* = \log_2(n_r)$. □

Compared to the static game, the threshold computation is much simpler as it only depends on the number of nodes remaining in the game.

We numerically evaluate the perfect Bayesian equilibrium using Matlab (Fig. 4.10). We compute the average number of nodes that cooperate in dynamic games of incomplete information given distributions of type and cost.

We observe that when the cost of cooperation γ increases, the number of nodes that cooperate decreases. The reason is that, in dynamic games, nodes have more information to optimize their decision and will thus avoid cooperating unless there is a large number of nodes in a game. The distribution of types also affects the number of cooperating nodes. We observe that a large population of nodes with high privacy ($\beta(5,2)$) cooperate less: nodes cooperate only if the privacy gain is large. We also observe that a larger number of nodes in a game, increases the probability of cooperation. In summary, the dynamic version of the game copes well with uncertainty by relying on the action of defecting nodes to improve the estimation of the potential privacy gain.

4.6 Protocols

In this section, we formally describe location privacy protocols, including PseudoGame protocols and evaluate them using simulations. As discussed in Section 4.3.1, pseudonym change games are usually composed of two parts: an initiation phase and a decision phase.

4.6.1 Initiation Protocols

The initiation phase aims at finding appropriate contexts to request pseudonym changes from nearby nodes. A context provides high location privacy if there is high node density and mobility unpredictability. The number of contexts providing high location privacy thus depends on the mobility of the nodes.

NaiveInitiation Protocol

A simple solution consists in issuing a pseudonym change request at every time step t when there is at least another node nearby. The sender can choose a silent period in the range $[sp_{min}, sp_{max}]$ that it attaches to the initiation message. We call this protocol the NaiveInitiation protocol (Protocol 1).

Protocol 1 NaiveInitiation.

1: **if** (At least one neighbor) and (not in silent period) **then**
2: Broadcast initiation message to change pseudonym.
3: **end if**

GainInitiation Protocol

In the GainInitiation protocol (Protocol 2), any node can initiate a pseudonym change by broadcasting an update message if a node has at least one neighbor and if its current location privacy is lower than the potential privacy gain. The sender can choose a silent period in the range $[sp_{min}, sp_{max}]$ that it attaches to the initiation message. This is a protocol similar to that in [145].

Protocol 2 GainInitiation.

1: maxGain = $\log_2$(number of neighbors)
2: **if** (At least one neighbor) and (current location privacy < maxGain) and (not in silent period) **then**
3: Broadcast initiation message to change pseudonym.
4: **end if**

4.6.2 Decision Protocols

Mobile nodes receiving the initiation message must decide whether to stop communicating for a silent period, defined in the initiation message, and change pseudonyms. The decision phase aims at making the best pseudonym change decision to maximize the level of privacy at a minimum cost. Below we describe several decision protocols including protocols proposed in previous work and protocols resulting from the aforementioned game-theoretic analysis.

Swing Protocol

Protocol 3 Swing.

Require: The current location privacy of node i is u_i^-
1: **if** (Receive Initiation message) or (Initiated change) **then**
2: **if** $u_i^- < \tilde{\theta}_i$ **then**
3: Change pseudonym and comply with silent period sp_{max}
4: **else**
5: Quit
6: **end if**
7: **else**
8: Keep pseudonym
9: **end if**

In the Swing protocol (Protocol 3) [145], the decision of mobile nodes to cooperate - or not - exclusively depends on their user-centric level of location privacy compared to a *fixed* threshold $\tilde{\theta}$. The cost of changing pseudonyms and the probability of cooperation of the neighbors are not considered in the computation of the threshold. Hence, this is a reactive model: users change pseudonyms only if their user-centric level of location privacy goes below the threshold.

Static PseudoGame Protocol

Protocol 4 Static PseudoGame.

Require: Node i knows the probability distribution $f(\theta)$
Require: The current location privacy of node i is u_i^-
1: **if** (Receive Initiation message) or (Initiated change) **then**
2: $n \Leftarrow estimate(n)$ //Number of neighbors
3: Calculate $\tilde{\theta}_i^*$ as solution of $\sum_{k=0}^{n-1} Pr(K=k)u_i(C, \mathrm{s}_{-i}) - u_i^- = 0$ wrt $\tilde{\theta}_i$, where $Pr(K=k) \Leftarrow \binom{n}{k} q^k (1-q)^{n-k}$ and $q \Leftarrow \int_0^{\tilde{\theta}_i} f(\theta_i) d\theta_i$
4: **if** $u_i^- \leq \tilde{\theta}_i^*$ **then**
5: Play C
6: Comply with silent period sp_{max}
7: **else**
8: Play D
9: **end if**
10: **else**
11: Keep pseudonym
12: **end if**

Our game-theoretic evaluation allows us to design PseudoGame protocols that extend the Swing protocol to consider optimal strategies of mobile nodes in a non-cooperative environment. The static PseudoGame protocol is based on our results for static n-player $\mathcal{I}$-games.

All nodes receiving the initiation message use the PseudoGame protocol to decide whether to change pseudonyms based on the number of neighbors and the probability of their cooperation (related to the distribution of user types $f(\theta_i)$). As described in Protocol 4 for any node i, the PseudoGame protocol assists mobile nodes in selecting the BNE strategy. Hence, after receiving the initiation message, the nodes calculate the equilibrium thresholds using

their location privacy level, the estimated number of neighbors, and their belief $f(\theta_i)$. The PseudoGame protocol extends the Swing protocol by computing the optimal threshold in a rational environment to determine when to change pseudonym.

Dynamic PseudoGame Protocol

Protocol 5 Dynamic PseudoGame.

Require: Node i knows the probability distribution $f(\theta)$
Require: The current location privacy of node i is u_i^-
if (Receive Initiation message) or (Initiated change) **then**
  $n \Leftarrow estimate(n)$ //Number of neighbors
  lastN $= n$
  for $t = 0$ to sp_{max} **do**
    if $u_i^- \geq \log_2(n)$ **then**
      Play D
      Quit
    end if
    $n =$ number of remaining nodes
    if $n = lastN$ **then**
      Play C
      Comply with silent period sp_{max}
    end if
    lastN $= n$
  end for
else
  Keep pseudonym
end if

The dynamic version of the PseudoGame protocol (Protocol 5) makes use of the action of other nodes as a signal to improve the decision making algorithm. It assists mobile nodes in selecting the PBE strategy.

All Cooperation Protocol

Protocol 6 AllCooperation.

if (Receive Initiation message) or (Initiated change) **then**
  Change pseudonym and comply with silent period sp_{max}
else
  Keep pseudonym
end if

The AllCooperation protocol (Protocol 6) is a straightforward solution in which players always cooperate when asked to change pseudonyms.

Random Decision Protocol

The Random protocol (Protocol 7) is a straightforward solution in which players decide randomly whether to cooperate or not.

Protocol 7 Random.

```
1: if (Receive Initiation message) or (Initiated change) then
2:     Throw a coin
3:     if Heads then
4:         Change pseudonym and comply with silent period sp_max
5:     end if
6: else
7:     Keep pseudonym
8: end if
```

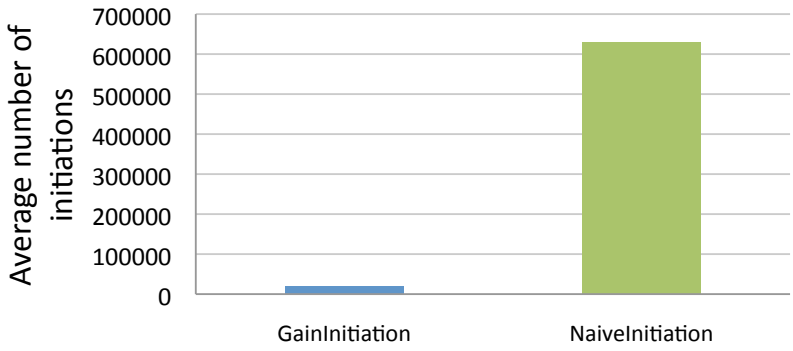


Figure 4.11: Average number of pseudonym change initiations for each initiation protocols using the dynamic PseudoGame.

Evaluation

To evaluate the ability of these protocols to mix pseudonyms, we simulate them in a mobile network by using the simulator[3] presented in Chapter 3. We consider the same simulation setup, i.e., mobility traces generated with Sumo [5] over a cropped map [6] of Manhattan of 6 km^2. We consider an initial distribution of user types $\beta(2,5)$, $\lambda = 0.0005$ and a cost of pseudonym change $\gamma = 0.3$. The results are averaged across 5 simulations.

Figure 4.11 shows the total number of games initiated by each initiation protocol. We observe that the NaiveInitiation protocol generates a larger number of games compared to the GainInitiation protocol. A large number of games will induce networking costs because of all the initiation messages but will also provide more opportunities to change pseudonyms. Yet, the quality of the contexts of the initiated games may be lower. For this reason, we evaluate below the achievable utility for both initiation protocols.

Figure 4.12 shows the average utility obtained with the different initiation and decision protocols. We observe that the initiation protocols do not affect the achievable utility of PseudoGame protocols. Intuitively, the reason is that PseudoGame protocols avoid inefficient pseudonym changes and thus adapt better to different contexts. In constrast, the NaiveInitiation protocol decreases the achievable utility of the AllCooperation, Swing and Random protocols because it increases the number of inefficient pseudonym changes.

In Figure 4.12, we also observe the achievable privacy obtained with different decision protocols. The dynamic PseudoGame achieves the highest level of location privacy among all

[3]The code is available at: http://mobivacy.sourceforge.net .

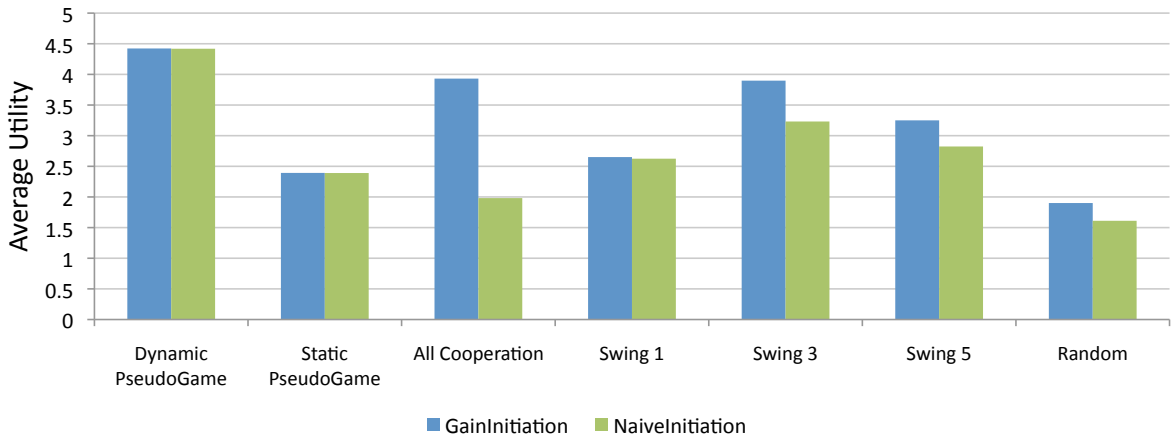


Figure 4.12: Average utility with each decision and initiation protocols. Swing 3 means the Swing protocol with a threshold $\tilde{\theta} = 3$.

protocols, showing that even in the presence of rational behavior high coordination is possible. We observe that in the case of the Swing protocol, a large threshold means that nodes will participate in many inefficient mix zones, whereas a small threshold means that nodes will have to wait too long before changing pseudonyms again. In this regard, $\tilde{\theta} = 3$ appears as an efficient static threshold. Finally, the static PseudoGame performs slightly worse than the Swing protocol, showing that rational behavior negatively affects the achievable privacy in this case. This can be notably observed in Fig. 4.13 that shows the average cost associated with the different protocols. Most of the time, the cost is larger with the NaiveInitiation protocol.

Comparing decision protocols, we observe that the dynamic PseudoGame protocol dramatically reduces the cost compared to other protocols, whereas for the Swing protocol, the cost increases with the threshold. The dynamic PseudoGame protocol provides the best trade-off between privacy and cost, showing that it efficiently deals with the uncertainty of incomplete information. In contrast, the static PseudoGame protocol performs poorly showing that, in the presence of uncertainty caused by static strategies, rationality may not reduce cost compared to Swing, but performs better than AllCooperation and Random.

4.7 Summary

We have considered the problem of rationality in location privacy schemes based on pseudonym changes. We introduced a user-centric model of location privacy to measure the evolution of location privacy over time. To evaluate the strategic behavior of mobile nodes, we proposed a game-theoretic model, the *pseudonym change game*. We first analyzed the n-player scenario with complete information and obtained NE strategy profiles. Then, using Bayesian game theory, we investigated the equilibria in the incomplete information game and derive the equilibrium strategies for each node for both static and dynamic strategies. In other words, we derive equilibria that predict the strategy of rational mobile nodes in order to achieve location privacy in a non-cooperative environment. This analysis results in the design of new protocols, the PseudoGame protocols, that can be used in practice to coordinate pseudonym

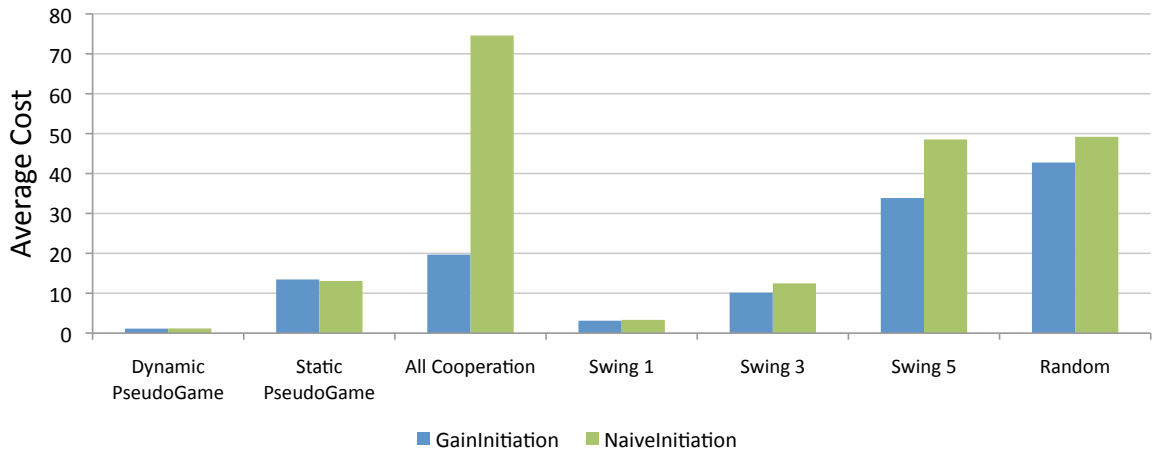


Figure 4.13: Average Cost with each decision and initiation protocols.

changes in distributed environments.

We numerically obtain a particularly interesting result: when uncertainty about others' strategies is high (i.e., static games), rational nodes care more about the successful unfolding of the game if the cost of pseudonyms is also high. This result indicates that cost, a parameter usually having negative consequences, can also have positive consequences: it considerably increases the success of pseudonym change coordination. We also showed that dynamic games, unlike static games, can take advantage of situations with a large number of players to achieve a large level of privacy. By means of simulations, we finally showed that dynamic games (relying on small signaling information from other nodes, i.e., defection), dramatically increase the coordination success of pseudonym changes. The dynamic PseudoGame protocol notably performs better than other protocols for the coordination of pseudonym changes and obtains an efficient trade-off between privacy and cost.

Publication: [95]

Chapter 5

Measuring Location Privacy: A Mean-Field Approach

> He who makes a beast of himself gets
> rid of the pain of being a man.
>
> Samuel Johnson

5.1 Introduction

The degree of location privacy provided by mix zones can be evaluated by measuring the entropy of mix zones and the tracking success of the adversary. Each mix zone acts as a confusion point for the adversary and makes tracking of mobile devices difficult. As observed in [120], the degree of location privacy also depends on how long an adversary can successfully track mobile nodes between confusion points. A longer tracking period may increase the likelihood that the adversary identifies nodes (i.e., the *distance to confusion* [118]). The *age of a pseudonym* refers to the lifetime of a given pseudonym.

Privacy is higher if the pseudonyms are short-lived. However, pseudonyms are costly as described in Chapter 2. Consequently, in many cases a node might consider that its level of privacy is still high enough and might prefer to not change its pseudonym, even if it is located in a mix zone.

Protocols that coordinate pseudonym changes usually limit the maximum age of pseudonyms. With centralized mix zone deployments, the maximum age of pseudonyms is upper bounded in the optimization algorithm. In distributed pseudonym change coordination, the user-centric location privacy model relates the maximum age of a pseudonym to the amount of entropy derived from the last successful pseudonym change. In both cases, we use simulations or empirical mobility data to evaluate the degree of privacy provided by location privacy protocols, in particular, to evaluate the achieved age of pseudonyms.

In this Chapter, we further analyze the distributed coordination of pseudonym changes and provide a framework for *analytically* evaluating privacy obtained with mix zones. This framework captures nodes mobility and evolution of age of pseudonyms over time. We model system dynamics and consequently provide *critical conditions* for the success of the multiple pseudonym approach. We validate our analytical results with simulations.

5.2 Related Work

Several metrics are proposed in the literature to quantify the level of location privacy [78, 88, 107, 117, 118, 120, 199, 202] based on concepts such as: entropy, probability of error or k-anonymity. Ideally, a location privacy metric should be able to measure the inability of a given adversary in accurately tracking the mobile users over space and time. There is ongoing research to identify a satisfactory metric [202]. In this Chapter, we do not attempt to derive another location privacy metric. Instead, we focus on analytically measuring an important aspect of location privacy protocols (related to the degree of privacy).

In wired mix networks, several metrics are proposed to measure anonymity [78, 199] based on the concept of entropy. Recent results have shown how traffic analysis attacks can be used to compute this entropy [212]. Age of pseudonyms is a problem specific to location privacy because the time period over which a given pseudonym is used can help identify the mobile devices (unlike mix networks where packets can be routed in many ways).

5.3 System Model

As introduced in the system and threat model of Chapter 2, we consider a global passive adversary and mobile devices equipped with peer-to-peer wireless interfaces. We consider a generic mobility model that captures all existing mobility models and define the notion of age of pseudonyms.

5.3.1 Mobility Model

We consider a random-trip mobility model characterized by the rate of encounters η, and the average number of nodes met in an encounter $\bar{N}$. The rate η determines the number of encounters with nearby nodes that occur on average. The average $\bar{N}$ is the average number of nodes that participate in each encounter. The meeting rate η and the average $\bar{N}$ depend on nodes' speeds and the topology of the underlying road network.

5.3.2 Location Privacy Model

There are several techniques to mitigate tracking. We consider the use of *multiple pseudonyms*: mobile nodes change their pseudonyms over time to reduce their long-term linkability.

Distance to Confusion or the Age of Pseudonyms

As observed in [120], the degree of location privacy not only depends on the location privacy achieved in mix zones by the nodes traversing it, but also on how long an adversary can successfully track mobile nodes between mix zones. A longer tracking period increases the likelihood that the adversary identifies the mobile nodes. Hence, mobile nodes should evaluate the distance over which they can be tracked by an adversary (i.e., the *distance to confusion* [118]) and act by deciding to change pseudonyms accordingly. To capture the notion of distance to confusion, we define the *age of a pseudonym* as the time period over which a given pseudonym is used.

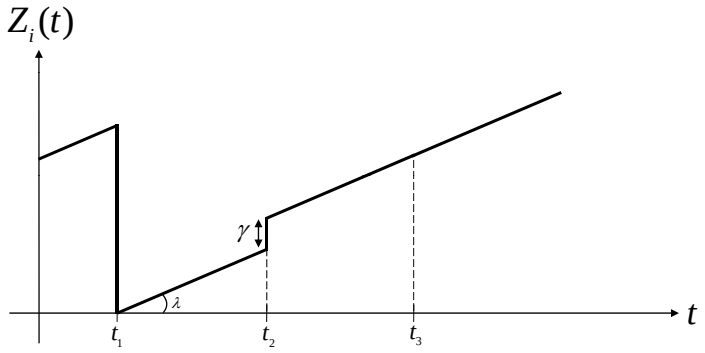


Figure 5.1: Example of evolution of age of pseudonyms. At t_1, node u_i successfully changes pseudonym with another node and the age of its pseudonym drops to zero. The age of pseudonym of node u_i then increases with rate λ. At t_2, pseudonym change fails and node u_i pays the cost γ of changing a pseudonym. At t_3, the node refuses to change its pseudonym.

In this work, we model the evolution of the age of pseudonyms over time $Z_i(t)$ for each mobile node u_i as a linearly increasing function of time with an *aging rate* λ_i:

$$Z_i(t) = \lambda_i \cdot (t - T_i^\ell) \tag{5.1}$$

where t is the current time and $T_i^\ell \leq t$ is the time of the last successful pseudonym change of mobile u_i. The value $Z_i(t)$ captures the age of the current pseudonym of user i at time t (Fig. 5.1). The aging rate λ_i mainly depends on the belief of node u_i with respect to the tracking power of the adversary and on the beaconing rate/range of node u_i. The higher the value of λ_i is, the faster the pseudonyms age. For simplicity, we consider that $\lambda_i = \lambda$, $\forall i$. Note that the function $Z_i(t)$ that models the age of pseudonym corresponds to the privacy loss function $\beta(t, T_i^\ell)$ of Chapter 4.

Strategies

With this location privacy model, mobile nodes request a pseudonym change when the age of their pseudonym is considered large and if there are other nodes in proximity. Nodes in proximity choose to cooperate (C) or defect (D) if their pseudonym age is large as well. Hence, the success of a pseudonym change depends on the state of the neighboring nodes. Asynchronous requests to change pseudonyms might cause failed attempts to achieve location privacy. Assume that n_c is the number of nodes that cooperate (change pseudonyms) in a meeting besides u_i and that $Z_i(t^-)$ is the age of u_i just before making its decision. Considering an encounter in a mix zone at time t, we write for node u_i:
If $C \wedge (n_c > 0)$,

$$T_i^\ell = t \tag{5.2}$$
$$Z_i(t) = 0 \tag{5.3}$$

If $C \wedge (n_c = 0)$,

$$Z_i(t) = Z_i(t^-) + \gamma \tag{5.4}$$

If D,

$$Z_i(t) = Z_i(t^-) \tag{5.5}$$

In other words, $Z_i(t)$ is reset to 0 when a pseudonym change is successful. If a node is alone in changing its pseudonym, then it pays the cost of changing pseudonym γ and the age of its pseudonym keeps increasing. The cost γ can be expressed as: $\gamma = \gamma_{acq} + \gamma_{rte} + \gamma_{sil}$, where γ_{acq} is the cost of acquiring new pseudonyms, γ_{rte} is the cost of updating routing tables, and γ_{sil} is the cost of remaining silent while traversing the mix zone. The cost γ is expressed in age units (i.e., time), causing an increase in the age of pseudonyms. The cost γ captures the failed opportunity of a pseudonym change and is thus an incentive to carefully manage pseudonyms. Finally, if a node defects, its pseudonym age is unchanged. Figure 5.1 illustrates how the age of pseudonyms evolves with time in the case of meetings between several nodes. With this model, nodes control the distance over which they can be tracked.

Nodes decide when to change pseudonyms next based on the time of their last successful pseudonym change T_i^ℓ. The probability distribution over the age $c_i(z)$ gives the probability of cooperation of each node u_i. For simplicity, we assume that the distribution is the same for all nodes and we write $c_i(z) = c(z)$. When several nodes meet, each decides whether to change its pseudonym with probability $c(z)$.

5.3.3 Metric

We are interested in measuring the success of the multiple pseudonym approach. A pseudonym change is successful only if it is coordinated with other nodes nearby. In order to evaluate the ability of nodes to synchronize, we measure the distribution of the age of pseudonyms in the network. We define $Z(t) \sim f(z,t)$ as a random variable that describes the density of probability for any age z. The cumulative distribution function (CDF) $F(z,t) = \int_z f(x,t)dx$ gives the fraction of nodes u_i at time t whose age of pseudonym is $Z_i(t) \leq z$.

5.4 Analytical Evaluation

In this section, we analytically derive the probability distribution of the age of pseudonyms, $F(z,t)$. To do so, we calculate the fraction of users whose age of pseudonyms is lower than z, i.e. $Pr\{Z \leqslant z\}$. We show that the evolution of the age of pseudonyms can be approximated by a dynamical system composed of a simple differential equation when the number of nodes N gets large.

5.4.1 Dynamical System

As discussed above, the random variable $Z(t)$ models the distribution of the age of pseudonyms at time t. The evolution of this random variable over time can be captured by a dynamical system composed of *drift* and *jump processes*. The goal of the *drift* and *jump* processes is to capture the dynamics of the age of pseudonyms by modeling possible variations of a pseudonym age. The drift models an increase in the age of the pseudonyms, and a jump models sudden changes in the age of the pseudonyms (e.g., upon using a new pseudonym).

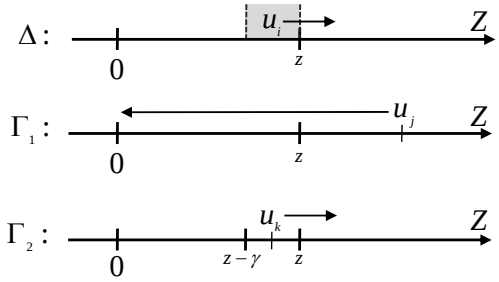


Figure 5.2: Illustration of the drift (Δ) and jump (Γ_1 and Γ_2) processes. We observe the three scenarios causing a change in the distribution $F(z,t)$.

Drift Process Δ

The drift process models the aging of pseudonyms over time as shown in Fig. 5.2. At each time step, the age of the pseudonym of every node is incremented with rate λ. Hence, for any fixed value z, the age of the pseudonyms of a fraction of nodes will pass above z and decrease $F(z,t)$. The drift directly depends on the aging rate λ and the density of the age of the pseudonyms:

$$\Delta = \lambda \frac{\partial F}{\partial z} \tag{5.6}$$

Jump Process Γ

The jump process captures the sudden variations in the age of pseudonyms. There are two possible scenarios, Γ_1 and Γ_2, that correspond to successful and failed attempts to change pseudonyms, respectively (Fig. 5.2). In the first type of jump Γ_1, a node u_j with a pseudonym age greater than z successfully changes its pseudonym with other nodes in proximity. Hence, the age of its pseudonym drops to 0. This happens with rate:

$$\Gamma_1 = \eta \int_0^\infty c(x)q(t)(1 - 1_{\{x \leq z\}}) \frac{\partial F}{\partial x}(x,t)dx \tag{5.7}$$

where $q(t)$ is the probability that at least one of the encountered nodes changes pseudonym; $c(z)$ is the probability of cooperation of user u_j given that its age of pseudonym is z, and η is the rate of meetings scaled by the number of nodes. Intuitively, Γ_1 (the rate at which any node u_j successfully changes pseudonym) depends on: (1) the rate of encounter between nodes η, (2) on the probability that u_j cooperates $c(z)$, (3) on the probability of meeting a nodes that cooperates $q(t)$, and (4) on the probability of having a pseudonym age larger than z. The integral captures the probability that a node u_j has a pseudonym age larger than z, cooperates, and meets at least one cooperative node, thus causing an increase in $F(z)$.

In the second type of jump process Γ_2, a user u_k with a pseudonym age between $z - \gamma$ and z changes pseudonym in an encounter with other nodes. However, none of the nodes in proximity cooperate, and the pseudonym change is a failure. Hence a pseudonym is wasted and user u_k suffers a cost γ: $Z_k(t) = Z_k^- + \gamma$ causing an increase in the number of users with

an age of pseudonym larger than z. This occurs with rate:

$$\Gamma_2 = \eta \int_{z-\gamma}^{z} c(x)(1-q(t))\frac{\partial F}{\partial x}(x,t)dx \tag{5.8}$$

Intuitively, Γ_2 (i.e., the rate at which u_k fails to change pseudonyms) depends on the rate of encounter η, on the probability that u_k cooperates, on the probability of meeting nodes that all defect $(1-q(t))$, and on the probability that u_k has a pseudonym age in the interval $[z-\gamma, z]$. The integral captures the probability that a node u_k has a pseudonym age in the interval $[z-\gamma, z]$, cooperates, and meets nodes that all defect, thus causing a decrease in $F(z)$.

5.4.2 Differential Equation

Taking into account the drift and jump processes, we obtain a dynamical system defined by a single differential equation. The cumulative distribution function $F(z,t)$, giving the fraction of nodes with an age of pseudonym smaller than z, is the unique solution of the following differential equation:

$$\frac{\partial F}{\partial t} = -\Delta + \Gamma_1 - \Gamma_2 \tag{5.9}$$

with boundary conditions: $F(\infty, t) = 1, \forall t$. Intuitively, on one hand, the drift Δ and the jump Γ_2 cause nodes to have an age larger than z, hence decreasing the fraction of nodes $F(z)$. For this reason, they are subtracted from $\frac{\partial F}{\partial t}$. On the other hand, the jump Γ_1 increases the number of nodes on the left size of z, hence increasing $F(z)$. For this reason, it is added to $\frac{\partial F}{\partial t}$.

As defined above, $q(t)$ is the probability that at least one of the encountered nodes cooperates. It can be calculated by considering the probability of meeting n nodes and the probability that at least one node cooperates:

$$q(t) = 1 - \sum_{n\geq 0} h_n(1-\bar{c}(t))^n = 1 - H(1-\bar{c}(t)) \tag{5.10}$$

where h_n is the probability of meeting n nodes (a meeting involves $n+1$ nodes: the node itself with the n encountered nodes), $H(\mathbf{z}) = \sum_{n\geqslant 0} \mathbf{z}^n h_n$ is the $\mathcal{Z}$-transform of h_n, and $\bar{c}(t)$ is the probability that an encountered node cooperates:

$$\bar{c}(t) = \int_0^{\infty} c(z) f(z,t) dz \tag{5.11}$$

Intuition of the equation above: The main idea is to replace all interactions between nodes with an average interaction. This can be done by using the principles of Mean Field theory. To do so, we consider the probability that each node has a certain age in the system (e.g., $f(z)$). Previous work [25, 56] has shown that such probability distribution function converges to a deterministic limit (mean field convergence) when N goes to infinity. The probability distribution function is known to satisfy an ordinary differential equation formed by drift and jump processes that capture possible transitions in the age of pseudonyms. In summary, by considering possible scenarios that affect the age of pseudonyms, we derive above differential equation characterizing the distribution of the age of pseudonyms.

5.5 Analytical Results

In this section, we solve the differential equation (5.9) characterizing the age of pseudonyms. We consider the system in the stationary regime (i.e., as t goes to infinity, we have $\frac{\partial F}{\partial t} = 0$) and evaluate how system parameters such as η, λ, θ, and $c(z)$ affect the distribution of the age of pseudonyms $F(z,t)$.

We assume that a node cooperates according to a simple threshold function. This means that if its age of pseudonym is smaller than a given threshold θ, it decides not to cooperate, whereas it cooperates with probability c_0 if its age of pseudonym is larger than θ:

$$c(z) = \begin{cases} 0 & z \leqslant \theta \\ c_0 & z > \theta \end{cases} \tag{5.12}$$

Intuitively, a node will tend not to cooperate as long as it estimates that the age of its pseudonym (or its distance to confusion) is sufficient. With this model, the threshold θ and the probability c_0 determine the inclination of each node to cooperate. For example, a low θ and a high c_0 mean that the nodes will often change their pseudonyms. These parameters directly affect the probability distribution of the age of pseudonyms. Consequently, we can fine tune the achievable level of privacy in the system.

As mentioned, we have $\frac{\partial F}{\partial t} = 0$ in the stationary regime. For simplicity, we derive Equation (5.9) with respect to z and as $\frac{\partial F}{\partial z}(z,t) = f(z,t)$, we obtain:

$$\begin{cases} \lambda \frac{\partial f}{\partial z} + \eta c(z) f(z) - \eta(1-q) c(z-\gamma) f(z-\gamma) = 0 \\ \\ \int_0^\infty f(z) dz = 1 \end{cases} \tag{5.13}$$

Considering the probability of cooperation $c(z)$ defined by Equation (5.12), the above differential equation must be solved in three intervals:

1. $z < \theta$: The probability of cooperation $c(z)$ is equal to 0 in this interval (i.e., nodes never cooperate). Hence the differential equation (5.13) becomes $\frac{\partial f}{\partial z} = 0$. The solution is then $f(z) = f(0)$.

2. $\theta \leqslant z < \theta + \gamma$: The probability of cooperation $c(z-\gamma)$ is equal to 0 in this interval and the differential equation is: $\lambda \frac{\partial f}{\partial z} + \eta c_0 f(z) = 0$. Considering the boundary condition $f(z = \theta)$, the solution in this interval is: $f(z) = f(0) e^{\frac{-\eta c_0}{\lambda}(z-\theta)}$.

3. $\theta + \gamma \leqslant z$: For these values of z, the differential equation (5.13) is a non-autonomous differential equation. We iteratively solve this differential equation by solving a series of autonomous differential equations in the interval $[0, \gamma]$. As illustrated in Fig. 5.3, we define m functions f_m, $m = 1, 2, 3, \cdots, \infty$, over the interval $[0, \gamma]$. For each interval, we obtain an autonomous differential equation as follows:

$$\begin{cases} \frac{\partial f_m(x)}{\partial x} + \frac{\eta c_0}{\lambda} f_m(x) - \frac{\eta c_0 (1-q)}{\lambda} f_{m-1}(x) = 0 \\ \\ f_m(0) = f_{m-1}(\gamma) \qquad m = 1, 2, 3, \cdots, \infty \end{cases} \tag{5.14}$$

In order to ensure the continuity of $f(z)$, we need to take into account the solution of

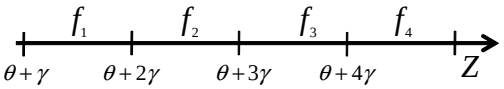


Figure 5.3: The definition of f_m over $[\theta + m\gamma, \theta + (m+1)\gamma]$ intervals.

$f(z)$ in the interval $[\theta, \theta+\gamma]$. Hence, we know that $f_0(\gamma)$ must be equal to $f(0)e^{\frac{-\eta c_0}{\lambda}(\gamma)}$. For $m = 1, 2, 3, \cdots, \infty$, the solution of the above system of iterative equations is:

$$f_m(x) = e^{\frac{-\eta c_0 x}{\lambda}} \left(f_{m-1}(\gamma) + \frac{\eta c_0 (1-q)}{\lambda} \int_0^x e^{\frac{\eta c_0 y}{\lambda}} f_{m-1}(y) dy \right) \tag{5.15}$$

Finally, we obtain the values of $f(z)$ by calculating $f_m(z - (\theta + m\gamma))$ for every interval $[\theta + m\gamma, \theta + (m+1)\gamma]$.

Let us define $\alpha = \eta c_0 / \lambda$ for simplicity. After some simplifications on Equation (5.15), we obtain:

$$f(z) = \begin{cases} f(0) & z < \theta \\ f(0)e^{-\alpha(z-\theta)} & \theta \leqslant z < \theta + \gamma \\ f(0)e^{-\alpha(z-\theta)} g(z) & \theta + \gamma \leqslant z \end{cases} \tag{5.16}$$

where $g(z)$ is a polynomial function as follows:

$$g(z) = \sum_{k=0}^{m} \frac{\alpha^k}{k!} e^{k\alpha\gamma} (1-q)^k (z - k\gamma - \theta)^k \tag{5.17}$$

for $\theta + m\gamma \leq z < \theta + (m+1)\gamma$. Recall that $f(0)$ can be calculated using the boundary condition presented in Equation (5.13). After some simplification, we obtain:

$$f(0) = \frac{1}{\theta + \frac{1 - e^{-\alpha\gamma}}{\alpha} + I} \tag{5.18}$$

where $I = \sum_{m=1}^{\infty} \sum_{k=0}^{m} e^{\alpha(\theta + k\gamma)} \frac{(1-q)^k \alpha^k}{k!} \int_{\theta+m\gamma}^{\theta+(m+1)\gamma} e^{-\alpha z} (z - \theta - k\gamma)^k dz$.

From the above equation, we observe that we need to calculate probability q (at least one node cooperates at the meeting point) in order to obtain $f(z)$. To do so, we must compute probability h_n of meeting n nodes and the probability of cooperation of a node $\bar{c}$ as shown in Equation (5.10).

5.5.1 Derivation of Probability q

Assume that the average number of nodes in a meeting point is $\bar{N}$. Usually, the probability of having n nodes in a meeting point follows a long tail distribution. We consider a Geometric distribution with parameter w for the probability of meeting n nodes. By definition of a Geometric distribution, the average number of nodes at a meeting point is $\bar{N} = \frac{w}{1-w}$. The probability of meeting n nodes is then:

$$h_n = w^n (1 - w) \tag{5.19}$$

Hence, the $\mathcal{Z}$-transform of h_n is:

$$H(\mathbf{z}) = \sum_{n \geqslant 0} \mathbf{z}^n w^n (1-w) = \frac{\mathbf{z}(1-w)}{1-\mathbf{z}w} \tag{5.20}$$

We also need to compute the average probability of cooperation $\bar{c}$. Using Equation (5.11), we obtain:

$$\begin{aligned} \bar{c} &= f(0)c_0 \left(\frac{1-e^{-\alpha\gamma}}{\alpha} + \int_{\theta+\gamma}^{\infty} e^{-\alpha(z-\theta)} g(y) dy \right) \\ &= f(0)c_0 \left(\frac{1-e^{-\alpha\gamma}}{\alpha} + I \right) \end{aligned} \tag{5.21}$$

Finally, q is obtained by computing $H(1-\bar{c})$, which is the value between 0 and 1 that satisfies the following equation:

$$\frac{c_0}{q} = \frac{1}{w} - (1-c_0) + \frac{\theta(1-w)}{w(\frac{1-e^{-\alpha\gamma}}{\alpha} + I)} \tag{5.22}$$

Our results in Equation (5.16) show that the probability density function $f(z)$ will first be uniform in the interval $[0, \theta]$. Then, on the small interval $[\theta, \theta+\gamma]$, it will decrease exponentially. For the other values of z, the probability $f(z)$ decreases according to an exponential distribution multiplied by a polynomial $g(z)$. Intuitively, it means that nodes will be evenly distributed below the threshold θ and for the other values of $z > \theta$ will have a long tail distribution.

With respect to probability q, we observe that it not only depends on cooperation parameters such as c_0 and θ, but also depends on the rate of encounters η, the average number of nodes in an encounter $\bar{N}$, and on the cost of changing pseudonym γ.

Example with $\gamma = 0$ and $c_0 = 1$

Assume that the cost of changing pseudonym $\gamma = 0$ and that $c_0 = 1$. The probability distribution function $f(z)$ can be rewritten as

$$f(z) = \begin{cases} f(0) & z < \theta \\ \\ f(0)e^{\frac{-\eta q}{\lambda}(z-\theta)} & \theta \leqslant z \end{cases} \tag{5.23}$$

where $f(0) = \frac{1}{\theta + \frac{\lambda}{\eta q}}$. We compute q by using Equation (5.10). Considering our threshold cooperation function $c(z)$, $\bar{c}(t)$ is:

$$\bar{c} = \int_{\theta}^{\infty} f(0) e^{-\frac{\eta q}{\lambda}(z-\theta)} dz = \frac{\lambda}{\lambda + \eta q \theta} \tag{5.24}$$

Finally q can be calculated by replacing $\bar{c}$ in Equation (5.10):

$$q = \frac{\lambda - \sqrt{\lambda(4\eta\theta w(1-w) + \lambda)}}{2(-1+w)\eta\theta} \tag{5.25}$$

We observe that in this simple example, the probability density function $f(z)$ is first uniform for $z < \theta$ and then decreases according to an exponential distribution.

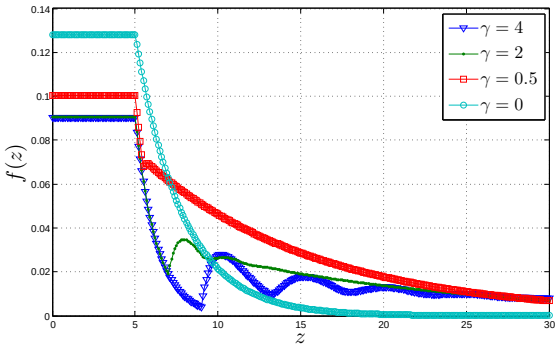


Figure 5.4: Probability distribution function $f(z)$ for different values of γ.

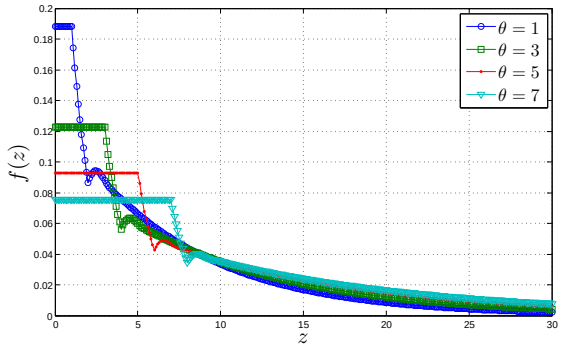


Figure 5.5: Probability distribution function $f(z)$ for different values of θ.

5.5.2 Numerical Evaluation

In this section, we evaluate numerically the analytical results of the previous section. In particular, we study how the system parameters affect the distribution of the age of pseudonyms, $f(z)$. Unless otherwise stated, we use the following values for the system parameters: $\bar{N} = 0.5$, $\theta = 5$, $\gamma = 1$, $\lambda = 1$, $\eta = 0.75$, and $c_0 = 1$.

As shown in Equation (5.16), the probability density function $f(z)$ has three different behaviors: it is first constant with value $f(0)$, then decreases exponentially with parameter $-\alpha$ and finally decreases according to an exponential multiplied by a polynomial which is different for every interval of size γ. We observe in Fig. 5.4 the three behaviors of $f(z)$ for different values of γ. For example, with $\gamma = 4$, we have $f(z) = f(0) = 0.09$ over the interval $[0, \theta]$. Then, $f(z)$ exponentially decreases until $\theta + \gamma = 9$. Finally, we observe that $f(z)$ oscillates because of the polynomial function (5.17) which is different for every interval of size γ. As z increases, the oscillation is attenuated because the exponential term dominates the polynomial function. Intuitively, the oscillation is caused by the jump process Γ_2: nodes with age of pseudonym belonging to $[z - \gamma, z]$ fail to coordinate and their age of pseudonym is thus increased by γ. In Fig. 5.4, we also observe the effect of different values of γ on the distribution $f(z)$. As γ decreases, the oscillations become less noticeable because the jump process Γ_2 affects fewer nodes (since the interval $[z - \gamma, z]$ becomes smaller). Moreover, in the case of $\gamma = 0$, we notice that there is no oscillation because Γ_2 does not affect any node. Note

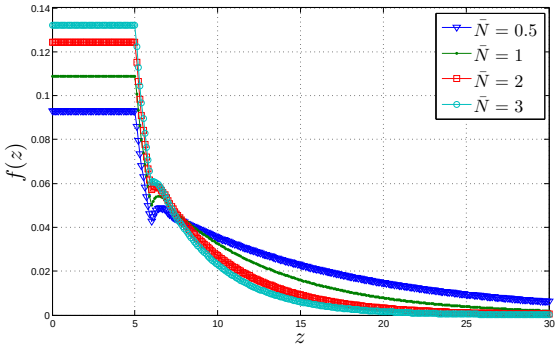


Figure 5.6: Probability distribution function $f(z)$ for different values of $\bar{N}$.

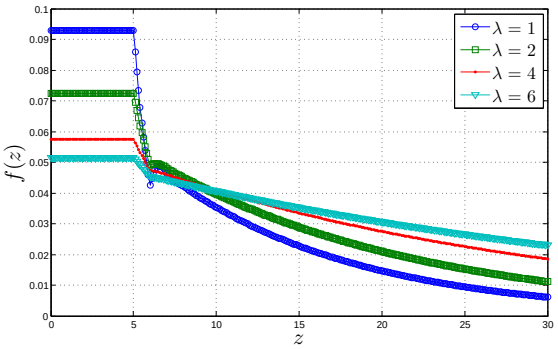


Figure 5.7: Probability distribution function $f(z)$ for different values of λ.

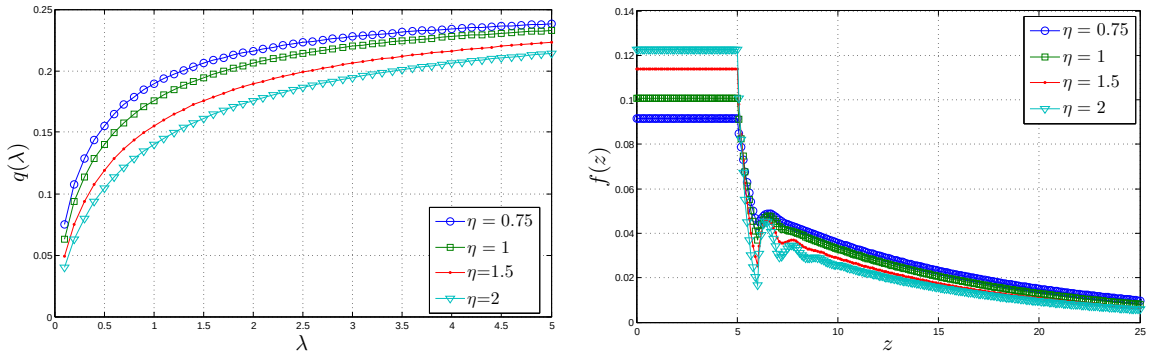
that when γ decreases, more nodes have an age of pseudonym smaller than the threshold θ.

Figure 5.5 shows the effect of different θ on $f(z)$. We observe that with larger values of θ, the number of nodes with age of pseudonym below θ increases. A system designer can thus fine tune θ to vary the population of nodes with age of pseudonym smaller than θ. As θ increases, we notice that the average value of z increases as well, meaning that more nodes have a high age of pseudonym because nodes are less cooperative.

Figure 5.6 illustrates the effect of the average number of nodes $\bar{N}$ in meetings on $f(z)$. When $\bar{N}$ increases, the probability q to find a cooperative node increases and consequently the number of nodes with age of pseudonym below the threshold θ increases as well, meaning that in average the age of pseudonym is smaller.

Figure 5.7 illustrates the effect of the aging rate λ on $f(z)$. We observe that with a high λ (i.e., pseudonyms age faster), fewer nodes have an age of pseudonym below θ compared to lower values of λ.

Finally, we evaluate the influence of the rate of meetings η on $f(z)$ in Fig. 5.8. First, we focus on the probability q of encountering at least one cooperative node in Fig. 5.8 (a). As λ increases, nodes age faster and we observe that their probability of cooperation increases logarithmically. When the rate η increases, we observe that the probability of cooperation q decreases for any value of λ: the reason is that for larger values of η, both jump processes Γ_1 and Γ_2 occur more frequently. Because Γ_1 dominates Γ_2 (as it affects more nodes), a larger fraction of nodes will have an age of pseudonym below θ (Fig. 5.8 (b)). For this reason, for a



(a) (b)

Figure 5.8: Influence of the rate of meetings η: (a) Probability q that at least one node in a meeting cooperates. (b) Probability distribution function $f(z)$.

high η, fewer nodes cooperate, and q decreases.

The above results can help a system designer find the conditions for the emergence of location privacy. More specifically, the system designer can fine tune parameters such as θ, γ and λ in order to control the number of nodes with large age of pseudonym.

5.5.3 Validation with Simulations

In order to verify the relevance of our model, we compare our numerical evaluations with simulation results.

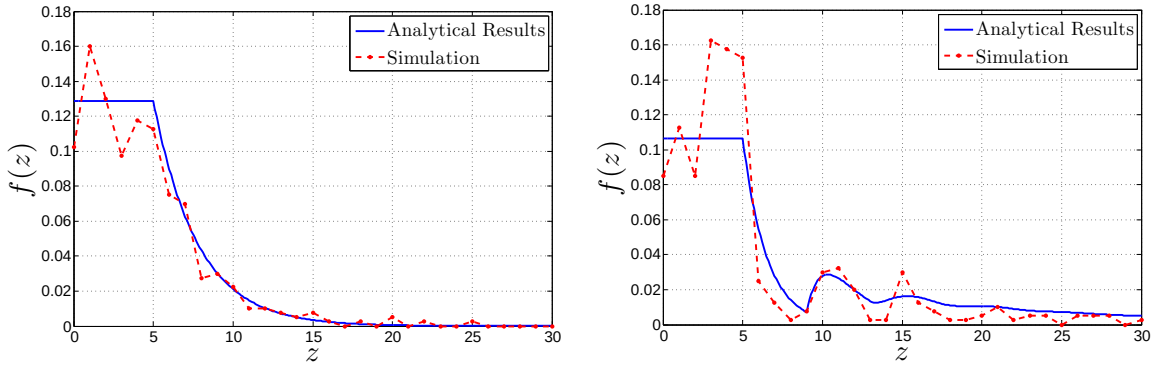
Simulation Setup

We consider a set of $N = 1000$ mobile nodes moving according to a random walk model [54]. The plane is composed of a grid of 10km×10km, where each step is of one meter. At every intersection, mobile nodes move from their current location to a new location by randomly choosing a direction. We consider that mobile nodes move with a constant speed. Directions are chosen out of $[0, 2\pi]$ with granularity $\pi/2$.

We consider that nodes are neighbors (i.e., in communication range) if they are within a fixed perimeter. We consider a communication range of 100m. Whenever a node has at least one neighbor, it must decide whether to cooperate or defect based on its value $Z_i(t)$ and the threshold cooperation function $c(z)$. After each iteration of the simulation, we compare the average of the current probability density function $f(z)$ to the average of $f(z)$ obtained in the 50 previous iterations. The simulation stops if the difference is smaller than 0.005, and otherwise runs at least for 200 iterations.

Simulation Results

Figure 5.9 compares $f(z)$ obtained with the numerical evaluation to the one obtained by simulation with two different values of $\gamma = 0$ and $\gamma = 4$. We consider the same value of η and $\bar{N}$ for the analytical and simulation results. The distribution of age obtained from the model shows a pretty good match with the distribution obtained with simulations. This means that



(a) (b)

Figure 5.9: Validation of numerical results: (a) $\gamma = 0$. (b) $\gamma = 4$.

our modeling assumptions succeeded in capturing the collective behavior of nodes changing their pseudonyms in mobile networks.

5.6 Summary

We developed a framework to analytically evaluate the age of pseudonyms. Our framework captures the mobility and interactions between nodes. With this model, we obtained critical conditions for the emergence of location privacy. In particular, we evaluated the importance of the probability of cooperation of the nodes (θ and c_0), their mobility (η and $\bar{N}$), the cost of pseudonyms γ, and the aging rate λ.

Some results are intuitive: the number of nodes with age of pseudonyms lower than the cooperation threshold increases with a large number of devices in meetings $\bar{N}$, with a large rate of encounters η, with a small cooperation threshold θ or with small aging rate λ. A less intuitive result is the influence of the cost γ on the distribution of the age of pseudonyms. We observe that a large cost γ creates oscillations in the distribution of the age of pseudonyms, indicating that it becomes harder to bound the age of pseudonyms. It is particularly interesting to note that our results enable to analytically relate different cooperation thresholds θ achieved in practice and the age of pseudonyms, e.g., the threshold at Nash equilibrium in non-cooperative environments (Chapter 4). The results of the model match well with simulations, meaning that our modeling assumptions succeeded in capturing the collective behavior of nodes changing their pseudonyms in mobile networks.

Publication: [94]

Part III

Privacy in Pervasive Social Networks

Chapter 6

Privacy of Communities

> Ce devant quoi une société se prosterne
> nous dit ce qu'elle est.
>
> Philipe Murray

6.1 Introduction

In addition to wireless interfaces that communicate with the infrastructure (e.g., cellular or WLAN), mobile devices are increasingly equipped with ad hoc communications capabilities (e.g., WiFi in ad hoc mode or Bluetooth). These wireless interfaces enable peer-to-peer interactions with nearby users and may fuel the development of context-aware applications: mobile devices can sense their environment and share data with other nearby devices.

Such peer-to-peer communications offer new ways to enhance social interactions between humans [15]. Wireless communications inherently depend on the geographic proximity of mobile devices and thus provide a geographic extension of online social networks. Both industry and academia are pushing towards the development of such context-aware applications. For example, new applications enable users to share information in real-time for local-area social networking [3, 4, 10, 176, 32], dating [1, 2, 135], gaming [156], or personal safety [177]. In most applications, users enter personal information in their smart phones, which is then shared with other phones nearby.

In real-life social interactions, humans naturally form groups called *communities* based on interests, proximity, or social relations [169]. Some communities are created dynamically for short periods of time, whereas other are persistent. In this work, we consider that users of mobile devices are grouped into user-defined *communities* based on user interests, social relation, or proximity. Such communities can be created dynamically when users are in proximity, or on online social platforms (e.g., groups in Facebook).

We study a communication primitive that enables users to share information with specific communities using ad hoc wireless communications. Users can subscribe to communities of interest to automatically receive messages sent to their community by other members. Friends or strangers can thus dynamically exchange *relevant* information when they are in proximity.

For privacy concerns, some users of online web services have an aversion towards sharing their contextual information with infrastructure-based services. Sharing personal information locally in a peer-to-peer fashion mitigates this problem, yet leaks personal information to

eavesdroppers. In particular, as the content of wireless communications can be eavesdropped, *data privacy* is at risk. Similarly, wireless broadcasting of messages leaks location information meaning that *location privacy* is in jeopardy [26]. In addition, the use of communities brings forth a new set of *community privacy* threats. Users from the same community could be linked, thus exposing their social relations [83]. Similarly, users' memberships in communities reveals their interests. Community privacy also affects location and data privacy as it might be easier to track an individual or to breach data privacy if it is known to belong to a specific community.

The upcoming generation of mobile computing requires mechanisms to identify groups of users while protecting their privacy. To do so, existing works suggest to *anonymously authenticate* other community members, thus protecting location and community privacy, and then derive a secret to protect data privacy. Several anonymous authentication techniques could be used, such as anonymous credentials [47, 48], group/ring signatures [35, 60, 188], private set intersections [91], affiliation-hiding envelopes [130] and secret handshakes [21, 127]. These methods thwart many privacy threats, but, besides secret handshakes, none of them achieves the desired privacy protection. In some cases, community membership is not hidden [47, 48, 35, 60, 188], others do not provide a suitable solution to our problem [91], and yet in others, anyone can communicate with community members [130]. Secret handshakes [129] appear to be the best solution to anonymously authenticate community members.

In this work, we tackle the main drawback of secret handshakes: cost. Secret handshakes are interactive protocols that require exchange of several messages and execution of several cryptographic operations to authenticate other group members. Although secret handshake schemes work well for Internet-based scenarios, in pervasive social networks, users will encounter a large number of other devices for a *short period of time*, e.g., dozens of encounters per minute in a city center, thus *frequently* invoking group-identification mechanisms. In particular, as interactions are short-lived, these group-identification mechanisms have to operate fast. In addition, battery constraints of mobile devices hinder the use of *computation* and *communication* intensive operations.

Most secret handshake schemes are linkable: users use persistent pseudonyms to initiate a secret handshake. Previous work on secret handshakes suggested the use of one-time pseudonyms to obtain unlinkability. However, such single-use pseudonyms may require too much storage. Some solutions provide cryptographic unlinkability using zero-knowledge proofs [129], or Key-Private Group Key Management Schemes [128]. However, such approaches entail a high cost. Other solutions include heuristic unlinkability approaches where users achieve unlinkability by rotating through a small set of pseudonyms, by setting strict time limits on the use of each pseudonym, by using k-anonymous techniques [224], or by associating different pseudonyms with different locations or aspects of a user's activity [127]. In this work, we consider these heuristic solutions as a starting point. Yet, the achievable privacy is unclear and other schemes could be derived.

In view of these observations, our contributions are:

1. We define the notion of community privacy that captures the privacy threat induced by communities. We provide a framework based on challenge-response protocols to formally evaluate the interactions between mobile users using community pseudonyms and an adversary aiming at breaking community privacy.

2. We propose new schemes to *efficiently* identify communities of users in a *privacy-preserving fashion*. These schemes complement existing schemes used in secret hand-

shakes by focusing on the specific requirements of mobile wireless networks, i.e., low cost and unlinkability. These schemes provide unlinkability by using short-term community pseudonyms and authentication based on symmetric cryptography. We show how asymmetric cryptography can be used only for specific operations in order to reduce cost.

3. We evaluate the privacy and cost of proposed and existing schemes by using the community privacy formalization. Our results shed light on the unlinkability achieved with short-term pseudonyms and the consequences on the privacy provided by secret handshakes.

6.2 Related Work

We discuss the properties of cryptographic techniques that could protect the privacy of communities.

Anonymous Authentication As previously discussed in Chapter 2, there are several mechanisms for anonymously authenticating users such as *anonymous credentials* [48], *group signatures* [35, 60], and *ring signatures* [188]. Although these techniques protect the authenticator's privacy, they expose group membership and hence leak community membership to any eavesdropper.

Private Broadcast Encryption and Affiliation-Hiding Envelopes There are several mechanisms for privately sharing information with members of a group. *Key-private encryption* (also called private broadcast encryption) [23] allows users to send encrypted messages to any member of a group such that the ciphertext hides the identity of the intended recipients from anyone except authorized recipients. Affiliation-hiding envelopes [130] address the same issue but with regards to interactive protocols: a receiver can read a message only if it satisfies the policy of the sender. *Hidden credentials* [38] are a special form of affiliation-hiding envelope schemes that also hide receivers' affiliations from all parties (including senders), and hide senders' policies from non-authorized receivers. Similar to affiliation-hiding envelopes, *oblivious signature-based envelopes* (OSBEs) [146] enable a sender to send an encrypted message to a receiver such that the receiver can open the message only if it is authorized and also guarantees that the sender cannot tell whether the receiver read the message.

Although such schemes could be used to privately share information with other community members, anyone (not even owning any credentials) can send messages to a given group. This property is undesirable in our setting as it makes it difficult to prevent spam. Only members of a community should be authorized to communicate with other members.

Private Set Intersection *Private set intersection* (PSI) protocols [68, 91] allow users to discover the intersection of two input sets without disclosing any information about further elements. PSI protocols could be used to detect other group members in proximity: two parties input their community membership as a list of community identifiers, execute the PSI interactive protocol and obtain the communities in common. One problem with PSI protocols is that memberships are not certified, so users can include identifiers of communities they do not belong to. Authorized PSI protocols [52] solve this problem by strengthening the requirements: the computed intersection of inputs must contain certified elements only.

Another problem with PSI protocols appears when multiple parties are involved: PSI protocols compute the intersection of the input sets of all parties. In order to obtain pairwise community discovery, PSI protocols thus require an exponential number of operations. For these reasons, PSI protocols are not suitable to solving the problem considered in this Chapter.

Nearby Friend Algorithms Previous work investigated algorithms to alert users of nearby friends [12]. Mobile devices determine their position and share it with friends using ad hoc or centralized communications. One disadvantage of this approach is that friends always learn each other's location, regardless whether they are actually nearby. To solve this issue, Zhong *et al.* propose a privacy-preserving buddy-tracking application where users can learn their friends' locations only if their friends are actually nearby [227].

These protocols assume that users know their friends and regularly communicate in order to compare their locations privately. Such technique would not scale to the setting of community detection because communities are composed of numerous users that would have to regularly communicate with each other. In this work, we investigate a related problem in which users can discover others that share common interests in a spontaneous fashion (i.e., when in vicinity). The nearby friend problem could actually be solved using communities: users could create a community of friends, and could detect them by checking for the proximity of members of that community.

Secret Handshakes and Affiliation-Hiding Authenticated Key Exchange Affiliation-hiding authentication schemes were introduced as secret handshakes [21]: they allow two members of the same group to authenticate each other in a way that hides their affiliation from all others. Such property can be used to anonymously authenticate members of communities. As secret handshake schemes are only entity authentication schemes, new schemes were proposed to provide Affiliation-Hiding Authenticated Key Exchange (AH-AKE) [127]. These schemes have the advantage of providing higher security guarantees as they output an authenticated session key after the handshake. The state-of-the-art AH-AKE scheme [127] guarantees perfect forward secrecy, robustness against man-in-the-middle attacks and was recently extended to support linear complexity for the discovery of multiple communities [154]. It achieves minimal costs of 3 communication rounds and two (multi) exponentiations per party and community. There is still no solution for group (multi-party) secret handshakes in this setting.

We argue that the main cost factor in secret handshakes schemes is due to asymmetric cryptography. One way to reduce cost is to rely instead on symmetric cryptography. Schemes based on symmetric cryptography make it more difficult to support revocation and could be linkable [213]. Hence, in order to provide revocation support, we restrict the use of asymmetric cryptography to certain situations. In order to achieve unlinkability, we rely on techniques similar to those proposed in the unlinkable secret handshakes schemes.

6.3 System Model

We introduce the assumptions made throughout the Chapter.

6.3.1 Network Model

Like in previous Chapters, we consider a network composed of personal mobile devices equipped with wireless interfaces that can communicate with each other in a peer-to-peer fashion upon coming in radio range (e.g., WiFi in ad hoc mode or Bluetooth). These peer-to-peer wireless communications complement communications with an infrastructure such as cellular or WLAN. We define $\mathcal{U} = \{U_1, ..., U_n\}$ as the set of users in the network, where n is the total number of users. For simplicity, we consider that each user owns a single communication device. We study a discrete time system with initial time $t = 0$ and consider the location of users at each discrete time instant t.

Mobile devices can identify each other with their link-layer identifier (i.e., MAC address), network-layer identifier (i.e., IP address) or application-layer identifiers (i.e., usernames or cookies). Let $ID_k(t)$ refer to all identifiers of user U_k at time t; we call $ID_k(t)$ a *user pseudonym*. As commonly assumed, user pseudonyms may change over time [26, 108].

We assume the presence of a trusted central authority (CA) run by an independent third party. Among other things, the CA loads authentication material in mobile devices and revokes misbehaving users. Users are preloaded by the CA with authentication credentials (i.e., asymmetric keys) and can encrypt or sign their messages by using asymmetric cryptography. The CA may not always be available to mobile users because of communication costs, limited network coverage or scalability issues. In practice, the CA can be a generic online platform (e.g., Verisign, Facebook) or a cellular operator.

6.3.2 Community Model

We consider that users can form *communities* [169, 177] structured around groups of people sharing *common interests*, such as professions, locations and social relations. Users can be represented as nodes in a graph connected by edges capturing shared interests. A community can then be defined as the union of several complete subgraphs that share many of their nodes [169]. Communities can be centrally formed, for example, online in the CA like groups on online social networks, or created in a distributed fashion by users in the network. Note that users can belong to several communities.

Let us define $\mathcal{C} = \{C_1, C_2, ..., C_m\}$ as the set of communities where m is the number of communities in the network. Each C_i is composed of a set of users $C_i = \{U_k\}$ and has private credentials SK_i (i.e., a *secret*) known to all community members. At time t, a community is identified by one or multiple community pseudonyms $P_{C_i}(t) = \{p_{i,j}\}$ where j is the jth pseudonym. Unless otherwise stated, we focus for simplicity on a time period during which community pseudonyms do not change and write $P_{C_i}(t) = P_{C_i}$. The set of all community pseudonyms in the system is then $\mathcal{P} = \bigcup_i P_{C_i}$. *Community pseudonyms* are generated using a *community pseudonym scheme*. In Section 6.6, we discuss various techniques to generate community pseudonyms based on the community secret.

We consider that communities have different *profiles* depending on the common interests characterizing members, their mobility profiles and communication rate. For example, a community profile may be characterized by a set of points of interest (POIs), e.g., restaurants for a fine food community. We consider that each user belongs to a fixed number of communities n_c. In summary, a community is established by defining its members, its name, its secret and its pseudonym scheme.

6.3.3 Communication Model

In order to automatically detect the presence of other users, mobile devices periodically broadcast proximity beacons containing the users' identities and time:

$$U_k \rightarrow * : ID_k(t) \mid t \tag{6.1}$$

where $ID_k(t)$ is the user pseudonym of U_k at time t. We consider a contention-based beaconing mode similar to Ad Hoc mode of IEEE 802.11. When U_k receives a beacon from U_j, U_k can start interacting with U_j.

In addition to standard unicast interactions, we consider that mobile users can exchange information with groups of users (communities). When another user is in proximity, users broadcast messages to all communities they belong to. Nearby users can then automatically detect messages sent to communities they subscribed to. To do so, community packets broadcasted by user U_k to community C_i at time t have the following format:

$$U_k \rightarrow C_i : ID_k(t) \mid p_{i,j} \mid msg \tag{6.2}$$

where $p_{i,j}$ is the j-th community pseudonym of C_i at time t (i.e., the destination of the packet) and msg is the message that may be encrypted. Community pseudonyms $p_{i,j}$ are used by receivers to detect whether a message is sent to a community they belong to and decide whether to read/decrypt the message. If a user wants to reply to a community packet, it can use standard unicast communications:

$$U_j \rightarrow U_k : ID_j(t) \mid ID_k(t) \mid msg \tag{6.3}$$

We consider that communications are multi-hop to a certain maximum hop-count defined by the system. Hence, traditional routing algorithms can be used [42] and community members located multiple hops away from the sender can receive messages.

The battery of mobile devices is affected by the number of communications and the computation overhead. Hence, *energy efficiency* is crucial in the design of algorithms used in the scope of pervasive social networks. In the following, we will aim at designing community pseudonym schemes that rely on as few operations and messages as possible.

6.3.4 Application Example

Numerous applications could take advantage of community information and real-time localized messages to enhance their context awareness. Consider the following example of a privacy-conscious geo-social network application. A group of users interested in technology create a community on their preferred online social network. Each member of the community receives keying material that is uploaded on their mobile device. Mobile devices broadcast community pseudonyms so that members of the same community can automatically detect their proximity. The notion of community enables to easily identify messages of other friends.

Upon detecting the proximity of another community member, mobile devices can take different actions depending on user preferences. Typically, mobile devices log received messages and decide whether to interrupt users' current activities. They can also perform pre-defined actions such as automatically share articles about technology users recently read. Later, users can check their logs to learn articles that others liked. Unlike other sources of information online (such as an RSS feed), this system is influenced by the geographic proximity of users and thus provides novel insights.

6.4 Threat Model

Based on the broadcasted packets, an adversary $\mathcal{A}$ could jeopardize user privacy. In particular, $\mathcal{A}$ may detect the members of communities, obtain users' locations and extract the content of messages. In this Chapter, we address these threats and focus on designing mechanisms to protect the privacy of communities.

We consider both a *passive* and an *active* adversary. A passive adversary collects from the network the messages broadcasted by the devices and obtains communication *traces*. At best, the passive adversary is *global*: it has complete coverage and tracks users throughout the entire network. So far in the thesis, we exclusively considered passive adversaries. In this Chapter, we extend our threat model to active adversaries because mobile devices contain keying information that may affect the privacy of other nodes as well. An active adversary can compromise mobile devices in order to extract their secrets. An adversary has thus incentive to actively compromise nodes to improve its tracking ability. In addition to compromising devices, $\mathcal{A}$ may also be a legitimate member of the network and know a priori the secret of several communities. We assume that $\mathcal{A}$ knowns $\mathcal{C}$ and define $P_{C_i}^{\mathcal{A}}$ the set of pseudonyms of C_i known by $\mathcal{A}$: for all communities C_i the adversary belongs to, we have $P_{C_i}^{\mathcal{A}} = P_{C_i}$.

We use the concept of *oracles* to abstract possible actions of the adversary and its strength. Based on collected mobility traces, the passive adversary $\mathcal{A}$ can attempt to break user privacy using *traffic analysis* attacks. $\mathcal{A}$ tracks users' locations and packets over time and tries to infer the relation between community pseudonyms and communities. For example, locations visited by a user or a group of users that are in proximity may leak their community membership to the adversary. We call s the number of packets collected by the passive adversary. Let $\mathcal{O}_P = \mathsf{TrafficAnalysis}(s)$ be a passive oracle that captures the traffic analysis attack of a passive adversary. The adversary inputs the s collected messages to $\mathcal{O}_P$ that outputs a mapping between community pseudonyms and specific communities.

In addition to the capabilities of a passive adversary, we consider an *active* adversary that can compromise devices using various means. Active attacks enable the adversary to run selective surveillance strategies in order to ascertain community membership information. We consider four oracles to model the attacks of an active adversary interacting with a mobile device D.

- $\mathsf{Query}(D, \mathcal{C})$: $\mathcal{A}$ sends packets to D about communities $\mathcal{C}$ by forging or replaying overheard packets. A unicast reply from D reveals the membership of the device to a community in $\mathcal{C}$.
- $\mathsf{Reveal}(D, \mathcal{C})$: $\mathcal{A}$ tries to obtain private credentials of all communities $\mathcal{C}$ of D by hacking into the hardware (e.g., reading the memory of D).
- $\mathsf{Join}(D, \mathcal{C})$: $\mathcal{A}$ tries to join communities $\mathcal{C}$ of D to obtain their private credentials, e.g., using social engineering attacks.
- $\mathsf{Create}(D, \mathcal{C})$: $\mathcal{A}$ creates several communities $\mathcal{C}$ and invites D to join them. With this knowledge of community membership of D, $\mathcal{A}$ may identify pseudonyms of other communities D belongs to.

We simulate the set of oracles to which the active adversary has access by considering an active Oracle $\mathcal{O}_A \subset \{\mathsf{Q}, \mathsf{R}, \mathsf{J}, \mathsf{C}\}$, where Q, R, J and C represent respectively the oracles Query,

Reveal, Join and Create. The adversary inputs devices D to $\mathcal{O}_A$ that then outputs a mapping between community pseudonyms and specific communities. $\mathcal{A}$ thus learns community memberships of input devices D and obtain all the corresponding community pseudonyms.

6.5 Protecting Privacy

In this section, we discuss different privacy-preserving mechanisms proposed to protect privacy of users of pervasive social networks. Recall that packets exchanged between users contain three parts: a user pseudonym $ID_k(t)$, a community pseudonym $p_{i,j}$ and a message msg. We first discuss how to provide unlinkability of user pseudonyms, how to protect data privacy of messages and, finally, address the unlinkability of community pseudonyms.

6.5.1 Location Privacy

In order to avoid traceability, users can change their pseudonyms $ID_k(t)$ over time [26, 108]. As a pseudonym changed by an isolated node can be trivially guessed by an external party, pseudonym changes should be coordinated among mobile nodes in regions called *mix zones* [26]. To protect against the spatial correlation of location traces, mix zones can also conceal the trajectory of mobile nodes to the external adversary by using silent/encrypted periods [122, 145], or (iii) regions where the adversary has no coverage [43].

The effectiveness of a mix zone, in terms of the location privacy it provides, depends on the adversary's ability to relate mobile nodes that enter and exit the mix zone [26]. Previous works show that mix zones are more effective when used by a large number of devices that have unpredictable mobility [28, 122]. The main drawback of mix zones is that they induce a cost for mobile users in terms of quality-of-service: for example, in a *silent mix zone* [123], mobile nodes cannot communicate and access services.

In this work, we consider that mobile users coordinate their pseudonym changes as defined in the dynamic game of Chapter 4 and thus effectively protect their location privacy at a low cost.

6.5.2 Data Privacy

In some cases, users may broadcast private messages to other members of the same community, and in others they may share their messages with everyone. Without lack of generality, we consider two types of communities: *public* and *private* communities. In public communities, any user can join the community and send messages to all members. Such messages can thus be read by anyone. In contrast, only members of a private community can read the content of packets sent to that community. Hence, users must be authorized to join a private community. Upon authorization, each private community member receives appropriate credentials to protect data privacy. In the following, we focus on private communities as they can be generalized to public communities by revealing private credentials.

Several mechanisms can be used to encrypt communications. A simple solution is to rely on the central authority to set up shared secrets: upon registration in the network, every user is given by the central authority a symmetric key SK_i for each community C_i it belongs to. As the secret key is common to every user in a community, users can use it to encrypt their packets and rely on community pseudonyms to detect messages sent to

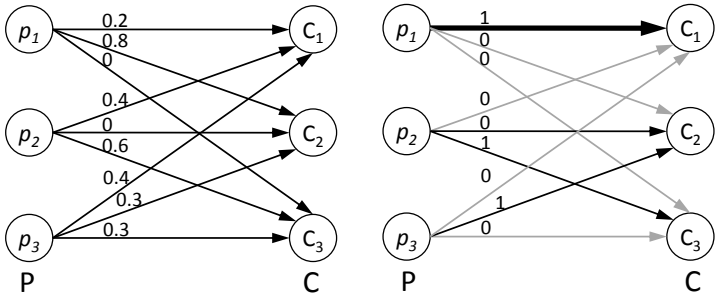


Figure 6.1: Illustration of result of attack by passive and active adversaries. (a) Weighted bipartite graph $\mathbf{G}$. (b) Weighted bipartite graph $\mathbf{G}'$ resulting from the combination of $\mathbf{G}$ and $\mathbf{G_A}$. The edge between p_1 and C_1 (in bold) belongs to $\mathbf{G_A}$ and is obtained from an interaction with Oracle O_A. Assuming that each community uses a single pseudonym, the adversary can rely on $\mathbf{G_A}$ to update the weights of other edges in $\mathbf{G}$. In this example, the adversary learns all mappings between community pseudonyms and communities.

a community they belong to. Another approach consists in deriving shared secrets in a distributed fashion [17, 39, 61, 87, 165, 208].

In this work, we consider that mobile users obtain certificates from the central authority and can derive shared secrets in a distributed or centralized fashion. Users also receive a symmetric key for each community they belong to.

6.5.3 Community Privacy

In this section, we discuss the properties that community pseudonyms should satisfy in order to preserve community privacy while enabling users to verify whether a message is destined for a community they belong to. To do so, we formalize the study of community privacy, define requirements for achieving community privacy and propose a framework to study the effect of various pseudonym schemes on community privacy.

Security proofs usually formalize the actions of the adversary and then derive security properties to protect against such an adversary [20]. In this Chapter, we follow a similar methodology to establish whether a community pseudonym scheme is resistant to passive and active attacks. We make use of a challenge-response methodology to evaluate the ability of the adversary to break privacy properties. A challenger provides the adversary with a privacy challenge. The adversary then attempts to solve the challenge by using oracles as a source of information and then inferring the link between pseudonyms and communities. If the adversary succeeds, it breaks the privacy of the scheme.

Description of Passive Attacker

A passive adversary collects s packets and infers the relation between sniffed community pseudonyms and communities by using the oracle $\mathcal{O}_P$. This relation depends on the way community pseudonyms are used (i.e., the community pseudonym scheme) by mobile devices. The number of collected messages s is a system parameter indicating the strength of the passive adversary. The application of oracle $\mathcal{O}_P$ to s messages results in a weighted bipartite

graph $\mathbf{G}$ (Fig. 6.1 (a)) representing a partial mapping between community pseudonyms $\mathcal{P}$ and communities $\mathcal{C}$, i.e., the vertices are divided into the set of pseudonyms $\mathcal{P}$ and the set of communities $\mathcal{C}$. Every edge connects a vertex in $\mathcal{P}$ to one in $\mathcal{C}$ and is weighted by the probability (estimated by the adversary) of linking a specific pseudonym to a specific community. There are multiple ways to obtain such graphs and we discuss them in Section 6.7.

Description of Active Attacker

An active adversary can extract information about communities by interacting with the oracle $\mathcal{O}_A$. We call $W(D)$ the result of the application of the oracle $\mathcal{O}_A$ to device D. $W(D)$ is a mapping between communities and community pseudonyms (similar to $\mathbf{G}$). An *active attack* is a set of calls to $\mathcal{O}_A$ by $\mathcal{A}$: $\{W_1(D_1), ..., W_\ell(D_\ell)\}$ where ℓ is the number of interactions and a system parameter indicating the strength of $\mathcal{A}$. Information obtained from an active attack can be represented with a weighted bipartite graph $\mathbf{G_A}$ similar to $\mathbf{G}$. If there is an edge between a vertex in $\mathcal{P}$ to one in $\mathcal{C}$ in the graph $\mathbf{G_A}$, then its weight is either 1 or 0 as the information learned from an active attack is absolute. At each interaction, the adversary changes the device that is input to the Oracle. Note that two interactions $W_i(D_1), W_j(D_2), i \neq j$ may produce identical mappings. Similarly, an interaction may be unsuccessful and output an empty result, $W_i(D) = \emptyset$.

The set of community pseudonyms of C_i known to $\mathcal{A}$ $P_{C_i}^{\mathcal{A}}$ increases depending on the number of interactions of $\mathcal{A}$ with $\mathcal{O}_A$. In general, the outcome of the active attack $\mathbf{G_A}$ is directly related to the number of different devices D input to the Oracle (i.e., the number of devices under attack). In other words, the larger the graph $\mathbf{G_A}$ is, the more successful is the active attack.

An active adversary can combine information from $\mathbf{G}$ and $\mathbf{G_A}$ into a more accurate graph $\mathbf{G}'$. Based on information in $\mathbf{G_A}$, the adversary can update the probability (i.e., weight) of the edges from $\mathcal{P}$ to $\mathcal{C}$. This is illustrated in Fig 6.1 (b) in case a single pseudonym is assigned to each community.

We derive two properties that community pseudonym schemes must satisfy to provide community privacy: *community anonymity* (CAN) and *community unlinkability* (CUN).

Community Anonymity

Definition 4. *For any community C_i, there is* community anonymity *at time t if and only if for all pseudonyms $p_{i,j}$ of C_i, only members of C_i are able to deterministically verify that $p_{i,j}$ is a valid pseudonym of C_i.*

Community anonymity (CAN) guarantees that users cannot be linked by third parties to the communities they belong to, i.e., community pseudonyms do not affect the anonymity of the community members.

To formalize community anonymity, we define a *challenge-response* game between the adversary and the challenger.

1. $\mathcal{A}$ collects s messages, interacts with $\mathcal{O}_P$ and obtains $\mathbf{G}$.
2. $\mathcal{A}$ interacts ℓ times with oracle $\mathcal{O}_A$ and obtains $\mathbf{G_A}$.
3. $\mathcal{A}$ queries challenger one community $C_0 \notin \mathbf{G_A}$ (i.e., for which the adversary does not have a mapping with probability 1).

4. Challenger selects at random $b \in \{0,1\}$. If $b = 0$, it sends $p \in C_0$ to $\mathcal{A}$, else it sends $p \notin C_0$.

5. $\mathcal{A}$ decides whether p belongs to C_0 and outputs b'.

Step 2 is not executed by a passive adversary. We define the advantage of the adversary for community anonymity as:

$$Adv_{s,\ell}^{CAN}(\mathcal{A}) = Pr(\mathcal{A} \text{ is correct}) - \frac{1}{2} \tag{6.4}$$

where ℓ is the length of the interaction of the adversary with oracle $\mathcal{O}_A$, and s is the number of messages collected by the adversary. The advantage measures the relation between the number of interactions ℓ, the number of messages collected s, and the probability of success of the adversary. If the adversary guesses uniformly at random ($Pr(\mathcal{A} \text{ is correct}) = 1/2$), then the advantage is $Adv_{s,\ell}^{CAN} = 0$. Similarly, if the adversary surely knows the answer ($Pr(\mathcal{A} \text{ is correct}) = 1$), then the advantage is $Adv_{s,\ell}^{CAN} = 1/2$. Thus, the advantage belongs to the interval $[0, 1/2]$. When there is no advantage provided by the community pseudonym scheme ($Adv_{s,\ell}^{CAN} = 0$), the adversary's best decision is to guess uniformly at random. For any advantage greater than 0 and smaller than one half, the adversary wins the game probabilistically.

Community Unlinkability

Definition 5. *For any community C_i, there is* community unlinkability *at time t if and only if for any two pseudonyms $p_{i,j}$ and $p_{i,k}$ of C_i, only members of C_i are able to deterministically verify that $p_{i,j}$ and $p_{i,k}$ belong to the same community.*

Community unlinkability (CUN) guarantees that users of the same community cannot be linked to each other or tracked by third parties, i.e., community pseudonyms do not affect traceability of community members.

We use again a *challenge-response* game.

1. $\mathcal{A}$ collects s messages, interacts with $\mathcal{O}_P$ and obtains $\mathbf{G}$.

2. $\mathcal{A}$ interacts ℓ times with oracle $\mathcal{O}_A$ and obtains $\mathbf{G_A}$.

3. $\mathcal{A}$ queries challenger one community $C_0 \notin \mathbf{G_A}$.

4. Challenger computes $C_1 = \mathcal{C} \setminus \{C_0 \cup \mathbf{G_A}\}$. Then, it selects at random $b \in \{0,1\}$ and $d \in \{0,1\}$ and sends $p \in C_b$ and $p' \in C_d$ to $\mathcal{A}$.

5. $\mathcal{A}$ decides whether p and p' belong to the same community and outputs yes/no.

Step 2 is not executed by a passive adversary. We define the advantage of the adversary for community unlinkability as:

$$Adv_{s,\ell}^{CUN}(\mathcal{A}) = Pr(\mathcal{A} \text{ is correct}) - \frac{1}{2} \tag{6.5}$$

Implications

Based on the definitions of CAN and CUN, we obtain the following theorem:

Theorem 6. *Community unlinkability implies community anonymity, but community anonymity does not imply community unlinkability.*

Proof. We prove Theorem 1 by contradiction. Assume first that unlinkability does not imply anonymity: $CUN \nrightarrow CAN$. This implies that there are probabilistic algorithms that *do not* breach community unlinkability but that do breach community anonymity. In other words, there are probabilistic algorithms that cannot decide better than a random guess whether any two pseudonyms belong to the same community, but that can decide better than a random guess their respective communities.

Let one such algorithm for breaching community anonymity be $A_{CAN}(p_j, C_i)$ that outputs *yes* if $p_j \in C_i$ and *no* if $p_j \notin C_i$. We have for some communities C_i:

$$\sigma \quad = \quad Pr(A_{CAN}(p_j, C_i) \text{ is correct}) > \frac{1}{2}$$

Given A_{CAN}, we can construct a probabilistic algorithm, $A_{CUN}(p_j, p_k)$, for deciding whether any two community pseudonyms belong to the same community or not. We define $A_{CUN}(p_j, p_k, C_i)$ as an algorithm that decides whether p_j and p_k belong to community C_i. We have:

$$Pr(A_{CUN}(p_j, p_k) \text{ is correct}) = \sum_{i}^{m} Pr(C_i) \cdot Pr(A_{CUN}(p_j, p_k, C_i) \text{ is correct})$$

where $Pr(C_i)$ is the probability that the challenger chooses a target community C_i and m is the number of communities. Note that $\sum_{i=1}^{m} Pr(C_i) = 1$.

Consider that $A_{CUN}(p_j, p_k, C_i)$ is the following algorithm:

1. Given community pseudonyms p_j and p_k.
2. Call $A_{CAN}(p_j, C_i)$ and guess if $p_j \in C_i$.
3. Call $A_{CAN}(p_k, C_i)$ and guess if $p_k \in C_i$.
4. Output yes if the two guesses say yes, else output no.

The probability of success of $A_{CUN}(p_j, p_k, C_i)$ is:

$$\begin{aligned}
\mu &= Pr(p_j, p_k \in C_i) Pr(A_{CUN}(p_j, p_k, C_i) \text{ is correct} | p_j, p_k \in C_i) \\
&+ Pr(p_j \notin C_i, p_k \in C_i) Pr(A_{CUN}(p_j, p_k, C_i) \text{ is correct} | p_j \notin C_i, p_k \in C_i) \\
&+ Pr(p_j \in C_i, p_k \notin C_i) Pr(A_{CUN}(p_j, p_k, C_i) \text{ is correct} | p_j \in C_i, p_k \notin C_i) \\
&+ Pr(p_j, p_k \notin C_i) Pr(A_{CUN}(p_j, p_k, C_i) \text{ is correct} | p_j, p_k \notin C_i) \\
&= \frac{1}{4}(\sigma^2) + \frac{1}{4}(\sigma^2 + (1-\sigma)^2) + \frac{1}{4}(\sigma^2 + (1-\sigma)^2)) + \frac{1}{4}(\sigma^2 + 2\sigma(1-\sigma)) \\
&= \sigma^2 + \frac{(1-\sigma)^2}{2} + \frac{\sigma(1-\sigma)}{2}
\end{aligned}$$

where σ^2 means A_{CAN} guesses p_j, p_k correctly, $(1-\sigma)^2/2$ models the two cases where A_{CAN} does not guess p_j, p_k correctly but A_{CUN} is correct and $\sigma(1-\sigma)/2$ models the two cases where A_{CAN} guesses one of p_j, p_k correctly and A_{CUN} is correct.

We observe that when $\sigma = 0.5$, we have $\mu = 0.5$, when $\sigma > 0.5$, we have $\mu > 0.5$ and when $\sigma = 1$, we have $\mu = 1$. Hence, for any distribution $Pr(C_i)$, we obtain that A_{CUN} succeeeds with probability greater than a random guess. This contradicts our initial hypothesis that the community identifier scheme is CUN. Thus, there can be no such algorithm and $CUN \rightarrow CAN$.

Second, assume that community anonymity implies community unlinkability: $CAN \rightarrow CUN$. If there are no probabilistic algorithms that can decide the respective community of a community pseudonym better than a random guess, then this implies that there are no probabilistic algorithms that can decide whether any two community pseudonyms belong to the same community better than a random guess. Formally, as we have CAN, there exists no algorithm $A_{CAN}(p_i, C_j)$ that is correct with a probability greater than $1/2$ for all p_i not known to the adversary, and for all communities C_j.

Assume that there exists a set of parameters $x_j \in X$ that uniquely identifies the community C_j and there is a mapping $R : X \rightarrow C$ from X to the set of all communities C. Let $A'(p_i, x_j)$ be the algorithm that decides if the pseudonym p_i belongs to a community with parameters x_j. Let $A''(x_j, C_k)$ be the algorithm that links a parameter x_j with a community C_k. If $A''(x_j, C_k)$ success with probability no greater than a random guess, then even if $A'(p_i, x_j)$ succeeds with probability greater than a random guess, it does not break CAN. But $A'(p_i, x_j)$ actually breaks CUN. This contradicts our initial assumption that $CAN \rightarrow CUN$, and thus, $CAN \nrightarrow CUN$.

□

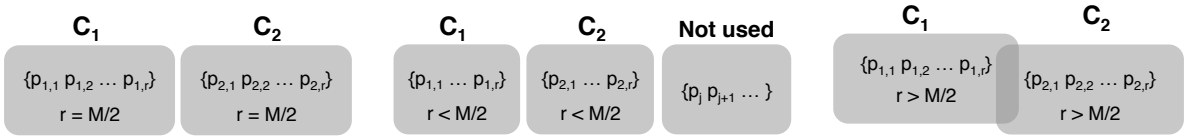
This theorem establishes the relation between CAN and CUN. It is consistent with existing results in the literature and validates our modeling assumptions. It shows that unlinkability is a stronger notion of community privacy in community identification protocols. Equivalently, applying the contrapositive property, Theorem 6 can be restated as: if CAN does not hold (i.e., there is an algorithm to break CAN), then CUN does not hold (i.e., there is also an algorithm to break CUN).

6.6 Community Pseudonym Schemes

In order to provide cost-efficient and private identification of user communities, we propose schemes of cryptographically generated community identifiers. The way these community pseudonyms are generated affects the verification cost and provided privacy.

In this section, we describe four classes of community pseudonym mechanisms and propose approaches to derive them. Then, we compare them with each other by evaluating their cost and security properties. We express the communication and computation cost needed to generate, store and transmit community pseudonyms with respect to the number of users in proximity and the number of communities per user. We consider that pseudonyms are defined over B bits. Thus, the space of possible pseudonyms has size $M = 2^B$.

We assume that a secret SK_i is shared among community members for all communities C_i and that when community pseudonyms change, the user pseudonym $ID_k(t)$ changes as well, and vice versa.



(a) (b) (c)

Figure 6.2: Community pseudonym schemes with multiple pseudonyms per community (assuming only two communities for illustration purposes). (a) Multiple pseudonyms over the entire domain. (b) Multiple pseudonyms over a shrunk domain. (c) Overlapping multiple pseudonyms, i.e., hints.

6.6.1 Single Pseudonym Schemes

In single pseudonym schemes, users use one *constant* pseudonym per community: $P_{C_i} = p_i$ where p_i is chosen uniformly at random in $\{0,1\}^M$ and $|P_{C_i}| = 1$. We consider two possible techniques to instantiate single pseudonym schemes.

- Each user uses a single pseudonym per community different from that of other users from the same community (similar to linkable secret handshake schemes).
- The same pseudonym is used by all users per community (similar to group signatures).

In practice, such schemes can be realized by using, for example, Hash functions. The pseudonym identifying a community can be the Hash of the secret of the community: $p_i = \mathcal{H}(SK_i)$, where $\mathcal{H}(.)$ is a Hash function such as SHA-2.

Such schemes have a low computation and communication overhead at the sender: the sender of a message has to select one pseudonym per community, i.e., $\mathcal{O}(m)$ lookups where m is the number of communities, and send one message per community it belongs to, i.e., $\mathcal{O}(m)$ communications. The receiver has to do more operations: for all community pseudonyms received from one neighbor, a receiver has to compare these community pseudonyms to all community pseudonyms of communities the receiver belongs to. The complexity of such *lookups* depends on the data structure used to store community pseudonyms (e.g., hashmaps or trees). The cost of lookups typically depends on memory requirements. As all messages are broadcasted, lookup operations are done for each device in communication range. Assuming hashmaps, the total number of lookups at the receiver is then: $\mathcal{O}(n_e m)$, where n_e is the number of nearby nodes (e stands for encounter).

One extension of single pseudonym schemes consists in relying on the concept of k-anonymity in order to achieve some unlinkability [224]. For each community pseudonyms of communities users belong to, users select $k-1$ other community pseudonyms that they send together with their messages. This technique increases the cost of the receiver as it increases the number of lookups to do. We obtain $\mathcal{O}(n_e k m)$ lookups. The higher the k, the higher the cost.

6.6.2 Multiple Pseudonym Schemes over the Entire Domain

With this class of schemes, each community C_i is identified by a *set* of pseudonyms known to all community members $P_{C_i} = \{p_{i,j}\}$ where j is the jth pseudonym of community C_i.

To send a packet to a community C_i, a user randomly selects a pseudonym from the set of pseudonyms P_{C_i}. Users receiving the packet determine whether the packet is sent by a member of their community by searching their local pseudonym repository for received community pseudonyms. If the receiver finds a match, it detects a message of interest and reads it. In practice, users must know all community pseudonyms of all communities they belong to.

The first type of multiple pseudonym scheme is defined over the entire domain of pseudonyms, i.e., over the M possible pseudonyms. It separates all possible pseudonyms across communities. In theory, we thus have: $|P_{C_i}| = M/m$ and $P_{C_i} \bigcap P_{C_j} = \emptyset, \forall i \neq j$. In practice, several mechanisms can be used to generate multiple pseudonyms over the entire domain.

- *Pre-Computed Schemes:* The CA randomly splits the set of all pseudonyms across communities as follows: every community C_i is assigned $\lfloor M/m \rfloor$ community pseudonyms.

- *Self-Generated Schemes:* Every user generates community pseudonyms on the fly. For each message sent to C_i, users choose a random number RND and using a Hash function compute a message authentication code: $p_{i,j}$ = RND || HMAC_{SK_i}(RND). For every message, a receiver verifies the HMAC with the key of all communities it belongs to. If the receiver can verify the hash, it knows that the message is sent to a specific community it belongs to.

Pre-computed schemes incur *storage costs*. For example, if pseudonyms are $B = 48$ bits and there are at most $m = 100000$ communities, every user must store $MB/8m \approx 16$ gigabytes per community it belongs to. The sender has to do $\mathcal{O}(m)$ lookups and broadcast $\mathcal{O}(m)$ messages. The advantage of pre-computed schemes is that the sender and receiver do not perform any computations. For every received community pseudonym, the receiver must do $\mathcal{O}(m)$ lookups. The large memory requirements ($\mathcal{O}(M)$) increases the cost of each lookup. Given n_e nearby nodes, the total number of lookups at the receiver is then: $\mathcal{O}(n_e m)$.

In self-generated schemes, the sender has computation costs, i.e., $\mathcal{O}(m)$ hashes and $\mathcal{O}(m)$ messages. The receiver also has a computation cost: users must hash all received messages with the secret key of all communities they belong to, i.e., $\mathcal{O}(n_e m^2)$ *Hash operations*. Then, each Hash result is compared to the received Hashes. The self-generated scheme thus introduces online computation costs.

A possible improvement that decreases overhead involves the use of Hash bins. Users can process their membership information before broadcasting it as follows: each user uses a Hash function $\mathcal{H}$ to map their memberships to an output into one of $\mathcal{B}$ bins. When two parties meet, the protocol must be run only between the membership credentials that were mapped by both parties to the same bin. Indeed, for every community C_i for which both users have membership credentials, both parties map these credentials to the same bin. As described in [154], this technique reduces the computation overhead to $\mathcal{O}(n_e m \log m)$, again at the expense of $\mathcal{O}(m)$ Hash operations at the sender.

A further improvement involves the use of Index-Hiding Message Encoding vectors (IHME) [154]. This technique uses polynomial interpolation to hide community membership in order to decrease the computation overhead to $\mathcal{O}(n_e m)$. Any of these two optimization mechanisms could be used to further reduce the cost of secret handshakes based on symmetric secrets.

6.6.3 Multiple Pseudonym Schemes over a Shrunk Domain

The next type of multiple pseudonym scheme shrinks the size of sets of community pseudonyms according to a shrink factor $h \in [0, 1]$. Hence, fewer community pseudonyms are assigned to each community (i.e., some pseudonyms are not assigned at all). The idea of shrinking pseudonym sets is to reduce costs and to make it more difficult for an adversary to relate community pseudonyms to communities (as some pseudonyms are not assigned at all). Formally, we have: $|P_{C_i}| = (M/m) \cdot h$ and $P_{C_i} \bigcap P_{C_j} = \emptyset, \forall i \neq j$. The smaller h is, the fewer community pseudonyms are used. Users then rotate in those sets to choose their community pseudonyms.

Two mechanisms can be used to generate multiple pseudonyms over a shrunk domain.

- *Pre-Computed Schemes:* The CA separates the set of pseudonyms across communities but only assigns a subset: every community C_i receives $\lfloor h \cdot M/m \rfloor$ community pseudonyms.

- *Self-Generated Schemes:* One technique is as follows: all users belonging to the same community generate a Hash chain in a synchronized fashion: $p_{i,1} = \mathcal{H}(SK_i)$, $p_{i,j+1} = \mathcal{H}(p_{i,j})$ for $1 < j < len$ where len is the length of the Hash chain.

In terms of the storage cost of a pre-computed scheme, it can be significantly lower compared to the scheme based on the entire domain. For example, users must now store only 160 megabytes of data if $h = 0.01$, $B = 48$ and $m = 100000$. The receiver cost is again $\mathcal{O}(n_e m)$ lookups.

With self-generated schemes, users can verify if a message is destined to them by checking whether the received community pseudonym belongs to one of the Hash chains they know. Hence, self-generated schemes enable verification by doing $\mathcal{O}(n_e m)$ lookups without online Hash operations.

6.6.4 Hints: Overlapping Multiple Pseudonyms

The third multiple pseudonym scheme relies on a method that we call *hints*. This method allows for an overlap in the set of pseudonyms used for each community. In other words, community pseudonyms can be used by more than one community. The idea of overlapping pseudonym sets is to create confusion for the adversary. We define the overlap factor $o \in [0, 1]$ as the fraction of community pseudonyms that may be shared by different communities. We use the term "hint" because a community pseudonym does not uniquely identify a community anymore but works rather as a hint to help receivers determine whether a messages is destined to them. We have: $P_{C_i} \bigcap P_{C_j} \neq 0$ for some $i \neq j$.

- *Pre-Computed Schemes:* The CA splits the set of all pseudonyms across communities but assigns some pseudonyms to multiple communities.

- *Self-Generated Schemes:* Hints can be implemented using the Hash method of self-generated schemes. In order to obtain some overlap, the output of the Hash is truncated to a smaller number of bits. As a consequence, several RND values will have the same Hash, thus creating collisions. The larger the truncation of the Hash is, the larger the overlap will be.

In addition to the cost of self-generated schemes over the entire domain, users will get some messages that are not destined to them because of the Hash collisions (i.e., false positives). If these messages are encrypted using the community secret, then users will have to unsuccessfully attempt decryption. The number of unsuccessful decryptions depends on the number of collisions and can be computed using the Birthday paradox. Hence, the larger the overlap is, the larger the cost of verification will be.

Table 6.1 summarizes the cost of the different schemes. Note that computations in our setting correspond to Hash operations, which is significantly lower than the cost of asymmetric cryptography operations. Similarly, lookup operations have a considerably lower cost than Hash operations.

We observe that it is possible to derive schemes that avoid online computations, such as single pseudonym schemes and pre-computed multiple pseudonym schemes. But, these schemes may suffer from drawbacks of trivial linkability or large storage costs. Other schemes, such as self-generated schemes, Hash bins and IHME overcome the problem of high storage costs by introducing online computations. Although Hash bins provide logarithmic complexity, they are communication-wise inefficient. The reason is that all bins must be transmitted even if users belong to only a few communities. IHME schemes are the most efficient but they are linkable because of the polynomial's uniqueness. Shrunk schemes reduce cost by decreasing the space of community pseudonyms. Hints attempt to provide confusion by overlapping community pseudonyms, but may be computation-wise inefficient because of the number of failed verifications.

Table 6.1: Cost of several community pseudonym schemes with m communities per participants, and n_e participants.

Technique	**Sender**			**Receiver**		
	Lookups	Computation	Communication	Lookups	Computation	Memory
Single pseudonym	$\mathcal{O}(m)$	$\emptyset$	$\mathcal{O}(m)$	$\mathcal{O}(n_e m)$	$\emptyset$	$\mathcal{O}(m)$
k-anonymity	$\mathcal{O}(km)$	$\emptyset$	$\mathcal{O}(km)$	$\mathcal{O}(n_e km)$	$\emptyset$	$\mathcal{O}(km)$
Pre-computed entire	$\mathcal{O}(m)$	$\emptyset$	$\mathcal{O}(m)$	$\mathcal{O}(n_e m)$	$\emptyset$	$\mathcal{O}(M)$
Self-generated entire	$\mathcal{O}(m)$	$\mathcal{O}(m)$	$\mathcal{O}(m)$	$\emptyset$	$\mathcal{O}(n_e m^2)$	$\mathcal{O}(m)$
Hash bins	$\mathcal{O}(m)$	$\mathcal{O}(m)$	$\mathcal{O}(m)$	$\emptyset$	$\mathcal{O}(n_e m \log(m))$	$\mathcal{O}(m)$
IHME	$\mathcal{O}(1)$	$\emptyset$	$\mathcal{O}(m)$	$\emptyset$	$\mathcal{O}(n_e m)$	$\mathcal{O}(m)$
Pre-computed shrunk	$\mathcal{O}(m)$	$\emptyset$	$\mathcal{O}(m)$	$\mathcal{O}(n_e m)$	$\emptyset$	$\mathcal{O}(hM)$
Self-generated shrunk	$\mathcal{O}(m)$	$\emptyset$	$\mathcal{O}(m)$	$\mathcal{O}(n_e m)$	$\emptyset$	$\mathcal{O}(len \cdot m)$
Hints	$\mathcal{O}(m)$	$\mathcal{O}(m)$	$\mathcal{O}(m)$	$\emptyset$	$\mathcal{O}(n_e m^2)$	$\mathcal{O}(m)$

6.6.5 Security Analysis

We study the security properties of schemes based on symmetric cryptography and show how asymmetric cryptography can be used in local regions to thwart misbehavior.

Forward Secrecy

The symmetric key shared by community members can be used as a digital credential to authenticate other community members and encrypt communications. However, such a widely

shared secret could leak if one community member is compromised or malicious. Hence, pseudonym schemes using symmetric cryptography do not provide *forward secrecy*: if the secret key of a community is leaked, then all community pseudonyms, even the ones generated before the leak [223], are no longer trustworthy. In addition, an adversary that obtains a community secret can break community privacy by observing the messages broadcasted by other users.

In order to protect forward secrecy, community pseudonym schemes based on symmetric key cryptography can be modified to *change over time* the shared secret of communities. As investigated in [223], symmetric key updates should be generated in mobile devices in a distributed fashion, e.g., relying on pseudo-random functions. The symmetric key rekeying can also be done relying on the asymmetric credentials of users as described in [175]. Such rekeying operations are also needed when new members join a community or existing members leave the community.

These rekeying operations require coordination among all community members. We argue that the cost of coordinating key updates is lower than the cost of relying on asymmetric cryptography to obtain similar properties. In the case of pervasive social networks, the symmetric keys of the communities could be changed at regular interval as suggested in [200] to minimize costs. In addition, mobile devices can communicate with the CA in order to check whether they are using the right current secret.

Revocation

The detection of misbehaving community members is hard if symmetric cryptography is used. For example, any community member can broadcast spam messages to other community members, and as the sender is not uniquely authenticated, he may be difficult to identify.

This problem can be overcome with the use of digital signatures within the secure channel established with the symmetric key of the community. In the event of a spamming attack, community members can require other members to use their PKI credentials in their messages [213]. This will induce a larger cost on all community members in the region of the network where the spamming attack is taking place. Users now authenticated with their personal credentials can be reported to a central server and revoked by using traditional revocation algorithms [226]. Such mechanism could also detect the presence of Sybil attacks.

6.7 Evaluation of Community Privacy

We evaluate privacy (in terms of CAN and CUN) provided by the different community pseudonym schemes with respect to passive and active adversaries. As previously described, an adversary collects information from the network and obtains $\mathbf{G}$ from the passive oracle, $\mathbf{G_A}$ from the active oracle and $\mathbf{G}'$ from the combination of the above two graphs.

6.7.1 Community Anonymity Analysis

Let us define $\rho =$ "$\mathcal{A}$ solves the CAN challenge". The probability that an adversary successfully answers a CAN challenge depends on the information the adversary might have about

the community pseudonym. Formally, we can write $Pr(\rho)$ as follows:

$$\begin{aligned} \sigma = Pr(\rho) &= Pr(\rho|p_b \in \mathbf{G_w})Pr(p_b \in \mathbf{G_w}) \\ &+ Pr(\rho|p_b \in \mathbf{G_f})Pr(p_b \in \mathbf{G_f}) \end{aligned} \tag{6.6}$$

where $\mathbf{G_f}$ is the subgraph of $\mathbf{G}$ containing all edges with weight equal to 0 or 1, $\mathbf{G_w}$ is the subgraph of $\mathbf{G}$ containing all edges with weight in $(0, 1)$ and p_b is the community pseudonym send to adversary by the challenger. More specifically,

$$Pr(\rho) = \sum_{C_i \notin \mathbf{G_A}} Pr(\text{“}\mathcal{A} \text{ picks } C_i\text{”})Pr(\rho_i) \tag{6.7}$$

where $Pr(\rho_i)$ is $Pr(\text{“}\mathcal{A}$ solves the CAN challenge for $C_i\text{”})$.

In other words, the adversary may know the community of p_b ($\mathbf{G_f}$) or have statistical information ($\mathbf{G_w}$) about the community of p_b. The probabilities $Pr(p_b \in \mathbf{G_x})$ depend on the type of adversary (i.e., passive or active), its strength (s and ℓ) as well as on the community pseudonym scheme.

Passive Adversary

Given Eq. (6.4) and (6.6), the advantage in the CAN challenge-response game for the passive adversary can be computed as follows:

$$Adv_s^{CAN} = Pr(\rho|p_b \in \mathbf{G_w}) - \frac{1}{2} \tag{6.8}$$

Indeed, in the passive case, the graph $\mathbf{G_f}$ is empty, as the adversary cannot be sure with probability 1 of the relation between community pseudonyms and communities. Hence, we have $Pr(p_b \in \mathbf{G_f}) = 0$, or equivalently, $Pr(p_b \in \mathbf{G_w}) = 1$.

Equation (6.8) shows that community anonymity exclusively depends on the information contained in $\mathbf{G_w}$, i.e., the ability of the adversary to exploit the collected messages s in order to profile communities and link community pseudonyms to communities.

Active Adversary

An active adversary can interact with the oracle $\mathcal{O}_A$ to discover the relation between some community pseudonyms and communities with probability 1. We compute the advantage of the adversary as follows:

$$Adv_{s,\ell}^{CAN} = \begin{cases} (Pr(\rho|p_b \in \mathbf{G'_w})\alpha' + (1-\alpha')) - \frac{1}{2} & \text{if } |v'_f| < |\mathcal{P}| - 1 \\ \frac{1}{2} & \text{else} \end{cases} \tag{6.9}$$

with

$$\alpha' = \frac{1}{2}\frac{1}{|V|}\sum_{v \in V}\frac{|e_w^v|}{|e^v|} + \frac{1}{2}\frac{1}{|V|}\sum_{v \in V}\frac{|e_w| - |e_w^v|}{|e| - |e^v|} \tag{6.10}$$

where $\mathbf{G'_w}$ is the subgraph of $\mathbf{G'}$ containing all edges with weight in $(0, 1)$, $\mathbf{G'_f}$ is the subgraph of $\mathbf{G'}$ containing all edges with weight equal to 0 or 1, V is the set of nodes in the communities of $\mathbf{G'}$, v'_f are the nodes in V with an edge with weight 1, e_w are the edges in $\mathbf{G'_w}$, e^v are

the edges connected to a node v, e_w^v are the edges in $\mathbf{G'_w}$ connected to a node v, $\mathcal{P}$ is the set of all community pseudonyms, α' is the probability that a pseudonym belongs to $\mathbf{G'_w}$ and $1-\alpha'$ is the probability that a pseudonym belongs to $\mathbf{G'_f}$. Equation (6.9) indicates that with probability $1-\alpha'$ the adversary knows the challenge and is always successful, whereas with probability α', the adversary must guess based on the information in $\mathbf{G'_w}$. The probability α' is composed of two parts. First, it depends on the probability that the challenger selects the community C_i queried by the adversary (1/2) and on the proportion of edges that belong to $\mathbf{G'_w}$ in C_i. Second, it depends on the probability that the challenger does not select the community C_i queried by the adversary (1/2) and on the proportion of edges that belong to $\mathbf{G'_w}$ in $\mathcal{C}-C_i$.

If $\alpha'=0$, meaning that community pseudonyms exclusively belong to $\mathbf{G'_f}$ (i.e., the number of interactions ℓ is large), then the advantage is $Adv_{s,\ell}^{CAN}=1/2$ indicating that the adversary can always successfully guess. If $\alpha'=1$, meaning that community pseudonyms exclusively belong to $\mathbf{G'_w}$, then the advantage is $Adv_{s,\ell}^{CAN}=Pr(\rho|p_b\in\mathbf{G'_w})-1/2$, the same as for the passive adversary.

6.7.2 Community Unlinkability Analysis

Let us define $\upsilon=$ "$\mathcal{A}$ solves the CUN challenge". As before, we can write:

$$\begin{aligned}\mu=Pr(\upsilon) &= Pr(\upsilon|p_b,p_d\in\mathbf{G_f})Pr(p_b,p_d\in\mathbf{G_f}) \\ &+ Pr(\upsilon|p_b,p_d\in\mathbf{G_w})Pr(p_b,p_d\in\mathbf{G_w}) \\ &+ Pr(\upsilon|p_b\in\mathbf{G_f},p_d\in\mathbf{G_w})Pr(p_b\in\mathbf{G_f},p_d\in\mathbf{G_w}) \\ &+ Pr(\upsilon|p_b\in\mathbf{G_w},p_d\in\mathbf{G_f})Pr(p_b\in\mathbf{G_w},p_d\in\mathbf{G_f})\end{aligned} \tag{6.11}$$

In other words, the probability of success of the adversary depends on the type of information it has on the challenges.

Passive Adversary

Given Eq. (6.5) and (6.11), the advantage in the CAN challenge-response game for the passive adversary can be computed as follows:

$$Adv_s^{CUN}=Pr(\upsilon|p_b,p_d\in\mathbf{G_w})-\frac{1}{2} \tag{6.12}$$

Theorem 6 shows that if the adversary is able to break the CAN challenge successfully, i.e., $\sigma\in(0.5,1]$, then the adversary can also break CUN. Hence, the probability of breaking the CUN challenge in the passive case η is:

$$\eta=Pr(\upsilon|p_b,p_d\in\mathbf{G_w})=\sigma^2+\frac{(1-\sigma)^2}{2}+\frac{\sigma(1-\sigma)}{2} \tag{6.13}$$

The advantage is obtained as follows: as with CAN, in the passive case, $\mathbf{G_f}$ is empty. Hence, we have $Pr(p_b,p_d\in\mathbf{G_f})=Pr(p_b\in\mathbf{G_w},p_d\in\mathbf{G_f})=Pr(p_b\in\mathbf{G_f},p_d\in\mathbf{G_w})=0$ and $Pr(p_b,p_d\in\mathbf{G_w})=1$. We can relate the probability of success υ to the probability of success σ in the CAN case. Indeed, the adversary can run the CAN challenge response protocol for both communities, and its success rate for those CAN challenges will determine its success rate for the CUN challenge. We obtain that η depends on the following probabilities:

the adversary can guess both CAN challenges correctly (σ^2); the adversary cannot guess both CAN challenges correctly but answers the CUN challenge correctly ($(1-\sigma)^2/2$); or the adversary can guess one of the CAN challenges correctly and CUN correctly ($\sigma(1-\sigma)/2$). Note that Theorem 6 identifies the general relation between CUN and CAN, whereas Equation (6.13) considers the particular case in which an adversary breaks CUN using the solution for CAN. More details about this expression can be found in the proof of Thm 6.

We observe that if $\sigma = 0$ (meaning that the adversary does not break community anonymity), then $\eta = 1/2$ and the advantage is minimum. Instead, if $\sigma = 1$, then the probability of success $\eta = 1$, indicating that the adversary has maximum advantage.

Active Adversary

If $\mathcal{A}$ is an active adversary, it discovers the relation between some community pseudonyms and communities. Using (6.6) and (6.11), we compute the advantage of the adversary as follows:

$$Adv_{s,\ell}^{CUN} = (1-\alpha')^2 + \eta\alpha'^2 + 2\sigma'\alpha'(1-\alpha') - \frac{1}{2} \tag{6.14}$$

We obtain the above formula by again relating the CUN advantage to the probability of success in the CAN challenge response game. With probability $(1-\alpha')^2$, the adversary is given two pseudonyms that it knows (in $\mathbf{G}'_{\mathbf{f}}$) and can always guess the CUN challenge correctly. With probability α'^2, the adversary does not know any of the pseudonyms and must guess with success η (like a passive adversary). Finally, with probability $\alpha'(1-\alpha')$, the adversary knows one of the two pseudonyms in the CUN challenge and has to guess the other pseudonym using the CAN probability of success σ'.

If $\sigma' = 1/2$ (meaning that the CAN algorithm does not help), then the advantage is $1 - \alpha' + \alpha'^2/2$, and depends exclusively on α'. If $\alpha' = 0$, then the advantage is maximal (0.5). Instead, if $\alpha' = 1$, then the advantage is minimal (0). If $\sigma = 1'$, then the advantage is maximum (0.5) for any value of α'. This indicates that if the adversary can solve the CAN challenge, it can also solve the CUN challenge.

6.7.3 Evaluation

In this section, we compute the CAN and CUN advantages for different community pseudonym schemes and adversaries based on numerical evaluations and simulations.

Numerical Evaluation

The CAN and CUN advantages depend on the CAN probability of success σ and the probability α'. In Fig. 6.3, we numerically evaluate the evolution of the CAN and CUN advantages with respect to those parameters. We observe in Fig. 6.3 (a) that the CAN advantage in the passive case increases linearly with σ. In practice, σ is an increasing function of s that depends on the attack of the adversary and on the community pseudonym scheme. In general, the higher the number of collected messages s is, the higher the advantage of the adversary is. In contrast, the CUN advantage (Fig. 6.3 (c)) increases non-linearly in the passive case as indicated in Eq. (6.13). Note that α' has no influence because the attack is passive.

In Fig. 6.3 (b) & (d), we plot the CAN and CUN advantages of an active adversary and observe that as α' decreases, the success of the adversary dramatically increases. This means

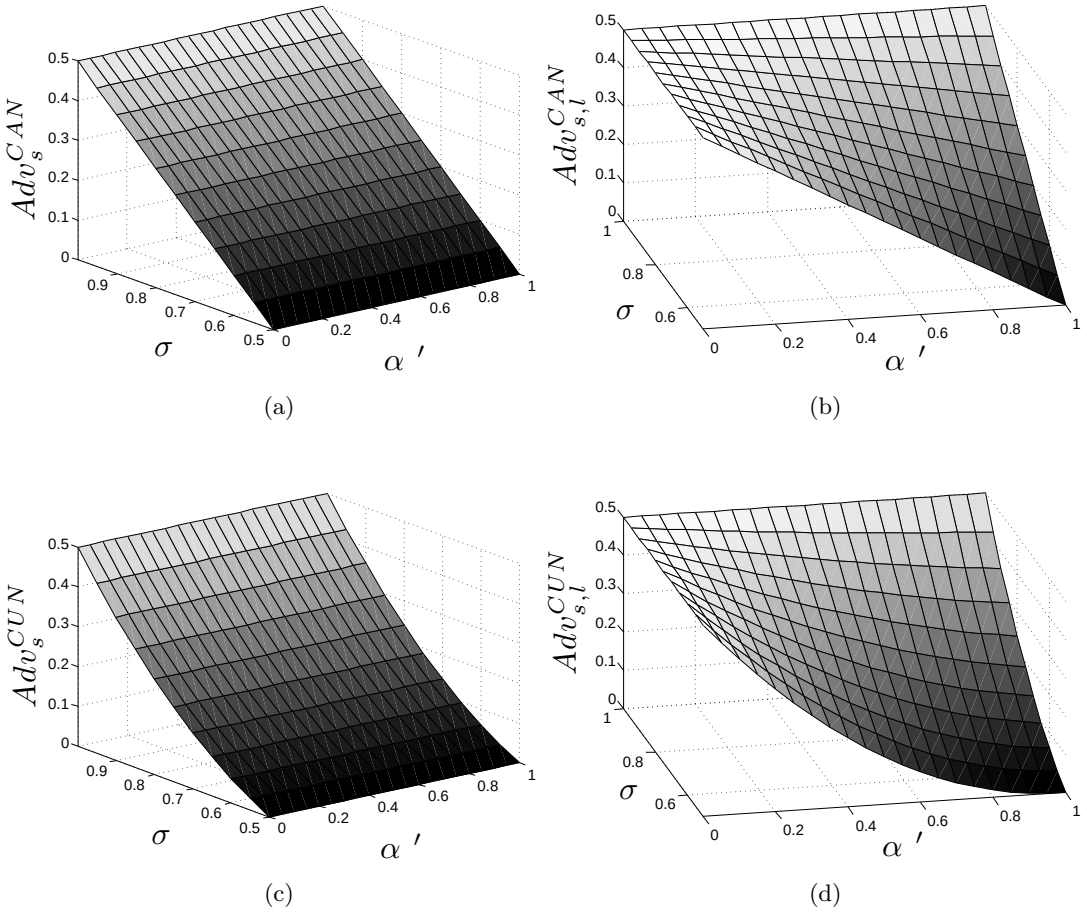


Figure 6.3: Numerical evaluation of the advantage of a passive and an active adversary. (a) Adv_s^{CAN}, (b) $Adv_{s,\ell}^{CAN}$, (c) Adv_s^{CUN} and (d) $Adv_{s,\ell}^{CUN}$

that compromising devices considerably helps the adversary. These plots give an interesting insight on how breaking CAN affects the success probability of CUN (assuming that the adversary breaks CUN using an algorithm to break CAN). These results provide a baseline to estimate the possible values of the advantage in different scenarios. Yet, the relation between σ and s is missing, as it depends on the attack and the community pseudonym scheme. In the following section, we investigate the relation between σ and s by simulating attacks on community pseudonym schemes and computing the advantage of the adversary for different values of s.

Simulation Setup

We simulate a pervasive social network composed of wireless mobile devices. The simulator models the mobility of users, the aforementioned community pseudonym schemes and an attack on community privacy by passive and active adversaries. We model $n = 50$ users moving on a grid of 1km $\times$ 1km where each step is one meter according to state-of-the-art random walk mobility model [218]. We consider that mobile nodes move with a constant

speed. Directions are chosen out of $[0, 2\pi]$ with granularity $\pi/2$. We leave it for future work to consider other mobility models. Two nodes are assumed to be in communication range if they are within a 100 meters. Mobile devices rely on the communication protocol described in Section 6.3 to interact. We consider that there are $m = 20$ communities, that each user belongs to 6 communities and that there are $M = 100000$ possible community pseudonyms. We consider that the adversary collects all messages broadcasted in a time slot t and that at every time slot all devices broadcast a message to all communities they belong to and if possible change community pseudonyms. We have $s = t \cdot n \cdot n_c$ where n_c is the number of communities per user.

Attack Description

The goal of the adversary is to obtain the graph $\mathbf{G}$ of the relation between community pseudonyms and communities. We consider a global passive adversary that collects community messages from the entire network. The attack consists of two parts: traffic analysis and community detection.

In the traffic analysis part, the adversary aims at obtaining a graph $\mathbf{G_e}$ where community pseudonyms are the nodes, and weighted edges indicate the possibility that community pseudonyms belong to the same community. To do so, the adversary links community pseudonyms based on the wireless communication patterns between nodes. For example, when two nodes A and B are in proximity and exchange their community pseudonyms, if the adversary observes a unicast communication occurring after the initial handshake, it can conclude that the two devices have at least one community in common. It can thus link all community pseudonyms broadcasted by device A to those of device B. A link with weight 1 indicates that two community pseudonyms may belong to the same community. Similarly, an adversary can link with weight 0 all community pseudonyms used by a given device, because such community pseudonyms must belong to different communities.

In this work, we assume that the adversary attempts to group community pseudonyms into communities using the community detection algorithm of [33] on graph $\mathbf{G_e}$. The problem of community detection in social graphs has been considerably investigated in previous works [33]: the challenge consists in efficiently detecting communities in large graphs that describe the relation between several entities. In some cases, community detection is difficult because the relation between entities in the graph is hidden. The so-called problem of *adversarial* community detection consists in detecting communities of privacy-conscious users [160].

If several community pseudonyms are linked to each other with weight 1, they are interpreted as a community by the community detection algorithm. The adversary can thus cluster community pseudonyms and obtains inferred communities. Yet, it still does not know the correspondence between real communities and inferred communities. An adversary must guess the relation between inferred communities and real communities. It can do so based on the profile of communities (e.g., some communities may frequently visit specific locations). In this work, we consider a very strong adversary that can correctly map inferred communities to the corresponding real communities. In reality, this is non-trivial, but we leave the study of such attacks for future work. In this setting, we compute the probability of success of the adversary by considering the amount of overlap between each inferred and real community and also considering the auxiliary information the adversary has about the relation of community pseudonyms and communities: for example, the adversary may know that several community

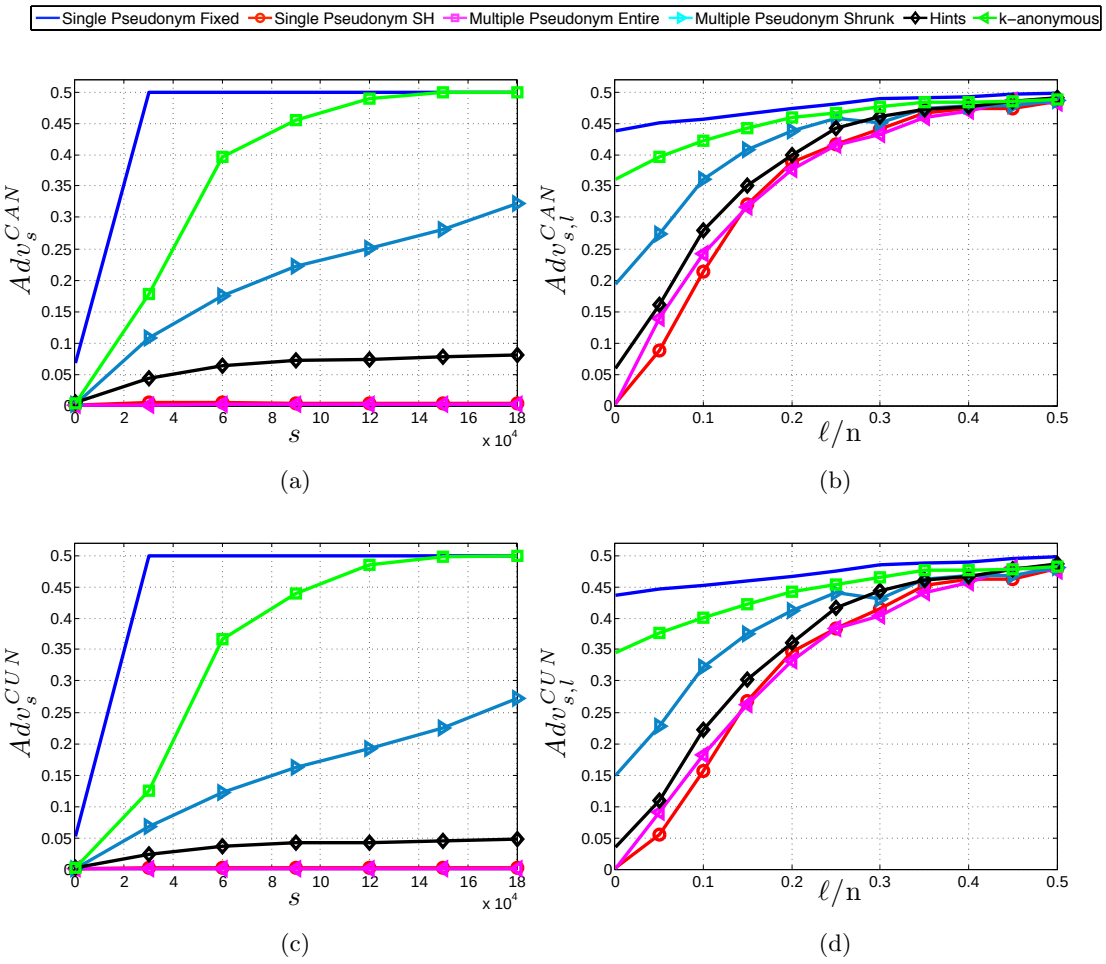


Figure 6.4: CAN and CUN Advantages obtained in simulations of a passive and active adversary for different community pseudonym schemes. (a) Adv_s^{CAN} with respect to the number of collected messages s and (b) $Adv_{s,l}^{CAN}$ with respect to the fraction of compromised devices ℓ/n. (c) Adv_s^{CUN} with respect to the number of collected messages s and (d) $Adv_{s,l}^{CUN}$ with respect to the fraction of compromised devices ℓ/n.

pseudonyms do not belong to the same community or belong to another community. This attack converts the graph $\mathbf{G_e}$ into $\mathbf{G}$.

In the case of an active attack, the attacker selects at random ℓ devices that it can compromise. For each of these devices, it discovers the entire relation between community pseudonyms and communities. We do not argue that the described attack is the best attack an adversary can perform. Still, we believe that the strong adversary model considered in this work (i.e., a global adversary that links inferred communities to real communities) provides insight into the ability to protect community privacy by comparing the performance of community pseudonym schemes.

Simulation Results

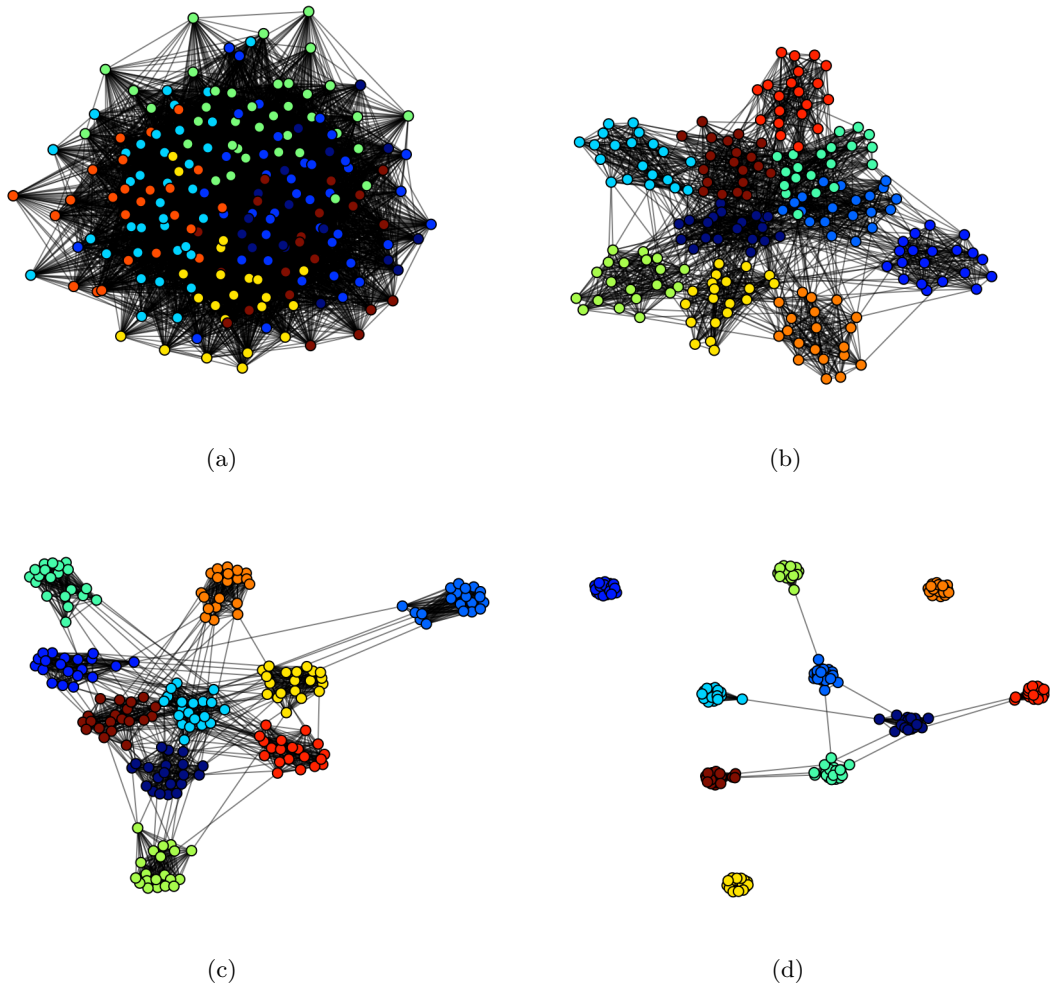
In Fig. 6.4 (a) and (c), we show the CAN and CUN advantages in the case of a passive adversary. We observe that for most schemes, the advantage increases with the number of collected messages s. The fixed single pseudonym scheme (*Single Pseudonym Fixed*) does not provide any community anonymity because, with such a scheme, it is trivial for the adversary to link all community pseudonyms together. In contrast, the single pseudonym scheme similar to linkable secret handshakes (*Single Pseudonym SH*) results in a low advantage. This is because the set of community pseudonyms broadcasted by mobile devices is static for a given user (i.e., always the same), but different from other users. Hence, linkable secret handshakes schemes effectively protect community privacy. Nevertheless, with such schemes, an adversary can trivially track users' whereabouts, thus jeopardizing location privacy.

The scheme resulting in the lowest advantage is the multiple pseudonym scheme over the entire domain (*Multiple Pseudonym Entire*). The reason is that the probability of reusing a community pseudonym tends to be very small as the domain of community pseudonyms of size M is large. This scheme also provides location privacy as it is not possible to link community pseudonyms to track users' whereabouts. As soon as M is decreased (*Multiple Pseudonym Shrunk* with $h = 0.01$), the advantage increases considerably. This shows that reusing community pseudonyms significantly reduces community privacy, as it makes it possible for an adversary to correlate different messages. Note that as the CAN and CUN advantage increase, its ability to track mobile devices also improves.

The negative effect of shrinking the community pseudonym set can be attenuated by using *Hints*. We implement Hints by considering the shrunk scheme with $h = 0.01$ and by selecting community pseudonyms for each community from the set of hM community pseudonyms without removing selected elements (i.e., with repetitions). Hints reduce the advantage compared to the shrunk scheme by introducing confusion in the set of community pseudonyms assigned to communities: community pseudonyms can be reused for different purposes. This technique provides community privacy by exploiting the limitations of community detection algorithms: these algorithms assume that each node in the graph belongs to a single community. Hence, Hints extend the lifetime of shrunk community pseudonym sets by reducing the advantage of the adversary.

The *k-anonymous* scheme complements the *single pseudonym SH* scheme by selecting $k - 1$ other community pseudonyms [224]. We consider that extra community pseudonyms are chosen from communities the sender does not belong to (e.g., pseudonyms eavesdropped in previous interactions). We observe that the k-anonymous scheme with $k = 3$ performs worse than the *single pseudonym SH* scheme. The graph $\mathbf{G_e}$ of the adversary (Fig. 6.5) shows that it can quickly distinguish communities. The reason is that adding $k - 1$ community pseudonyms to each message leaks additional information: the adversary learns that these groups of pseudonyms do not belong to the same community. With a k-anonymous scheme, the advantage increases even faster than the shrunk domain approach. Even if the $k - 1$ other community pseudonyms were chosen at random, the adversary could still statistically recognize the community pseudonyms of mobile devices out of the random values and k-anonymous schemes would at best achieve the same level of privacy than the raw single pseudonym SH scheme. In summary, k-anonymous schemes do not provide community privacy and can even be detrimental to privacy.

In Figures 6.4 (b) and (d), we show the advantage of the active adversary with respect to the fraction of compromised devices ℓ/n averaged across all values of s. We observe that the



(a) (b)

(c) (d)

Figure 6.5: Graph $\mathbf{G_e}$ resulting from an attack on a k-anonymous scheme with $n = 50$, $m = 10$ and $n_c = 3$. The color of a node indicates the community it belongs to. (a) $t = 1$, (b) $t = 2000$, (c) $t = 5000$ and (d) $t = 10000$.

adversary can rely on the information actively learned to infer the relation between community pseudonyms and communities. The increase is non-linear: even if devices are compromised at random, it is sufficient to compromise a fraction of those devices to have a significant impact on community privacy. In practice, an adversary may target devices that belong to a large number of communities in order to improve its effectiveness.

The advantages obtained through simulations of the attack with respect to community pseudonym schemes can be mapped to our numerical results in Fig. 6.3. Our numerical model allows us to evaluate the performance of the attack and the community pseudonym scheme. For example, in the case of an active attack, we know from our numerical results the minimum and maximum advantage the adversary could obtain for different values of α'. We can map the advantage obtained from simulations with a certain value of s and a certain number of compromised devices l to a point in Fig. 6.3 in order to evaluate the performance of the attack.

6.8 Conclusion

We have considered the problem of privacy in pervasive social networks exhibiting potential threats on location, data and community privacy. We considered that people may share information with other users based on a social graph and evaluated the privacy risks introduced by such social network information used atop peer-to-peer wireless networks. We identified the need to protect community privacy and proposed a framework based on challenge-response games to study it. An interesting outcome of the framework is the analytical relation obtained between community anonymity and community unlinkability. Although the relation between these two properties was previously studied [178], to the best of our knowledge, we are the first to analytically relate these two properties. We also showed how to use asymmetric cryptography in local areas in order to reduce cost and provide resilience to attacks.

By means of simulations, we evaluated the privacy provided by different privacy-preserving schemes. We obtained that shrinking the number of possible community pseudonyms significantly reduces the achievable privacy. This result outlines the delicate trade-off between the achievable community privacy and cost of community pseudonym schemes. Our analysis enables system designers to tune their shrunk scheme to a desired privacy level, by for example regularly changing the set of community pseudonyms in order to bound the adversarial advantage. We also showed how a technique called Hints can increase community privacy. We argued that reusing pseudonyms across communities can provide a good cost/privacy trade-off. We also demonstrated that k-anonymous schemes are detrimental to community privacy. In the future, we intend to investigate other communication models and study with practical implementations the cost introduced by community pseudonym schemes.

Publication: [93]

Conclusion

Malo periculosam libertatem quam quietum servitium.[a]

[a] I prefer liberty with danger to peace with slavery.

Latin Proverb

Modern communication technologies increasingly rely on contextual information in order to, notably, offer customized services. For example, when mobile users query search engines, search results can be customized to users' locations. Most mobile handsets are now equipped with localization technology and are thus compatible with such location-based services. In the future, one of the major shifts in communication technologies will be the adoption of ad hoc wireless communications complementing infrastructure-based communications. Such peer-to-peer wireless communications will further improve the environment awareness of mobile devices and enable a new breed of context-based services.

In this work, we study privacy issues that appear with the sharing of location information. We argue that location information indirectly reveals a large amount of personal data. In particular, third parties can learn users' whereabouts and jeopardize users' *location privacy*. New communication technologies require novel mechanisms for the protection of personal information. We observe that the increased pervasiveness of communication technologies may force such mechanisms to become increasingly distributed and complex. In such setting, mobile devices will tend to behave rationally in order to optimize their cost-privacy trade-off, thus jeopardizing most privacy-preserving protocols. In view of these observations, the novel contributions of this thesis are as follows.

In **Part I**, we push further the understanding of location privacy threats. In **Chapter 1**, we study the ability of third parties to de-anonymize location traces given a certain quantity of location information and given how location information is collected. In particular we consider the threat induced by location-based services: we show that operators of location-based services frequently find the identity and points of interest of their users based on a *small number* of location samples taken from users' everyday lives. We observe that the type of location information shared by users determines the privacy risks: if users access location-based services from personal locations (such as home or work places), they dramatically increase the risk to be identified. These results exhibit a peculiar property of location information, namely that the spatio-temporal correlation of location traces tends to be unique to individuals, and question the ability of privacy-preserving mechanisms to obfuscate highly correlated location traces. These results increase the awareness of location privacy threats and can thus help us design better privacy-preserving mechanisms.

In **Part II**, we study the privacy architecture of distributed wireless networks and present

solutions to preserve location privacy in this setting. In **Chapter 2**, we describe the multiple pseudonym approach and the mix zone concept used to protect location privacy. We explain how mobile devices should change their pseudonyms in a coordinated manner in order to successfully confuse an adversary tracking their whereabouts. We illustrate the approach by providing an example of pseudonym change for vehicular networks. As pseudonym changes introduce a cost to mobile devices and to network operators, we show that pseudonym change strategies should aim at achieving an efficient trade-off between privacy and cost.

In **Chapters 3** and **4**, we set up centralized and distributed approaches for pseudonym changes in order to study the achievable trade-off in different contexts, thus far a research area that has received little attention. Our centralized approach (**Chapter 3**) is modeled as an optimization problem taking into account the mixing effectiveness of mix zones, the distance between mix zones and their cost. By means of simulations, we show an increase in location privacy brought by the optimal placement of mix zones. Yet, we also observe that a global passive adversary can track a large fraction of the nodes. Even if this result may appear negative at first, the threat of a global passive adversary is unlikely due to the high cost to put in place such attack. In practice, an adversary would obtain lower coverage and noisy information about users' mobility. We show that such an adversary is considerably less effective at tracking mobile devices. A centralized approach assumes that users trust a central authority responsible for the establishment of security and privacy in the system. Our work, by making a possible algorithm public, contributes to the trustworthiness of the authority as it provides a basis for comparison.

In our distributed schemes (**Chapter 4**), we introduce a user-centric model of location privacy to measure the evolution of location privacy over time. This model is used to evaluate the strategy of mobile devices individually optimizing their trade-off between privacy and cost. We develop the pseudoGame protocol to model the behavior of rational nodes participating in pseudonym changes. Our game-theoretic analysis shows that rational behavior in the presence of uncertainty tends to decrease the achievable level of location privacy: rational nodes participate less in pseudonym changes because it is more difficult for them to predict the success of pseudonym changes. Using this insight, we develop a dynamic version of the pseudoGame protocol that relies on signaling information from other nodes to dramatically increase the achievable location privacy. By means of simulations, we show that the dynamic pseudoGame protocol outperforms any existing protocols at a much lower cost. These results shed light on the potential application of game-theory to privacy problems in distributed environments and indicate fundamental benefits and limits of the multiple pseudonym approach in mobile networks.

In **Chapter 5**, we further push the analysis of the distributed approach by providing a framework to *analytically* evaluate the privacy obtained with mix zones. Using mean-field approximation techniques, we show that it is possible to measure for different pseudonym change strategies the distribution of the time period over which a pseudonym is used, i.e., the distribution of the age of pseudonyms. With this result, we show that we can design pseudonym change protocols that bound the maximum age of pseudonyms of mobile devices in distributed networks. If the cost of pseudonym changes is high, then we note that it becomes more difficult to bound the distribution of the age of pseudonyms. These results encourage further application of approximation methods to analytically evaluate the achievable privacy in complex systems such as distributed wireless networks.

In **Part III**, we explore how peer-to-peer wireless communications could enhance social interactions between humans. In **Chapter 6**, we study the mechanisms required by so-called

pervasive social networks in order to anonymously identify communities of users. We propose new schemes that complement existing secret handshake schemes proposed in the research literature: we focus on the specific requirements of mobile wireless networks, i.e., low cost and unlinkability. These schemes provide authentication based on symmetric cryptography and unlinkability by using short-term community pseudonyms. We show how the use of asymmetric cryptography can be limited to local regions in order to reduce cost and still be resilient to attacks. We identify the problem of *community privacy* and provide a framework based on challenge-response protocols to formally evaluate the interactions between mobile users that use community pseudonyms and an adversary. With our framework, we analytically related the notion of community anonymity and unlinkability. We also consider an active adversary that can compromise mobile devices in order to obtain their private credentials. Our results indicate that an adversary may, based on observations from pervasive social networks, break the protection provided by several community pseudonym schemes. We also show that some schemes can efficiently protect community privacy.

Future Research Directions

The results of this thesis, with both their limitations and their promises, indicate that it is possible to design communication technologies that effectively protect privacy. By considering various fundamental approaches, we hope to have cleared the way for designing privacy solutions for emerging wireless networks and to inspire future work on the preservation of privacy. Many research directions can be pursued:

- The understanding of privacy threats heavily depends on the information available to an adversary. Hence, an interesting extension of the first part of the thesis would be to further study the influence of the background information of an adversary on its ability to infer information about users. In particular, analytical models capturing the strategy of the adversary, as well as the cost to implement attacks, would definitely improve the understanding of location privacy threats and have an impact on our experimental results.

- In the same vein, the models we developed in the second part of the thesis can be extended in various ways. Mobile devices protecting their privacy in a distributed fashion could rely on information about network conditions to further optimize their strategy. For example, the a priori knowledge of good mixing locations may influence the decision to change pseudonyms of rational devices. Such information could be captured by introducing fictitious strategies in our game model.

- We evaluated the pseudonym change approach based on simulations and numerical evaluations, as we could not evaluate them on real systems. The feedback from a real test bed of devices implementing the protocols suggested in this thesis would provide considerable insight on our theoretical and simulation results. Such a test bed could also be used to evaluate the coexistence of infrastructure-based communications and ad hoc communications. In particular, in the last part of the thesis, we suggested to combine asymmetric and symmetric cryptography to prevent misbehavior and this could be tested in real systems.

- We discussed how privacy threats depend on the context of mobile devices and the information shared by mobile devices. In order to take this into account in the design of privacy-preserving mechanisms, we used tools from other disciplines to better model different contexts. Our results show that such an approach can positively affect the design of privacy protocols. This is still a relatively untouched, yet burgeoning area of research that could be further explored.

Bibliography

[1] http://en.wikipedia.org/wiki/Bluedating.

[2] http://en.wikipedia.org/wiki/Lovegetty.

[3] Blue star project. http://www.csg.ethz.ch/research/projects/Blue_star.

[4] Social serendipity project. http://reality.media.mit.edu/serendipity.php.

[5] Sumo (simulation of urban mobility): An open-source traffic simulator. http://sumo.sourceforge.net.

[6] TIGER maps. http://www.census.gov/geo/www/tiger.

[7] IEEE P1609.2 Version 1. Standard for wireless access in vehicular environments - security services for applications and management messages. In *development*, 2006.

[8] 3rd Generation Partnership Project. 3GPP GSM R99. In *Technical Specification Group Services and System Aspects*, 1999.

[9] A. Acquisti, R. Dingledine, and P. Syverson. On the economics of anonymity. In *Financial Cryptography and Data Security*, 2003.

[10] Aka Aki. The discovery of a lifetime. http://www.aka-aki.com.

[11] WiFi Alliance. Wi-Fi CERTIFIED Wi-Fi Direct: Personal, portable Wi-Fi that goes with you anywhere, any time, 2010. http://www.wi-fi.org/Wi-Fi_Direct.php.

[12] A. Amir, A. Efrat, J. Myllymaki, L. Palaniappan, and K. Wampler. Buddy tracking–efficient proximity detection among mobile friends. *Pervasive and Mobile Computing*, 3(5):489–511, 2007.

[13] A. G. Amsterdam. Perspectives on the fourth amendment. *Minnesota Law Review*, 58:349, 1973.

[14] R. Anderson. *Security Engineering: A Guide to Building Dependable Distributed Systems*. Wiley Publishing, Inc., 2008.

[15] M. Arrington. The holy grail for mobile social networks, 2007. http://www.techcrunch.com/2007/09/11/the-holy-grail-for-mobile-social-networks.

[16] D. Ashbrook and T. Starner. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.

[17] N. Asokan and P. Ginzboorg. Key agreement in ad-hoc networks. *Computer Communications*, 23:1627–1637, 1999.

[18] G. Ateniese, J. Camenisch, M. Joye, and G. Tsudik. A practical and provably secure coalition-resistant group signature scheme. In *CRYPTO*, volume 1880, pages 255–270, 2000.

[19] G. Ateniese, A. Herzberg, H. Krawczyk, and G. Tsudik. Untraceable mobility or how to travel incognito. *Comput. Netw.*, 31(9):871–884, 1999.

[20] G. Avoine. Radio frequency identification: Adversary model and attacks on existing protocols. Lasec-report-2005-001, EPFL, 2005.

[21] D. Balfanz, G. Durfee, N. Shankar, D. Smetters, J. Staddon, and H.-C. Wong. Secret handshakes from pairing-based key agreements. In *IEEE Symposium on Security and Privacy*, pages 180–196, 2003.

[22] R. Barnes, A. Cooper, R. Sparks, and C. Jennings. IETF geographic location/privacy. http://www.ietf.org/dyn/wg/charter/geopriv-charter.html.

[23] A. Barth, D. Boneh, and B. Waters. Privacy in encrypted content distribution using private broadcast encryption. In *Financial Cryptography and Data Security*, 2006.

[24] D.L. Baumer, J. B. Earp, and J.C. Poindexter. Internet privacy law: A comparison between the United States and the European Union. *Computers & Security*, 23(5):400–412, 2004.

[25] M. Benaim and J.-Y. Le Boudec. A class of mean field interaction models for computer and communication systems. *Performance Evaluation*, 65(11-12):823–838, 2008.

[26] A. R. Beresford. *Location Privacy in Ubiquitous Computing*. PhD thesis, University of Cambridge, 2005.

[27] A. R. Beresford and F. Stajano. Location privacy in pervasive computing. *Pervasive Computing, IEEE*, 2(1):46–55, 2003.

[28] A. R. Beresford and F. Stajano. Mix zones: user privacy in location-aware services. In *Pervasive Computing and Communications Workshops*, pages 127–131, 2004.

[29] O. Berthold, A. Pfitzmann, and R. Standtke. The disadvantages of free MIX routes and how to overcome them. *Lecture Notes in Computer Science*, 2009:30–45, 2001.

[30] C. Bettini, X. S. Wang, and S. Jajodia. Protecting privacy against location-based personal identification. In *SDM*, 2005.

[31] Google Mobile Blog. Finding places "near me now" is easier and faster than ever, 2010. http://googlemobile.blogspot.com/2010/01/finding-places-near-me-now-is-easier.html.

[32] Official Nokia Blog. Nokia instant community gets you social, 2010. http://conversations.nokia.com/2010/05/25/nokia-instant-community-gets-you-social.

[33] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[34] R. Bohme, G. Danezis, C. Diaz, S. Kopsell, and A. Pfitzmann. Mix cascades vs. peer-to-peer: Is one concept superior. In *PET*, 2004.

[35] D. Boneh, X. Boyen, and H. Shacham. Short group signatures using strong diffie-hellman. In *CRYPTO*, volume 3152, pages 41–55, 2004.

[36] D. Boyd. Making sense of privacy and publicity. In *SXSW*, 2010. http://www.danah.org/papers/talks/2010/SXSW2010.html.

[37] R. Bradbury. *Fahrenheit 451*. Cornelsen, 2005.

[38] R. W. Bradshaw, J. E. Holt, and K. E. Seamons. Concealing complex policies with hidden credentials. In *CCS*, 2006.

[39] V. A. Brennen. The keysigning party howto, 2008. http://cryptnet.net/fdp/crypto/keysigning_party/en/keysigning_party.html.

[40] V. Brik, S. Banerjee, M. Gruteser, and S. Oh. Wireless device identification with radiometric signatures. In *MobiCom*, 2008.

[41] D. Brin. *The Transparent Society*. Perseus Books, 1998.

[42] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva. A performance comparison of multi-hop wireless ad hoc network routing protocols. In *MobiCom*, pages 85–97, 1998.

[43] L. Buttyan, T. Holczer, and I. Vajda. On the effectiveness of changing pseudonyms to provide location privacy in VANETs. In *ESAS*, 2007.

[44] L. Buttyan, T. Holczer, A. Weimerskirch, and W. Whyte. SLOW: A practical pseudonym changing scheme for location privacy in vanets. In *VNC*, 2009.

[45] L. Buttyan and J.-P. Hubaux. *Security and Cooperation in Wireless Networks*. Cambridge University Press, 2008.

[46] G. Calandriello, P. Papadimitratos, A. Lloy, and J.P. Hubaux. Efficient and robust pseudonymous authentication in vanets. In *ACM VANET*, 2007.

[47] J. Camenisch and E. Van Herreweghen. Design and implementation of the Idemix anonymous credential system. In *CCS*, 2002.

[48] J. Camenisch, S. Hohenberger, M. Kohlweiss, A. Lysyanskaya, and M. Meyerovich. How to win the clone wars: efficient periodic n-times anonymous authentication. In *CCS*, 2006.

[49] J. Camenisch and A. Lysyanskaya. Efficient non-transferable anonymous multi-show credential system with optional anonymity revocation. In *EUROCRYPT*, volume 2045, 2001.

[50] J. Camenisch and A. Lysyanskaya. Signature schemes and anonymous credentials from bilinear maps. In *CRYPTO*, volume 3152, pages 56–72, 2004.

[51] J. Camenisch and M. Stadler. Efficient group signature schemes for large groups. In *CRYPTO*, volume 1296, pages 410–424, 1997.

[52] J. Camenisch and G. Zaverucha. Private intersection of certified sets. In *Financial Cryptography and Data Security*, pages 108–127, 2009.

[53] J. Camenish, S. Hohenberger, and M. O. Pedersen. Batch verification of short signatures. In *EUROCRYPT*, volume 4515, pages 246–263, 2007.

[54] T. Camp, J. Boleng, and V. Davies. A survey of mobility models for ad hoc network research. *Wireless Communication & Mobile Computing (WCMC)*, 2(5):483–502, 2002.

[55] A. Cavoukian. Privacy by design, 2009. http://www.ontla.on.ca/library/repository/mon/23002/289982.pdf.

[56] A. Chaintreau, J.-Y. Le Boudec, and N. Ristanovic. The age of gossip: Spatial mean-field regime. In *ACM Sigmetrics*, 2009.

[57] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6:606–620, 2007.

[58] D. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 24(2):84–90, February 1981.

[59] D. Chaum. Security without identification: Transaction systems to make big brother obsolete. *Communications of the ACM*, 28(10), 1985.

[60] D. Chaum and E. van Heyst. Group signatures. In *EUROCRYPT Workshop on the Theory and Application of Cryptographic Techniques*, volume 547, pages 257–265, 1991.

[61] C.-H. O. Chen, C.-W. Chen, C. Kuo, Y.-H. Lai, J. M. McCune, A. Studer, A. Perrig, B.-Y. Yang, and T.-C. Wu. Gangs: gather, authenticate 'n group securely. In *MobiCom*, pages 92–103, 2008.

[62] S.F. Cheng, D.M. Reeves, Y. Vorobeychik, and W.P. Wellman. Notes on equilibria in symmetric games. In *Workshop on Game-Theoretic and Decision-Theoretic Agents*, 2004.

[63] B.Z. Chor, O. Goldreich, and E. Kushilevitz. Private information retrieval, 1998. US Patent 5,855,018.

[64] Cloudmade. Makes maps differently. http://cloudmade.com.

[65] S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge. Location disclosure to social relations: why, when, & what people want to share. In *SIGCHI*, page 90, 2005.

[66] Car2Car Communication Consortium. http://www.car-to-car.org.

[67] R. Cooper. *Coordination Games*. Cambridge Univ. Press, 1998.

[68] E. De Cristofaro and G.Tsudik. Practical private set intersection protocols with linear complexity. In *Financial Cryptography and Data Security*, 2010.

[69] M. Dahl, S. Delaune, and G. Steel. Formal Analysis of Privacy for Vehicular Mix-Zones. In *ESORICS*, pages 55–70, 2010.

[70] T. Dalenius. Finding a needle in a haystack - or identifying anonymous census records. *Journal of Official Statistics*, 2(3):329–336, 1986.

[71] B. Danev and S. Capkun. Transient-based identification of wireless sensor nodes. In *IPSN*, 2009.

[72] B. Danev, H. Luecken, S. Capkun, and K. Defrawy. Attacks on physical-layer identification. In *WiSec*, pages 89–98, 2010.

[73] G. Danezis. Mix-networks with restricted routes. In *PET*, pages 54–68, 2003.

[74] G. Danezis. *Better Anonymous Communications*. PhD thesis, University of Cambridge, January 2004.

[75] K. El Defrawy and C. Soriente. PEUC-WiN: Privacy enhancement by user cooperation in wireless networks. In *NPSec*, pages 38–43, 2007.

[76] C. Diaz, S. J. Murdoch, and C. Troncoso. Impact of network topology on anonymity and overhead in low-latency anonymity networks. In *PETS*, 2010.

[77] C. Diaz, S. Seys, J. Claessens, and B. Preneel. Towards measuring anonymity. In *PET*, pages 54–68, 2002.

[78] C. Diaz, S. Seys, J. Claessens, and B. Preneel. Towards measuring anonymity. In *Proceedings of PET*, volume 2482, 2002.

[79] R. Dingledine, N. Mathewson, and P. Syverson. Tor: the second-generation onion router. In *USENIX Security Symposium*, pages 21–21, 2004.

[80] John R. Douceur and Judith S. Donath. The sybil attack. In *IPTPS*, 2002.

[81] 5.9GHz DSRC. http://grouper.ieee.org/groups/scc32/dsrc/index.html.

[82] M. Duckham and L. Kulik. Location privacy and location-aware computing. *Dynamic & mobile GIS: investigating change in space and time*, pages 34–51, 2006.

[83] N. Eagle, A. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. In *National Academy of Sciences (PNAS)*, pages 15274–15278, 2009.

[84] EFF. Court rejects warrantless GPS tracking, 2010. http://www.eff.org/press/archives/2010/08/06-0.

[85] J. E. Ekberg. Implementing wibree address privacy. In *IWSSI*, 2007.

[86] K. Fall. A delay-tolerant network architecture for challenged internets. In *SIGCOMM*, 2003.

[87] A. Fiat and M. Naor. Broadcast encryption. In *CRYPTO*, pages 480–491, 1994.

[88] L. Fischer, S. Katzenbeisser, and C. Eckert. Measuring unlinkability revisited. In *ACM WPES*, 2008.

[89] Foursquare. Check-in, find your friends, unlock your city. http://foursquare.com.

[90] J. Franklin, D. McCoy, P. Tabriz, V. Neagoe, J. Randwyk, and D. Sicker. Passive data link layer 802.11 wireless device driver fingerprinting. In *USENIX*, 2006.

[91] M. Freedman, K. Nissim, and B. Pinkas. Efficient private matching and set intersection. In *EuroCRYPT*, pages 1–19, 2004.

[92] E. Frejinger. *Route choice analysis : data, models, algorithms and applications*. PhD thesis, EPFL,, 2008.

[93] J. Freudiger, M. Jadliwala, J.-P. Hubaux, P. Ginzboorg, and I. Aad. Privacy of communities in pervasive social networks. *Under Submission*, 2010.

[94] J. Freudiger, H. Manshaei, J.-Y. Le Boudec, and J.-P. Hubaux. On the age of pseudonyms in mobile ad hoc networks. In *Infocom*, 2010.

[95] J. Freudiger, M. H. Manshaei, J.-P. Hubaux, and D. C. Parkes. On non-cooperative location privacy: A game-theoretic analysis. In *CCS*, 2009.

[96] J. Freudiger, R. Neu, and J.-P. Hubaux. Privacy sharing of user location over online social networks. In *HotPETs*, 2010.

[97] J. Freudiger, M. Raya, M. Felegyhazi, P. Papadimitratos, and J.-P. Hubaux. Mix zones for location privacy in vehicular networks. In *WiN-ITS*, 2007.

[98] J. Freudiger, M. Raya, and J.-P. Hubaux. Self-organized anonymous authentication in mobile networks. In *SECURECOMM*, 2009.

[99] J. Freudiger, R. Shokri, and J.-P. Hubaux. On the optimal placement of mix zones. In *PETS*, 2009.

[100] J. Freudiger, R. Shokri, and J.-P. Hubaux. Evaluating the privacy risk of location-based services. In *Financial Cryptography and Data Security*, 2011.

[101] G. Friedland and R. Sommer. Cybercasing the joint: On the privacy implications of geo-tagging. In *HotSec*, 2010.

[102] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, 1991.

[103] D. C. Gazis. *Traffic Theory*. Kluwer Academic Publishers, 2002.

[104] P. Golle and K. Partridge. On the anonymity of home/work location pairs. In *Pervasive*, 2009.

[105] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[106] B. Greenstein, D. McCoy, J. Pang, T. Kohno, S. Seshan, and D. Wetherall. Improving wireless privacy with an identifier-free link layer protocol. In *MobiSys*, 2008.

[107] M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *MobiSys*, 2003.

[108] M. Gruteser and D. Grunwald. Enhancing location privacy in wireless LAN through disposable interface identifiers: a quantitative analysis. *Mob. Netw. Appl.*, 2005.

[109] S. F. Gurses. *Multilateral Privacy Requirements Analysis in Online Social Network Services*. PhD thesis, KU Leuven, 2010.

[110] J. Hall, M. Barbeau, and E. Kranakis. Enhancing intrusion detection in wireless networks using radio frequency fingerprinting. In *CIIT*, 2004.

[111] J. Halpern and V. Teague. Rational secret sharing and multiparty computation: extended abstract. In *STOC*, pages 623–632, 2004.

[112] J. Harsanyi. Games with incomplete information played by Bayesian players. *Management Science*, 1967.

[113] H. Hartenstein and K. Laberteaux. A tutorial survey on vehicular ad hoc networks. *IEEE Communications Magazine*, 46(6), 2008.

[114] C. Heller. Embracing post-privacy: Optimism towards a future where there is nothing to hide. In *Chaos Communication Congress*, 2008.

[115] M. Hellman and J. Raviv. Probability of error, equivocation, and the Chernoff bound. *IEEE Transactions on Information Theory*, 16(4):368–372, Jul 1970.

[116] A. Herzberg, H. Krawczyk, and G. Tsudik. On travelling incognito. In *Mobile Computing Systems and Applications*, pages 205–211, 1994.

[117] B. Hoh and M. Gruteser. Protecting location privacy through path confusion. In *SECURECOMM*, pages 194–205, 2005.

[118] B. Hoh, M. Gruteser, R. Herring, J. Ban, D. Work, J.-C. Herrera, A. M. Bayen, M. Annavaram, and Q. Jacobson. Virtual trip lines for distributed privacy-preserving traffic monitoring. In *MobiSys*, pages 15–28, 2008.

[119] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Enhancing security and privacy in traffic-monitoring systems. *IEEE Pervasive Computing*, 5(4):38–46, 2006.

[120] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Preserving privacy in GPS traces via path cloaking. In *CCS*, 2007.

[121] L. Huang, K. Matsuura, H. Yamane, and K. Sezaki. Enhancing wireless location privacy using silent period. In *ECNC*, 2005.

[122] L. Huang, K. Matsuura, H. Yamane, and K. Sezaki. Towards modeling wireless location privacy. In *Proceedings of PET*, 2005.

[123] L. Huang, H. Yamane, K. Matsuura, and K. Sezaki. Silent cascade: Enhancing location privacy without communication QoS degradation. In *Security in Pervasive Computing*, pages 165–180, 2006.

[124] M. Humbert, M. H. Manshaei, J. Freudiger, and J.-P. Hubaux. Tracking Games in Mobile Networks. In *Conference on Decision and Game Theory for Security*, 2010.

[125] A. Huxley. *Brave new world*. Vintage Canada, 2007.

[126] S. Izmalkov, S. Micali, and M. Lepinski. Rational secure computation and ideal mechanism design. In *FOCS*, pages 585–595, 2005.

[127] S. Jarecki, J. Kim, and G. Tsudik. Beyond secret handshakes: Affiliation-hiding authenticated key exchange. In *CT RSA*, pages 352–369, 2008.

[128] S. Jarecki and X. Liu. Unlinkable secret handshakes and key-private group key management schemes. In *ACNS*, pages 270–287, 2007.

[129] S. Jarecki and X. Liu. Private Mutual Authentication and Conditional Oblivious Transfer. *CRYPTO*, pages 90–107, 2009.

[130] S. Jarecki and X. Liu. Affiliation-hiding envelope and authentication schemes with efficient support for multiple credentials. *Automata, Languages and Programming*, pages 715–726, 2010.

[131] A. Johnson, J. Feigenbaum, and P. Syverson. Preventing active timing attacks in low-latency anonymous communication. In *PETS*, 2010.

[132] T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, 15(1):52–60, February 1967.

[133] D. Kaplan. *Informatique, libertés, identités*. fyp Editions, 2010.

[134] J. Katz. Bridging game theory and cryptography: Recent results and future directions. In *TCC*, 2008.

[135] M. Khiabani. Metro-sexual, 2009. http://bit.ly/theranMetroSexual.

[136] T. Kohno, A. Broido, and K.C. Claffy. Remote physical device fingerprinting. *TDSC*, 2, 2005.

[137] E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *STACS*, 1999.

[138] J. Krumm. Inference attacks on location tracks. In *Pervasive Computing*, May 2007.

[139] J. Krumm. A survey of computational location privacy. In *Personal and Ubiquitous Computing*, 2008.

[140] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[141] R. Kurzweil. *The singularity is near: When humans transcend biology*. Viking Adult, 2005.

[142] K. Laberteaux and Y.-C.Hu. Strong VANET security on a budget. In *ESCAR*, 2006.

[143] M. Langheinrich. Privacy by design—principles of privacy-aware ubiquitous systems. In *Ubicomp*, pages 273–291, 2001.

[144] R. Laroia. Future of wireless? the proximate Internet. Keynote presentation, 2010. http://www.cedt.iisc.ernet.in/people/kuri/Comsnets/Keynotes/Keynote-Rajiv-Laroia.pdf.

[145] M. Li, K. Sampigethaya, L. Huang, and R. Poovendran. Swing & swap: user-centric approaches towards maximizing location privacy. In *WPES*, pages 19–28, 2006.

[146] N. Li, W. Du, and D. Boneh. Oblivious signature-based envelope. *Distributed Computing*, 17(4):293–302, 2005.

[147] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition using relational Markov networks. In *IJCAI*, 2005.

[148] L. Liao, D. J. Patterson, D. Fox, and H. Kautz. Learning and inferring transportation routines. *Artificial Intelligence*, (171):311–331, 2007.

[149] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37:145–151, 1991.

[150] Loopt. Discover the world around you. http://loopt.com.

[151] D. Lyon. *Surveillance as social sorting: privacy, risk, and digital discrimination*. Routledge, 2003.

[152] A. Lysyanskaya, R. Rivest, A. Sahai, and S. Wolf. Pseudonym systems. In *SAC*, volume 1758, pages 184–199, 1999.

[153] C. Y. T. Ma, D. K. Y. Yau, N. K. Yip, and N. S. V. Rao. Privacy vulnerability of published anonymous mobility traces. In *MobiCom*, 2010.

[154] M. Manulis, B. Pinkas, and B. Poettering. Privacy-Preserving Group Discovery with Linear Complexity. In *ACNS*, pages 420–437, 2010.

[155] L. A. Martucci, M. Kohlweiss, C. Andersson, and A. Panchenko. Self-certified sybil-free pseudonyms. In *WiSec*, 2008.

[156] Game Mobile. http://www.gamemobile.co.uk/bluetoothmobilegames.

[157] M. F. Mokbel, C.-Y. Chow, and W. G. Aref. The new casper: Query processing for location services without compromising privacy. In *VLDB*, 2006.

[158] M. Mouly and M. B. Pautet. *The GSM system for mobile communications*. Telecom Publishing, 1992.

[159] Y. De Mulder, G. Danezis, L. Batina, and B. Preneel. Identification via location-profiling in GSM networks. In *WPES*, 2008.

[160] S. Nagaraja. The impact of unlinkability on adversarial community detection: Effects and countermeasures. In *PETS*, 2010.

[161] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Security and Privacy*, 2009.

[162] J. Nash. Non-cooperative games. *Annals of Mathematics*, 1951.

[163] United Nations. Article 12. The Universal Declaration Of Human Rights, 1948.

[164] H. Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books, 2010.

[165] R. Nithyanand, N. Saxena, G. Tsudik, and E. Uzun. Groupthink: usability of secure group association for wireless devices. In *Ubicomp*, pages 331–340, 2010.

[166] S. J. Ong, D. C. Parkes, A. Rosen, and S. Vadhan. Fairness with an honest minority and a rational majority. In *Theory of Cryptography Conference (TCC)*, 2009.

[167] G. Orwell. *1984*. Editions Underbahn Ltd., 2006.

[168] B. Palanisamy and L. Liu. MobiMix: Protecting location privacy with mix-zones over road networks. In *IEEE Conference on Data Engineering (ICDE)*, 2011.

[169] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, June 2005.

[170] J. Pang, B. Greenstein, R. Gummadi, S. Seshan, and D. Wetherall. 802.11 user fingerprinting. In *MobiCom*, 2007.

[171] P. Papadimitratos, L. Buttyan, T. Holczer, E. Schoch, J. Freudiger, M. Raya, Z. Ma, F. Kargl, A. Kung, and J.P. Hubaux. Secure vehicular communication systems: design and architecture. *IEEE Communications Magazine*, 46(11):100–109, 2008.

[172] P. Papadimitratos, L. Buttyan, J-P. Hubaux, F. Kargl, A. Kung, and M. Raya. Architecture for secure and private vehicular communications. In *IEEE Conference on ITS Telecommunications ITST*, 2007.

[173] P. Papadimitratos, V. Gligor, and J.-P. Hubaux. Securing vehicular communications - assumptions, requirements, and principles. In *Proceedings of Workshop on Embedded Security in Cars (ESCAR)*, 2006.

[174] P. Papadimitratos, A. Kung, J.-P. Hubaux, and F. Kargl. Privacy and identity management for vehicular communication systems: A position paper. In *Proceedings of Workshop on Standards for Privacy in User-Centric Identity Management*, 2006.

[175] D.G. Park, C. Boyd, and S. J. Moon. Forward secrecy and its application to future mobile communications security. In *Public Key Cryptography*, pages 433–445, 2004.

[176] Patently Apple. iGroups: Apple's new iPhone social app in development, 2010. http://www.patentlyapple.com/patently-apple/2010/03/igroups-apples-new-iphone-social-app-in-development.html.

[177] E. Paulos and E. Goodman. The familiar stranger: anxiety, comfort, and play in public places. In *CHI*, pages 223–230, 2004.

[178] A. Pfitzmann and M. Kohntopp. Anonymity, unobservability, and pseudonymity - a proposal for terminology. In *Proceedings of International Workshop on Design Issues in Anonymity and Unobservability*, volume 2009 of *Lecture Notes in Computer Science*, pages 1–9. Springer, 2001.

[179] M. Piorkowski. Sampling urban mobility through on-line repositories of GPS tracks. In *HotPlanet*, 2009.

[180] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. A parsimonious model of mobile partitioned networks with clustering. In *ComsNets*, pages 1–10, 2009.

[181] K. B. Rasmussen and S. Capkun. Implications of radio fingerprinting on the security of sensor networks. In *SecureComm*, 2007.

[182] M. Raya and J.-P. Hubaux. The security of vehicular ad hoc networks. In *SASN*, 2005.

[183] M. Raya and J-P. Hubaux. Securing Vehicular Ad Hoc Networks. *Journal of Computer Security, Special Issue on Security of Ad Hoc and Sensor Networks*, 15(1):39 – 68, 2007.

[184] M. Raya, M. H. Manshaei, M. Felegyhazi, and J.-P. Hubaux. Revocation Games in Ephemeral Networks. In *CCS*, 2008.

[185] M. Raya, P. Papadimitratos, V. D. Gligor, and J.P. Hubaux. On data-centric trust establishment in ephemeral ad hoc networks. In *Infocom*, pages 1238–1246, 2008.

[186] M. Raya, P. Papadimitratos, and J.-P. Hubaux. Securing vehicular communications. In *IEEE Wireless Communications Magazine*, 2006.

[187] D. M. Reeves and M.P. Wellman. Computing best-response strategies in infinite games of incomplete information. In *Uncertainty in artificial intelligence*, pages 470–478, 2004.

[188] R. Rivest, A. Shamir, and Y. Tauman. How to leak a secret. In *ASIACRYPT*, 2001.

[189] rococo. Powering proximity. http://www.rococosoft.com.

[190] C. Roth, S. M. Kang, M. Batty, and M. Barthelemy. Commuting in a polycentric city. Technical report, CNRS, 2010.

[191] Raphael D. Sagarin and T. Taylor. *Natural Security: A Darwinian Approach to a Dangerous World*. University of California Press, 2008.

[192] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. on Knowl. and Data Eng.*, 2001.

[193] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *PODS*, 1998.

[194] D. Samfat, R. Molva, and N. Asokan. Untraceability in mobile networks. In *MobiCom*, pages 26–36, 1995.

[195] K. Sampigethaya, L. Huang, M. Li, R. Poovendran, K. Matsuura, and K. Sezaki. CARAVAN: Providing location privacy for VANET. In *Proceedings of Embedded Security in Cars (ESCAR)*, 2005.

[196] F. Schauer. Fear, risk and the first amendment: Unraveling the chilling effect. *B.U. L. Rev.*, 58:685, 1978.

[197] J. Schiller and A. Voisard. *Location-Based Services*. Morgan Kaufmann Publishers, 2004.

[198] E. Schoch, F. Kargl, T. Leinmuller, S. Schlott, and P. Papadimitratos. Impact of pseudonym changes on geographic routing in VANETs. In *ESAS*, 2006.

[199] A. Serjantov and G. Danezis. Towards an information theoretic metric for anonymity. In *PET*, 2002.

[200] S. Setia, S. Koussih, S. Jajodia, and E. Harder. Kronos: A scalable group re-keying approach for secure multicast. In *IEEE S&P*, pages 215–228, 2002.

[201] C. Shapiro and H. R. Varian. *Information rules: a strategic guide to the network economy*. Harvard Business Press, 1999.

[202] R. Shokri, J. Freudiger, and J.-P. Hubaux. A distortion-based metric for location privacy. In *WPES*, 2009.

[203] R. Shokri, J. Freudiger, and J.-P. Hubaux. A unified framework for location privacy. In *HotPETs*, 2010.

[204] D. Solove. A Taxonomy of Privacy. *University of Pennsylvania Law Review*, 154:477, 2005.

[205] D. Solove. *The Future of Reputation: gossip, rumor, and privacy on the internet*. Yale University Press, 2007.

[206] D. Solove. I've got nothing to hide' and other misunderstandings of privacy. *San Diego Law Review*, 44, 2007.

[207] D. Solove. *Understanding Privacy*. Harvard University Press, 2008.

[208] M. Steiner, G. Tsudik, and M. Waidner. Key agreement in dynamic peer groups. *IEEE Transactions on Parallel and Distributed Systems*, 2000.

[209] Peuple Suisse. Article 13: Protection de la sphère privée. Constitution Fédérale de la Confédération Suisse, 1999.

[210] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10:557–570, 2002.

[211] G. Toth and Z. Hornak. Measuring anonymity in a non-adaptive, real-time system. In *PET*, 2004.

[212] C. Troncoso and G. Danezis. The bayesian analysis of mix networks. In *CCS*, 2009.

[213] G. Tsudik and S. Xu. A flexible framework for secret handshakes. In *PETS*, pages 295–315, 2006.

[214] O. Ureten and N. Serinken. Wireless security through RF fingerprinting. *Canadian J. Elect. Comput. Eng.*, 32, 2007.

[215] G.W. van Blarkom, J.J. Borking, and J.G.E. Olk. *Handbook of Privacy and Privacy-Enhancing Technologies. (The Case of Intelligent Software Agents).* College bescherming persoonsgegevens, 2003.

[216] H. Varian. Economic aspects of personal privacy. White paper, UC Berkeley, 1996.

[217] S. Vasudevan, J. Kurose, and D. Towsley. On neighbor discovery in wireless networks with directional antennas. In *Infocom*, 2005.

[218] M. Vojnovic and J.-Y. Le Boudec. Perfect simulation and stationarity of a class of mobility models. In *Infocom*, 2005.

[219] S. D. Warren and L. D. Brandeis. The right to privacy. *Harvard Law Review*, 4:193, 1890.

[220] A. F. Westin. *Privacy and freedom.* London, 1967.

[221] L. Wittgenstein, P.M.S. Hacker, and J. Schulte. *Philosophical investigations.* Wiley-Blackwell, 2009.

[222] Q. Xu, T. Mak, J. Ko, and R. Sengupta. Vehicle-to-vehicle safety messaging in DSRC. In *VANET*, 2004.

[223] S. Xu. On the security of group communication schemes based on symmetric key cryptosystems. In *SASN*, pages 22–31, 2005.

[224] S. Xu and M. Yung. K-anonymous secret handshakes with reusable credentials. In *CCS*, pages 158–167, 2004.

[225] K. Zeng. Pseudonymous PKI for ubiquitous computing. In *EuroPKI*, pages 207–222, 2006.

[226] P. Zheng. Tradeoffs in certificate revocation schemes. *SIGCOMM Comput. Commun. Rev.*, 33(2):103–112, 2003.

[227] G. Zhong, I. Goldberg, and U. Hengartner. Louis, Lester and Pierre: Three protocols for location privacy. In *PETS*, pages 62–76, 2007.

[228] S. Zhong, L. E. Li, Y. G. Liu, and Y. R. Yang. Privacy-preserving location-based services for mobile users in wireless networks. Technical report, State University of New York at Buffalo, 2005.

Index

Julien Freudiger

Research and Teaching Assistant
EPFL-IC-LCA
Station 14
1015 Lausanne, Switzerland

julien.freudiger@epfl.ch
http://people.epfl.ch/julien.freudiger
+41 21 693 46 68

Personal

Born in Caracas, Venezuela on July 27, 1981. Citizen of Switzerland, France and Venezuela.
Languages: French (native), Spanish (native), English (fluent), German (basic).

Education

Ph.D. in Communication Systems, *Sept. 2006 – Dec. 2010* - EPFL
"When Whereabouts is No Longer Thereabouts: Location Privacy in Wireless Networks"
Advisor: Prof. J.-P. Hubaux

M.Sc. in Communication Systems, *March 2006* - EPFL
Specialization in Compuer Security, *Grade*: 5.9/6.0
Thesis "Performance of VoIP traffic on WCDMA HSUPA" at Qualcomm Inc, CA, USA
Advisors: Prof. J.-Y. Le Boudec and S. Lundby (Qualcomm Inc., CA, USA)

Professional Experience

Research and Teaching Assistant, *Sep. 2006 – present*
School of Computer and Communication Sciences (IC), EPFL

Software Engineer Intern, *Sep. 2005 – Mar. 2006*
Qualcomm Inc., San Diego, CA, USA

Hardware Engineer Intern, *Feb. 2004 – Oct. 2004*
Infineon Technologies, Munich, Germany

Professional Activities

Reviewer for scientific journals and conferences
IEEE Journal on Selected Areas in Communications (JSAC), IEEE TDSC, IEEE TVT, IEEE TMC, WISEC, MobiHoc, MobiCom, INFOCOM, WiOpt, VANET, NetEcon

Technical Program Committee member
PETS 2011

Teaching

Teaching Assistant
Security and Cooperation in Wireless Networks, Prof. J.-P. Hubaux – Fall 07, 08, 09, 10
Mobile Networks, Prof. J.-P. Hubaux – Spring 07, 08, 09, 10
Computer Networks, Prof. J.-P. Hubaux – Fall 08, 09
C++ Programming, Prof. J. Sam – Spring 07

Student Assistant
Principles of Digital Communications, Prof. R. Urbanke – Spring 05
Computer Architecture, Prof. P. Ienne – Fall 03 & Spring 04

Supervised Student Projects

Ahmed Fawaz, *Evaluation of Pseudonym Change Protocols*, Summer 2010
Fabien Dutoit, *Computer Surveillance: a Global Information Recorder*, S'10
Pierre-Alexandre Lombard, *Sniffing and Mining Wireless Communications*, F'09
Selma Chouaki, *Location Privacy amidst Location-Based Services*, F'09
Thibaut Beuret, *Espionnage des Reseaux Sociaux Mobiles*, S'09
Sebastien Epiney, *Real-Time Eavesdropping of Wireless Communications*, S'09
Fabien Dutoit, *Tracage des Cookies par les Publicitaires en Ligne*, F'08
Vincent Salzgeber, *On the Effectiveness of Mix Zones with Realistic Mobility Traces*, F'08
Sylvain Luiset, *Eavesdroping Wireless Communications with Mobile Devices*, F'08
Mahdad Kamal, *Optimal Design of Spectrum Management Using Stackelberg Games*, S'08
Alexandre de Tenorio, *On the Economics of Location Privacy*, S'08
Aurelia Rochat, *Tracking Mobile Nodes in the Web2.0 Era*, S'08

Supervised Master Theses

Antoine Amiguet, *Track People's Habits Online: Data Mining LBSs*, S'10
Raoul Neu, *Private Sharing of User Location over Online Social Networks*, S'10
Marcel Sutter, *Evaluation of Non-Cooperative Location Privacy Protocols*, F'09
Mathias Humbert, *Location Privacy Amidst Local Eavesdroppers*, S'09

Volunteer Experience

Member of I&C Graduate Student Association, 2007 – current
Organized trip to Bolivia with PhD students (meetings with activists), 2009
Speaker at La langue de bois, Radio Frequence Banane, 2009
Organized Beer tasting event, EPFL, 2008
Organized Petanque tournament, EPFL, 2008

Publications

Conferences, Workshops, Book Chapters

1. J. Freudiger, R. Shokri, and J.-P. Hubaux, **Evaluating the Privacy Risk of Location-based Services**, Financial Cryptography and Data Security, March 2011.

2. M. Humbert, M. H. Manshaei, J. Freudiger, and J.-P. Hubaux, **Tracking Games in Mobile Networks**, IEEE GameSec, November 2010.

3. R. Shokri, C. Troncoso, C. Diaz, J. Freudiger, and J.-P. Hubaux, **Unraveling an Old Cloak: k-anonymity for Location Privacy**, ACM WPES, October 2010.

4. N. Vratonjic, J. Freudiger, and J.-P. Hubaux, **Integrity of the Web Content: The Case of Online Advertising**, Usenix CollSec, August 2010.

5. R. Shokri, J. Freudiger, and J.-P. Hubaux, **A Unified Framework for Location Privacy**, HotPETs, July 2010.

6. J. Freudiger, R. Neu, and J.-P. Hubaux, **Private Sharing of User Location over online Social Networks**, HotPETs, July 2010.

7. J. Freudiger, M. H. Manshaei, J.-Y. Le Boudec, and J.-P. Hubaux, **On the Age of Pseudonyms in Mobile Ad Hoc Networks**, IEEE INFOCOM, March 2010.

8. J. Freudiger, M. H. Manshaei, J.-P. Hubaux, and D. C. Parkes, **On Non-Cooperative Location Privacy: A Game-theoretic Approach**, ACM CCS, October 2009.

9. R. Shokri, J. Freudiger, M. Jadliwala, and J.-P. Hubaux, **A Distortion-based Metric for Location Privacy**, ACM WPES, October 2009.

10. J. Freudiger, M. Raya, and J.-P. Hubaux, **Self-Organized Anonymous Authentication in Mobile Ad Hoc Networks**, SecureComm, September 2009.

11. J. Freudiger, R. Shokri, and J.-P. Hubaux, **On the Optimal Placement of Mix-Zones**, PETS, July 2009.

12. J. Freudiger, N. Vratonjic, and J.-P. Hubaux, **Towards Privacy-Friendly Online Advertisement**, IEEE W2SP, May 2009.

13. P. Papadimitratos, L. Buttyan, and T. Holczer, E. Schoch, J. Freudiger, M. Raya, Z. Ma, F. Kargl, A. Kung and J.-P Hubaux, **Secure vehicular communication systems: design and architecture - [topics in automotive networking]**, IEEE Communications Magazine, 2008.

14. M.H. Manshaei, J. Freudiger, M. Félegyházi, P. Marbach and J.-P. Hubaux, **On Wireless Social Communities**, INFOCOM, April 2008.

15. M. H. Manshaei, J. Freudiger, M. Felegyhazi, P. Marbach, and J.-P. Hubaux **Wireless Social Community Networks: A Game-Theoretic Analysis**, Int. Zurich Seminar on Communications (IZS), January 2008.

16. J. Freudiger, M. Raya, M. Félegyházi, P. Papadimitratos, and J.-P. Hubaux, **Mix-Zones for Location Privacy in Vehicular Networks**, Workshop on Wireless Networks for Intelligent Transportation Systems (WiN-ITS), August 2007.

17. M.H. Manshaei, M. Félegyházi, J. Freudiger, J.-P. Hubaux, and P. Marbach, **Spectrum Sharing Games of Network Operators and Cognitive Radios**, in Cognitive Wireless Networks: Concepts, Methodologies and Visions - Inspiring the Age of Enlightenment of Wireless Communications, F. Fitzek & M. Katz Eds. Springer, 2007.

Theses

18. J. Freudiger, **Performance of VoIP traffic on WCDMA HSUPA**, Master Thesis, Qualcomm Inc, San Diego, USA, Mar. 2006.

19. J. Freudiger, **Brain States Analysis for Direct Brain-Computer Communication**, Bachelor Thesis, EPFL, Switzerland, Mar. 2003.

Patents

21. I. Aad, J. Freudiger, M. Jadliwala, J.-P. Hubaux, M. Raya, K. Leppanen, and M. T. Turunen, **Method and Apparatus for triggering user communications based on privacy information**, US Patent Nr.: 12/718521, Mar. 2010.

22. T. Luo, E. Chaponniere, and J. Freudiger, **Message Remapping and Encoding**, US Patent Nr.: 09/666529, Feb. 2007.