Shall I post this now? Optimized, delay-based privacy protection in social networks

Javier Parra-Arnau, Félix Gómez Mármol, David Rebollo-Monedero and Jordi Forné

Abstract—Despite the several advantages commonly attributed to social networks such as easiness and immediacy to communicate with acquaintances and friends, significant privacy threats provoked by unexperienced or even irresponsible users recklessly publishing sensitive material are also noticeable. Yet, a different, but equally significant privacy risk might arise from social networks profiling the online activity of their users based on the timestamp of the interactions between the former and the latter. In order to thwart this last type of commonly neglected attacks, this paper proposes an optimized deferral mechanism for messages in online social networks. Such solution suggests intelligently delaying certain messages posted by end users in social networks in a way that the observed online-activity profile generated by the attacker does not reveal any time-based sensitive information, while preserving the usability of the system. Experimental results as well as a proposed architecture implementing this approach demonstrate the suitability and feasibility of our mechanism.

Index Terms—Time-based profiling, online social networks, privacy-enhancing technology, Shannon's entropy, privacy-utility trade-off.

1 Introduction

Information and communication technologies (ICT) have revolutionized our lives, leading to an unprecedented societal transformation aimed to reach the so called "digital era". In that sense, we are witnessing today how social networks are paving the way to reach such transformation by influencing and even modifying the way we interact with each other and behave amongst us. Amid the plethora of advantages brought by social networks we find the easiness to communicate with friends and acquaintances, the easiness to share thoughts, opinions and experiences in any format (plain text, pictures, audio, video, etc.) and even the immediate reaction in case of emergency or catastrophe.

Yet, despite their proven convenience, online social networks might also pose serious privacy risks [1], most of the times due to irresponsible or unexperienced users who recklessly post private or sensitive information exposing themselves (and sometimes maybe even their friends and connections in the social network) [2], [3] to undesired and unexpected situations (bullying, bribery, identity theft, etc.) [4], [5].

Likewise, an equally significant privacy threat inherent to social networks might also become a burden to the constant increase of their wide deployment and acceptance. However, unlike the previous one, such threat is not based on the content itself published by the end users and, therefore, it might not be as evident as the aforementioned one. Whenever we interact with any social network (post a comment on Facebook, write a message in Twitter, etc.), regardless of the content associated with such interaction, it is reasonably easy for the social network to log a timestamp stating the instant when the interaction occurred. By doing so, the social network is able to build, almost effortlessly, an activity profile of its users based on the timestamps of each of the interactions conducted by such users within the social network.

Profiling users based on their online activity prompts non-negligible privacy concerns. Some examples that illustrate the kind of information that could be inferred from an activity profile include, for instance: when a user normally wakes up and goes to bed, whether a user is unemployed or not, whether they are single or married, and whether they are on holidays or not.

The disclosure of the timing of a message clearly heightens the risk of privacy when considered in the context of additional information obtainable from a user. In combination with geotagging and considering also the contents of the message posted, accurate timing may reveal accurate behavioral patterns, in terms of when and for how long a particular individual does what, and whether these patterns exhibit a particular trend over time. When timing is added to the wealth of data shared across numerous information services, which a privacy attacker could observe and cross-reference, such attacker may more easily infer, even if in a statistical sense, circumstances and trends affecting sensitive aspects of an individual's life, including

- *J. Parra-Arnau, D. Rebollo-Monedero and J. Forné and are with the Department of Telematics Engineering, Universitat Politècnica de Catalunya, C. Jordi Girona 1-3, E-08034 Barcelona, Spain. E-mail: javier.parra,david.rebollo,jforne@entel.upc.edu.*
- *F. Gómez Mármol is with NEC Laboratories Europe, Kurfürsten-Anlage 36, 69115 Heidelberg, Germany. E-mail: felix.gomez-marmol@neclab.eu.*

Manuscript prepared October, 2015.

health status, religious beliefs, social relationships or work performance.

Of special relevance are also the inferences that an attacker may draw when certain background knowledge (e.g., cultural and religious patterns and habits) is available to them. For example, a recent report [6] indicates that, during Ramadan, Facebook and Twitter users in the Middle East are in general most active after iftar time. In both social networks, however, significant differences are observed depending on the country. For instance, Qatar and the Emirates reach peaks of activity just after the iftar, while other countries like Saudi Arabia are most active around midnight[1]. In short, based on this information, the type of attack explored here could undoubtedly help an adversary to ascertain whether a user is Muslim or not, and thus it could seriously compromise their privacy.

With the purpose of thwarting profiling attacks based on the posting times, the paper at hand investigates a data disturbance approach in the form of an optimized message-deferral mechanism. The mechanism under study enables users to delay a number of their messages (without loss of generality, interactions with social networks), hindering an attacker in its efforts to compromise their privacy from their activity profiles. The adversary model assumed in this paper considers an attacker who, based on those profiles, strives to target peculiar users, or said otherwise, users who deviate from the typical, common behavior.

When a user adheres to our mechanism, the profile observed by such attacker (which in our case, as we will see later, is not limited to the social networking site, but broadened to any entity able to collect such timing information), differs from the original, genuine user profile of online activity in such a way that it appears to be much more common and therefore less valuable to the adversary.

The paper is organized as follows: Sec. 2 analyzes general privacy risks and attacks affecting social networks, and examines several privacy-enhancing technologies (PETs) that may help counter time-based profiling attacks. Our optimized deferral mechanism is introduced and described in Sec. 3, while Sec. 4 specifies the building blocks of an architecture implementing our solution. In turn, Sec. 5 studies two specific utility metrics for our approach, namely, expected message delay and messages storage capacity. A comprehensive set of experiments demonstrating the feasibility of our proposal has been conducted and its outcomes are shown in Sec. 6. Finally, Sec. 7 underlines some concluding remarks as well as future research directions.

1. Aggregated Facebook and Twitter activity profiles are shown in [7] per country, during and before Ramadan.

2 State of the Art

In this section, we briefly explore *general* privacy risks and attacks that may occur in online social networks. Then, we review several PETs that could be used to cope with the *specific* time-based profiling attacks illustrated in the previous section.

2.1 Privacy Risks and Attacks in Social Networks

A traditional view on privacy risks and attacks emanates from vulnerabilities in systems presumably protecting confidential data by means of access control policies. These systems may resort to cryptographic protocols implementing services of authentication, access control, confidentiality, and integrity, indispensable when the data to be protected flows across an open medium. A great deal of the vastly abundant literature on cybersecurity concerns such type of traditional security and privacy risks.

Online social networks, a modern, widely popular repository for a wealth of personal, potentially sensitive data, are clearly subject to most forms of traditional risks and attacks, as any other online information system would. Somewhat less obvious is the fact that the particular nature of social networks exposes them to a number of privacy vulnerabilities distinctive of this particular type of online service. In order to better outline the context of the work presented here, in the following, we would like to make a succinct digression on privacy risks and attacks that affect online social networks due to their specific nature, beyond traditionally well-known vulnerabilities universally common to information systems.

Whether those vulnerabilities constitute glaring risks inherent to the mode of operation of the network, or require considerable effort on behalf of an attacker, is often a matter of the level of sophistication of the attack, and the various resources available to the attacker. The quantity and quality of the effort required by an attack is an important pragmatic question addressed in the assumptions adopted in the following descriptions, whose details can be found in the accompanying references.

A fundamental category of privacy attacks distinctively directed against social networks draws upon the principle of *identity theft*. By impersonating a user, an attacker may establish online (friendship) relationships with known registered contacts, in order to gain access to confidential information otherwise restricted to related peers. That information may be about the impersonated user or their contacts. Two variations of this attack are studied in [8], with various degrees of sophistication, possibly involving profile cloning, potentially aggravated by means of *automated crawling* through the online social network, or even across sites, and the automated breaking of CAPTCHA codes. The authors offer empirical evidence of the plausibility

of these attacks in Facebook, StudiVZ, MeinVZ and XING.

Another class of attacks, related to the previous category on identity theft, involves Sybil attacks [9]. In the context of peer-to-peer networks, and other community-based online systems, a *Sybil attack* is an attack wherein a user forges a large number of identities, in order to subvert the underlying trust model or reputation system, and thus gain a disproportionately large influence.

These attacks are relevant in online social networks because they effectively constitute collaborative recommender systems relying on user content ratings, often implemented by means of "like" and "dislike" annotations. Hence, malicious Sybil attackers may outvote honest users in order to alter the suggested relevance of content to better conform with their personal interests, possibly affecting the popularity and reputation of other members of the social network. Mechanisms conceived to counter Sybil attacks in online social networks are explored, for instance, in [10], [11]. The main countermeasure allows the forgery of many identities, but precludes the creation of excessive trust relationships.

A final example of types of privacy attacks specific to online social network encompasses those referred to as *neighborhood attacks* [12], [13]. Under the usual model of an online social network as a graph, with vertices representing users, and edges representing relationships among them, we concordantly define the (1-)neighborhood of an individual as the induced subgraph consisting of all immediately adjacent vertices. Even if the identities of the individuals in the overall graph were purposefully hidden, an attacker with knowledge of the neighborhood subgraph of a known user could still attempt to match the subgraph structure and successfully reidentify the user in question, thus violating the supposed anonymity. Moreover, if several users were reidentified in this manner, knowledge of the anonymized graph would enable this attacker to infer possible direct relationships between them, relationships that may also be construed as confidential information. Strategies to mitigate the effect of those attacks are the subject of the aforementioned work [12], [13].

2.2 Privacy-Enhancing Technologies against Time-based Profiling Attacks

To the best of our knowledge, there is no privacy-enhancing mechanism *specifically conceived* to counter the time-based profiling attack introduced in Sec. 1. In this section, we review some general-purpose technologies that might be adopted to tackle this kind of attacks. Partly inspired by [14], we classify these technologies into three categories: encryption-based methods, approaches based on trusted third parties (TTPs) and data-perturbative techniques.

In traditional approaches to privacy, users or designers decide whether certain sensitive information is to be made available or not. On the one hand, the availability of this data enables certain functionality, e.g., sharing pictures with friends on a social network. On the other hand, its unavailability, traditionally attained by means of access control or encryption, produces the highest level of privacy. In the scenario considered in this work, the use of encryption-based techniques could limit access to the content of the messages posted on a social network, by providing or not a cryptographic key permitting their deciphering. Nevertheless, even though this key was not provided, an attacker with access to the encrypted messages could still be able to jeopardize user privacy — encryption may conceal the content of such messages, but it cannot hide the time instants when they were posted.

A conceptually-simple approach to protect user privacy consists in a TTP acting as an intermediary or *anonymizer* between the user and an untrusted information system. In this scenario, the system cannot know the user ID, but merely the identity of the TTP itself involved in the communication. Alternatively, the TTP may act as a *pseudonymizer* by supplying a pseudonym ID′ to the service provider, but only the TTP knows the correspondence between the pseudonym ID′ and the actual user ID. In online social networks, the use of either approach would be unappropriated as users of these networks are required to be logged in. Although the adoption of TTPs to this end would therefore be ruled out, users themselves could provide a pseudonym at the sign-up process, thus playing the role of a pseudonymizer. In this line, some sites have started offering social-networking services where users are not required to reveal their real identifiers[2].

Unfortunately, none of these approaches may prevent an attacker from profiling a user based on message content, and ultimately inferring their real identity. In its simplest form, reidentification is possible due to the personally identifiable information often included in the messages posted. However, even though no identifying information is included, pseudonyms could also be insufficient to protect both anonymity and privacy. As an example, suppose that an observer has access to certain behavioral patterns of online activity associated with a user, who occasionally discloses their ID, possibly during interactions not involving sensitive data. The same user could attempt to hide under a pseudonym ID′ to exchange information of confidential nature. Nevertheless, if the user exhibited similar behavioral patterns, the unlinkability between ID and ID′ could be compromised through these similar patterns. In this case, any past profiling

2. SocialNumber (http://www.socialnumber.com) is an example of such networks, where users must choose a unique number as identifier.

inferences carried out for the pseudonym ID′ would be linked to the actual user ID.

Another class of PETs relying on trusted entities is anonymous-communication systems (ACSs). In anonymous communications, one of the goals is to conceal who talks to whom against an adversary who observes the inputs and outputs of the anonymous communication channel. Mix systems [15], [16], [17] are a basic building block for implementing anonymous-communication channels. These systems perform cryptographic operations on messages such that it is not possible to correlate their inputs and outputs based on their bit patterns. In addition, mixes delay and reorder messages to hinder the linking of inputs and outputs based on timing information.

In the context of our work, ACSs may hide the link between social networking sites and users, and therefore may protect user privacy against the intermediary entities enabling the communications between them. We may distinguish between two cases — the case where messages are public, and the case where messages are kept private or available to authorized users. In the former case, ACSs obviously cannot provide any privacy guarantees, as user online activity is publicly available. In the latter case, the use of anonymous communications might contribute to privacy enhancement provided that the attacker is not the social-networking site[3].

Among a variety of privacy and threat models that have been proposed for ACSs [18], [19], [20], [21], [22], the important case when the adversary knows all the senders (inputs) and receivers (outputs) would render the anonymous system useless under the time-based profiling attack at hand: it would be enough for this adversary to observe the messages generated by the target user. In other words, under the assumption of an external and global attacker [20], [23], an ACS would not be an appropriate approach to thwart an adversary who strives to profile users based on their online activity.

An alternative to hinder an attacker in its efforts to profile users consists in perturbing the information they disclose when communicating with an information system. The submission of false data, together with the user's genuine data, is an illustrative example of data-perturbative mechanism. In the context of information retrieval, query forgery [24] prevents privacy attackers from profiling users accurately based on the *content* of queries, without having to trust neither the service provider nor the network operator, but obviously at the cost of traffic overhead. A software implementation of query forgery is the Web browser add-on TrackMeNot [25]. This popular add-on exploits RSS feeds and other sources of information to extract keywords, which are then used to generate false queries. The add-on gives users the option to choose how to forward such queries. In particular, a user may send bursts of bogus queries, thus mimicking the way people search, or may submit them at predefined intervals of time.

Clearly, the perturbation of user profiles for privacy protection may be carried out not only by means of the insertion of bogus activity, but also by suppression. An example of this latter kind of perturbation may be found in [26], [27], where the authors propose the elimination of tags as a privacy-enhancing strategy in collaborative-tagging applications. Tag suppression allows users to enhance their privacy to a certain degree, but it comes at the expense of degrading the semantic functionality of those applications, as tags have the purpose of associating meaning with resources.

The data-perturbative mechanisms described above aim to prevent an attacker from profiling users based on their *interests*. Although these mechanisms could also be used to avoid profiling attacks based on the *time instants* when users communicate through social networks, we believe that they would not be adopted in practice — users of social networks would be reticent to eliminate their comments and to generate fake comments, as these actions would have a significant impact on the information-exchange functionality provided by social networks.

3 Privacy Protection via Message Deferral

This section presents the deferral of messages as a PET. The description of this technology is prefaced by a short illustration of time-based profiling attacks in social networks (including a brief explanatory use case), and followed by a succinct introduction of the concepts of soft privacy and hard privacy. Afterwards, we propose a model for representing user activity and describe the assumptions about the privacy attacker assumed in this work. Finally, we define a quantifiable measure of privacy and utility, and present a formulation of the trade-off between these two aspects.

3.1 Illustration of Time-based Profiling Attacks in Online Social Networks

The disclosure of the timing activity of a user may prompt serious privacy concerns, especially when this information is considered in combination with additional data about them. Together with location tagging and the content of the posted messages themselves, the exposure of precise timing activity may uncover behavioral patterns from which a privacy attacker might learn when and for how long a particular individual does what, and where, and whether these patterns show a particular trend over time. When said timing information is added to the data available at other online services such as search engines,

3. Clearly, if the attacker was the social networking platform, any information disclosed by the user would be known to the adversary.

multimedia sharing platforms and e-mail, an attacker that might cross-reference this information may find it easy to ascertain situations and trends affecting several sensitive aspects of a person, including, for example, health status, financial situation, social relationships, work performance, or changes in political preferences. The following use case illustrates the kind of inferences and privacy threats that the sole disclosure of timing information may cause.

3.1.1 Use case: Inference of Religious Beliefs

Isabella Kaya, a student originally from Turkey, has just finished her M.S. degree at the School of Law, University of Texas. Since she was a teenager, our fictional character has been registered with the most popular social networks. Generally she is quite active. In her Twitter and Instagram profiles, her followers can find pictures of her dog and, more recently, comments and congratulations for her graduation. During the Ramadan month, however, her behavior in the networks is altered: Isabella is Muslim and during that period of time, her online activity is notably increased at noon. Due to the fast, she has clearly more opportunities to log into the social networks at that time of the day.

A couple of months ago, Isabella applied for a position in a prestigious law firm. The Department of Human Resources of this firm, similarly to many other companies, often uses social networks to get a glimpse of the candidate outside the confines of a CV, cover letter and interview. Although Isabella posts around 20 messages a day and is aware that firms might snoop on them, she is not worried about a possible invasion of her privacy: she is very reserved and respectful with her comments, and does not have any compromising pictures or nothing blameworthy in her more than 8 years of activity. However, she keeps a constant eye on the comments that others may publish in her profile. Now that she is looking for a job, this control is even stricter.

Isabella had an interview yesterday. Although everything went smoothly, she was surprised by the excessive interest of the interviewer in the origin of her surname. Because she had heard of a few cases of discriminatory practices against the Muslim community by this company, she merely responded her surname was European so as not to reduce her chances of getting the position.

Not satisfied with the response, the interviewer's curiosity could lead him, in a hypothetical case, to examine her profiles in the social networks. Although he would not find any comment that might uncover her religious beliefs, again hypothetically he could confirm his intuition by conducting a basic search on her publicly available social-network profiles. In particular, he could notice that *exactly* from 18 June to 17 July (period of the last Ramadan) Isabella's online activity follows a distinct, characteristic pattern, and observe that this same behavior is exhibited *precisely* during the Ramadan month of the previous year (from 29 June to 28 July), and the one from two years ago (from 9 July to 8 August), and so it goes on for the last 8 years of activity, all available at her public Twitter and Facebook accounts. Also hypothetically, this could be the reason why she did not get the job in the end.

3.2 Soft Privacy and Hard Privacy

The privacy research literature [28] recognizes the distinction between the concepts of *soft privacy* and *hard privacy*. In a soft-privacy model, users entrust an external entity or TTP to safeguard their privacy. That is, users put their trust in an entity which will hereafter be in charge of protecting their private data.

In the literature, numerous attempts to protect user privacy have followed the traditional method of anonymous communications, which is based on the suppositions of soft privacy. Additional examples of PETs building on this model are anonymizers and pseudonymizers. The main drawbacks of all these technologies, as we commented in Sec. 2, are that they come at the cost of infrastructure and are not completely effective [29], [30], [31], [32]. Besides, even in those cases where we could fully trust in the effectiveness of an entity, that entity could be legally enforced to reveal the information it has access to [33]. The AOL search data scandal [34] is another example that shows that the trust relationship between users and TTPs may be broken. In short, whether privacy is preserved or not under this model depends on the trustworthiness of the data controller and its capacity to manage the entrusted data.

On the other extreme is the hard-privacy model, where users mistrust any communicating entity and thus endeavor to reveal as little private information as possible. In the application scenario at hand, hard privacy means that users need not trust an external entity such as the social networking provider or the network operator. Mechanisms providing hard-privacy guarantees primarily rely on data perturbation and operate on the user side. An archetypal example is TrackMeNot, a Web browser extension installed on the user's machine that aims at perturbing their Web search profile through the submission of false queries. As we shall see next, the privacy-preserving technology proposed here leans on this model.

3.3 Message Deferral

In the introductory section, we emphasized the risk of profiling based on the time instants when users submit messages to a social networking site. In particular, we mentioned that, building on this online behavior, an adversary could extract an accurate snapshot of their profiles of activity throughout time and thus could compromise user privacy.

In this situation, we propose a data disturbance approach consisting of the *deferral of messages* as a conceptually-simple mechanism that may thwart this kind of profiling attacks. The proposed mechanism allows users to delay the submission of certain messages, by storing them locally and afterwards sending them to the social-network provider in question. The application of this mechanism may help users protect their privacy to a certain extent, at the cost of no infrastructure, and without having to trust neither the service provider nor any other external entity. Since privacy protection takes place exclusively on the user side, our mechanism contributes to the principle of *data minimization* [4] and avoids any potential leakage by external privacy systems, social-networking sites, Internet service providers (ISPs), proxies, routers and other networking entities. In a nutshell, it provides hard-privacy guarantees, meaning that the protection offered by the mechanism is robust in the presence of untrusted or not fully trusted external entities like the above might be.

Delaying messages may therefore allow certain privacy protection, but this inevitably comes at the expense of data-storage capacity and, more importantly, the utility of the services provided by the online social network. As an example, consider a user posting a tweet[5] to confirm a meeting this evening. If this tweet was postponed, the confirmation could arrive late and, if so, the information-exchange functionality would be useless. In short, the deferral of messages poses a trade-off between the contrasting aspects of privacy on the one hand, and utility on the other. Fig. 1 shows a conceptual depiction of our mechanism.

In the coming sections, we shall investigate the deferral of messages as a technique that may preserve users' privacy against an attacker who tries to profile them based on their posting times. Note that this is in contrast to other types of profiling attacks that exploit the *content* of the information disclosed, rather than the *time* when this information is revealed.

Naturally, this latter kind of user profiling may occur in conjunction with the former, but the degree of sophistication and computational efforts are presumably much higher for the former type of attacks, i.e., those that capitalize on content information. Mainly for this reason, online social networking services and microblogging services like Twitter and Facebook are more prone to time-based profiling. In these information systems, an attacker would have to analyze the content of posts, where, in addition to text, users often include images and videos. Processing all these

4. According to [35], the data-minimization principle means that a data controller, e.g., the social-networking platform, should restrict the collection of personal data to what is strictly necessary to achieve its purpose. Also, it implies that the controller should store the data only for as long as is necessary to fulfil the purpose for which the information was collected.

5. A tweet is a message sent using Twitter.

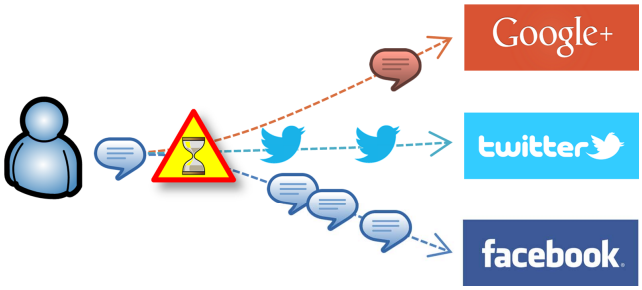


Fig. 1: Message deferral as a mechanism to protect the privacy of the online activity of a user by delaying the submission of certain messages.

data and extracting features from them would require far more computational efforts[6] than simply retrieving the timestamp field of those posts. A Web application that exemplifies the ease with which time-based profiles can be built is [36].

Despite the potential occurrence of these time-based profiling attacks and the evident privacy risks they entail, we acknowledge that, within the context of *certain* social-networking applications, users may not be willing to tolerate a degradation of the intended functionality due to message deferral. This is the case, for example, of real-time conversations, which may not be particularly conducive to our privacy mechanism. We believe, however, that many other uses of the social networks may allow it.

3.4 Adversary Model

In order to evaluate the level of privacy provided by our mechanism, it is fundamental to specify the concrete assumptions about the attacker, that is, its capabilities, properties or powers. This is known as the *adversary model* and its importance lies in the fact that the level of privacy provided is measured with respect to it.

Next, we describe the adversary model assumed in this work, in terms of the application scenario considered, the type of adversaries able to profile users, the way these adversaries model user activity, and the objective behind the construction of these activity models.

- **Scenario**. First, we consider a typical scenario where users are required to be logged into a social networking site for their messages to be posted. This could be the case of Google Plus, Twitter and Facebook. In addition, we may reasonably assume that users of these applications provide their real identifiers to create their accounts. We must hasten to stress that, even though a user employs pseudonyms, the content of the messages exchanged or the knowledge of their "friends" in those social networks may lead an attacker to ascertain the actual identity of this user.

6. This is in contrast to other information systems where user data (e.g., tags, queries or ratings) are simpler to process.

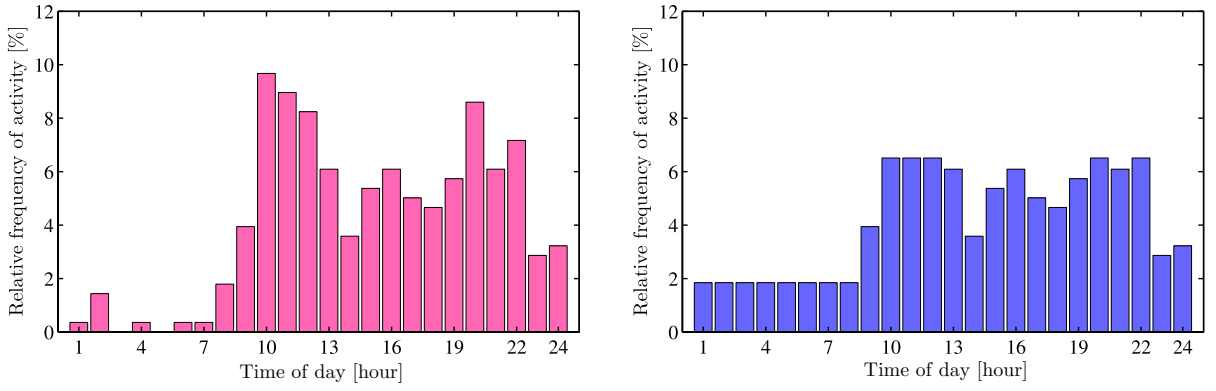


Fig. 2: Actual user profile (a) and apparent user profile (b). Both profiles represent the profile of activity across a day, in particular, the percentage of messages posted between 0 a.m. and 1 a.m., 1 a.m. and 2 a.m., and so on.

- **Privacy attackers**. In this scenario, any entity capable of capturing users' messages is regarded as a potential privacy attacker. This includes the social network provider, the Internet service provider (ISP), and the intermediary entities (switches, routers, firewalls) enabling the communications between users and social networking sites. Besides, since posted messages are often publicly available[7], any entity able to collect this information is also taken into consideration in our adversary model.
- **User-profile model**. We assume that the attacker represents behavioral patterns of online user activity as probability mass functions (PMFs). Conceptually, a user profile may be interpreted as a histogram of relative frequencies of messages across a day, week, month or year. The proposed user-profile model is a natural, intuitive representation in line with the models used in many information systems to characterize user profiles [27], [37], [38], [39], [40].

  In our adversary model, we distinguish between two kinds of profiles. On the one hand, the user's genuine profile, and on the other, the profile perceived from the outside, which results from delaying certain messages before posting them. Hereafter, we shall refer to these two profiles as the *actual* profile q and the *apparent* profile t. That said, in this work we shall assume that the attacker is unaware or ignores the fact that the observed, perturbed profile does not reflect the actual behavior of the user. Fig. 2 provides an example of such profiles. In this figure we represent the profile of online activity of a user within 1-hour slot throughout one day.
- **Objective behind profiling**. Finally, our adversary model contemplates what the attacker is after when profiling users. According to [40], and in line with the technical literature of profiling [41], [42], we assume that the attacker's ultimate goal is to target peculiar users. Put differently, we consider an adversary that aims to find users who deviate significantly from the average and common activity profile.

The goal of profiling, together with the assumptions about the scenario and the user-profile representation, constitute the adversary model upon which our privacy metric builds.

3.5 Privacy Metric of Online Activity

Next, we justify the Shannon entropy and the Kullback-Leibler (KL) divergence as measures of privacy when an attacker aims to target uncommon users based on their profiles of activity. The rationale behind the use of these two information-theoretic quantities as privacy metrics is documented in greater detail in [40].

Recall that Shannon's entropy $\mathrm{H}(t)$ of a discrete random variable (r.v.) with PMF $t = (t_i)_{i=1}^n$ on the alphabet $\{1, \ldots, n\}$ is a measure of the uncertainty of the outcome of this r.v., defined as

$$\mathrm{H}(t) = -\sum t_i \log t_i.$$

Throughout this work, all logarithms are taken to base 2, and subsequently the entropy units are *bits*. Given two probability distributions t and p over the same alphabet, the KL divergence is defined as

$$\mathrm{D}(t \,\|\, p) = \sum t_i \log \frac{t_i}{p_i}.$$

The KL divergence is often referred to as *relative entropy*, as it may be regarded as a generalization of the Shannon entropy of a distribution, relative to another. Conversely, Shannon's entropy is a special case of KL divergence, as for a uniform distribution u on a finite alphabet of cardinality n,

$$\mathrm{D}(t \,\|\, u) = \log n - \mathrm{H}(t). \tag{1}$$

Leveraging on a celebrated information-theoretic rationale by Jaynes [43], the Shannon entropy of an

7. Messages exchanged on Twitter are publicly visible by default.

apparent user profile, modeled as a PMF, may be regarded as a measure of privacy, or more accurately, anonymity. The leading idea is that the method of types [44] from information theory establishes an approximate monotonic relationship between the likelihood of a PMF in a stochastic system and its entropy. Loosely speaking, the higher the entropy of a profile, the more likely it is that the more users behave according to it. Under this interpretation, entropy is a measure of anonymity, *not* in the sense that the user's identity remains unknown, but only in the sense that higher likelihood of an apparent profile, believed by an external observer to be the actual profile, makes that profile more common, hopefully helping the user go unnoticed, less interesting to an attacker whose objective is to seek peculiar users.

If an aggregated histogram of the whole population of users were available as a reference profile p, the extension of Jaynes' argument to relative entropy would also give an acceptable measure of anonymity. Recall that KL divergence is a measure of discrepancy between probability distributions, which includes Shannon's entropy as the special case when the reference distribution is uniform. Conceptually, a lower KL divergence hides discrepancies with respect to a reference profile, say the population's, and there also exists a monotonic relationship between the likelihood of a distribution and its divergence with respect to the reference distribution of choice, which enables us to deem KL divergence as a measure of anonymity in a sense entirely analogous to the above mentioned.

Under this interpretation, the Shannon entropy is therefore interpreted as an indicator of the *commonness* of similar profiles. As such, Shannon's entropy appears as a meaningful anonymity measure since it effectively captures the attacker's goal behind profiling. We should hasten to stress that the Shannon entropy is a measure of *anonymity* rather than privacy, in the sense that the obfuscated information is the uniqueness of the profile behind the online activity, rather than the actual profile itself.

3.6 Formulation of the Trade-Off between Privacy and Message-Deferral Rate

In this section, we present a formulation of the optimal privacy-utility trade-off posed by our message-deferral mechanism.

In our mathematical model, we represent the messages of a user as a sequence of independent and identically distributed (i.i.d.) r.v.'s taking on values in a common finite alphabet of n time periods, namely the set $\{1, \dots, n\}$ for some integer $n \geqslant 2$. As an example, the set of time periods could be the hours of a day or a week, or the days of a month. According to this model, we characterize the *actual* profile of a user as the common PMF of these r.v.'s, $q = (q_1, \dots, q_n)$. In conceptual terms, our model of user profile is a normalized histogram of messages over those time periods.

Based on this model, we quantify the *initial* privacy level as the Shannon entropy of the user's actual profile, $\mathrm{H}(q)$. For the sake of tractability, we measure utility as the *deferral rate* $\varphi \in [0, 1)$, that is, the ratio of the number of messages that a user is willing to delay to the total number of messages.

When a user accepts delaying their tweets, comments or, in general, messages, their actual profile q is seen from the outside as the apparent profile $t = q - s + r$, according to a *storing strategy* s and a *forwarding strategy* r. These strategies are two n-tuples that would tell the user when to retain those messages and when to release them. More specifically, the i-th component of the storing strategy is the fraction of messages that this user should store at time period i. Similarly, r_i is the proportion of messages to total number of messages that the user should forward at time i. Clearly, these two strategies must satisfy that $s_i, r_i \geqslant 0$, $q_i - s_i + r_i \geqslant 0$, for all i, and that $\sum s_i = \sum r_i = \varphi$ so that t is a PMF.

According to this notation, we denote by $\mathrm{H}(t)$ the (*final*) privacy level and define the *privacy-deferral function* as

$$\mathcal{P}(\varphi) = \max_{\substack{r,s \\ r_i \geqslant 0,\, s_i \geqslant 0, \\ q_i - s_i + r_i \geqslant 0, \\ \sum s_i = \sum r_i = \varphi}} \mathrm{H}(q - s + r), \tag{2}$$

which models the optimal trade-off between privacy and message-deferral rate.

The optimization problem inherent in this definition belongs to the extensively studied class of convex optimization problems [45]. Most of these problems do not have an analytical solution and thus need to be solved numerically. For this, there exist a number of extremely efficient methods, such as interior-point algorithms. The problem formulated here, however, turns out to be a particular case of a more general optimization problem, for which interestingly there is an explicit closed-form solution, albeit piecewise [46].

In practice, this means that we shall be able to find an analytical expression for the *optimal* storing and forwarding strategies, i.e., those strategies that maximize user privacy for a given φ. Later on, in Sec. 5.1, we shall show that (2) is a particularization of this latter problem.

4 Architecture

As we commented on Sec. 3.2, our PET leverages on the hard-privacy model. In essence, this means that users seek to safeguard their privacy themselves, since any communicating entity (e.g., the network provider, social-networking platform, ISP) may be regarded as a potential attacker. Because our mechanism takes

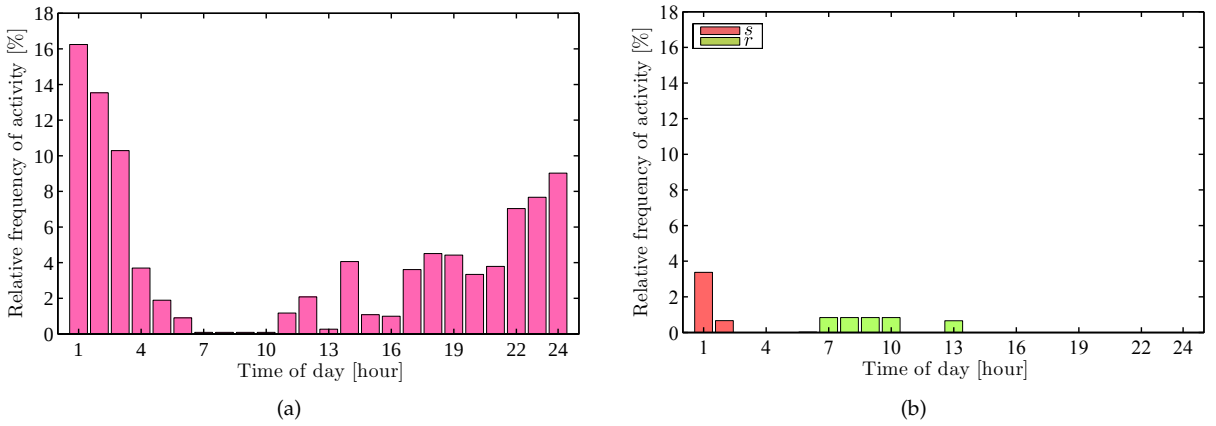


Fig. 3: Example of user profile (a) and its optimal storing and forwarding strategies r and s (b) for a deferral-message rate $\varphi = 4\%$.

place on their side and therefore does not rely on any external party, it offers hard-privacy protection.

In this section, we specify the building blocks of an architecture implementing our privacy-enhancing, message-deferral mechanism. As we shall see later, the system architecture revolves around a module that computes the optimal forwarding and storing strategies from (2). In essence, the proposed system will employ these two strategies to distort the actual profile in a way that user privacy is maximized. We would like to stress that, since our data-perturbative mechanism is optimized for any message-deferral rate, any perturbation introduced in the actual profile will always be in the direction of providing a better privacy protection. In other words, and in contrast to randomized perturbative mechanisms, deviations from the actual profile caused by our mechanism *always* guarantee an improvement in privacy.

The architecture proposed in this section provides high-level functional aspects so that our PET can be implemented as software running on the user's local machine, for example, in the form of a Web-browser extension. Specifically, our architecture builds on the aforementioned hard-privacy model, which implies that users need not trust any external entity to protect their privacy. We only assume, however, that users trust the piece of software that implements our mechanism, in terms of the data it collects and its execution, exactly as they trust their Web browser.

Our assumptions about the proposed architecture are described next:

- First, we assume that both the user and the adversary use the same time periods, for example, 24 uniformly distributed time slots within a day. This implies that the profile computed on the user's side coincides with the profile built by the attacker.
- Secondly, according to equation (2), our approach needs the user's actual profile q to compute the optimal storing and forwarding strategies. Because of this, we contemplate a training period before our architecture starts delaying messages. However, since the attacker might learn about the user profile during this training period, the user could alternatively provide the software with an estimate of their profile.
- Lastly, we suppose that, in the estimation of the relative histogram, the components of the user profile remain stable after the training phase. We acknowledge, however, that a practical implementation of our mechanism should take into account that the user activity may vary significantly over time.

Before we proceed with the description of our architecture, we shall provide an example showing what the optimal storing and forwarding strategies mean in practice. For this, consider the profile q depicted in Fig. 3(a), which corresponds to a user with initial privacy risk $\mathcal{P}(0) \simeq 4.2775$ bits. If this user decided to delay $\varphi = 4\%$ of their messages, the relative privacy gain would be around 5.18%. That is, in this particular case we observe that the privacy gain would be, interestingly, greater than the delay rate introduced.

The optimal strategies are illustrated in Fig. 3(b). The storing strategy suggests buffering 3.37% and 0.63% of messages at time instants 1 and 2, respectively[8]. On the other hand, the forwarding strategy recommends extracting 0.84% of the total number of messages from the buffer at time periods 7, 8, 9 and 10, and 0.64% of the messages at time 13.

In Fig. 4 we depict the proposed architecture, which consists of a number of modules, each of them performing a specific task. From a general perspective, this figure shows a user interacting with a social networking site, an entity that basically stores the messages generated by this and other users. Next, we provide a functional description of the modules of this architecture.

- **User-profile constructor**. It is responsible for the estimation of the user's profile. Specifically, this module receives the messages the user generates,

8. Those time instants are, in fact, time periods of one hour each. In particular, the time index i consists in the interval $(i-1, i]$.

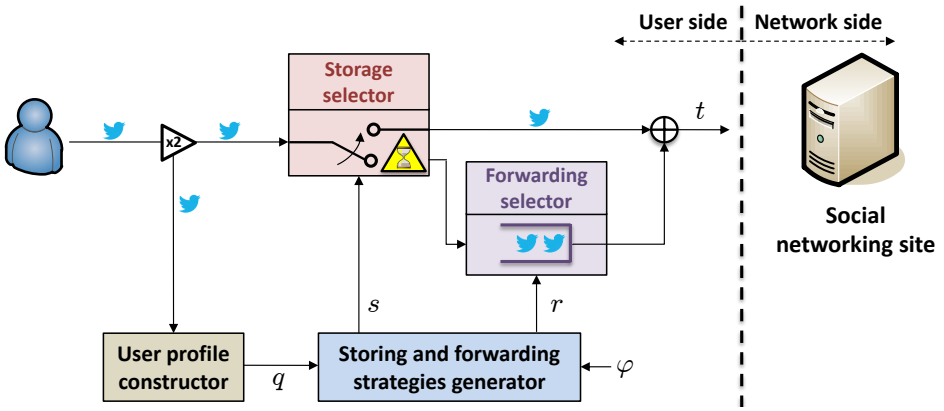


Fig. 4: Architecture implementing the message-deferral mechanism.

and computes a histogram of relative frequencies of these messages within, for example, 1-hour slot throughout one day. Afterwards, this profile is submitted to the *storing and forwarding strategies generator*.

We would like to emphasize that this module is active even when the user explicitly declares their profile. Since the profile specified by the user may not be an accurate reflection of their online behavior, our architecture may decide, after the training phase, to replace it with the profile implicitly inferred from the posted messages.

- **Storing and forwarding strategies generator**. This module is the core of the architecture as it is responsible for computing the solution to the optimization problem inherent in function (2). To this end, this component is first provided with the user profile and the message-deferral rate. Secondly, the module uses this information to compute the optimal tuples of storing and forwarding; and finally, those tuples are given to the *storage selector* module and to the *forwarding selector* block.
- **Storage selector**. The functionality of this module is to warn the user when they should delay messages[9]. Specifically, at time period i, with probability s_i/q_i the user should send a message to the buffer implemented in the *forwarding selector* module. On the other hand, with probability $1 - s_i/q_i$, this message should be submitted directly to the social networking site.
- **Forwarding selector**. This block includes a buffer where messages are stored. Its main functionality is to output messages from this buffer according to the optimal forwarding strategy r. In particular, this module would operate as follows: throughout time slot i, the module would send $\alpha\, r_i$ messages from the buffer to the service provider, where α represents the total number of messages generated within the time period covered by the profile, e.g., one day.

  This block also considers the possibility of assigning priorities to messages. For instance, it could be necessary that certain messages stored in the buffer have different levels of priority. As an example, those messages generated during working hours could have a higher likelihood of leaving the buffer. Other alternatives include first in, first out (FIFO), last in, first out (LIFO) and uniformly-random extraction. This last option is precisely the one considered in Sec. 5.

9. This would be, in fact, transparent to the user. The software installed on the user's machine would decide whether a message is to be delayed or not.

5 Expected Delay and Message-Storage Capacity

In Sec. 3.6 we characterized the optimal privacy-utility trade-off posed by message deferral, in terms of the Shannon entropy of the apparent profile as measure of privacy, and the message-deferral rate as measure of utility. In that same subsection, we also mentioned that the optimization problem characterizing this trade-off is a particular case of a more general optimization problem for which there exists a closed-form solution. Although this allows us to obtain analytically our optimal storing and forwarding strategies for a given deferral rate, users would certainly benefit from more meaningful metrics of loss in usability than this fraction of messages delayed. In other words, it would be interesting and even necessary to investigate more elaborate and informative utility measures, capturing the actual impact that our mechanism would have.

Motivated by this, in this section we examine more sophisticated metrics such as the expected delay experienced by messages and the capacity of the buffer where these messages are stored. Further, we investigate how they relate each other, under the premise that messages are output from the buffer uniformly at random, that is, without considering any kind of priority such as FIFO or LIFO.

This section is structured as follows. First, Sec. 5.1 examines some interesting results derived from the more general optimization problem examined in [46].

Then, Sec. 5.2 presents a mathematical analysis modeling the utility metrics mentioned above, namely, the expected delay and buffer capacity. Finally, Sec. 5.3 provides an example illustrating the theoretical results obtained in the previous subsection.

5.1 Preliminaries

The optimization problem investigated in [46] is a resource allocation problem that arises in the context of privacy protection in recommendation systems. In the cited work, the authors model the privacy-utility trade-off posed by a data-perturbative mechanism consisting in the forgery and the elimination of ratings. Specifically, the privacy risk $\mathcal{R}$ is measured as the KL divergence between the apparent profile of *interests*[10] and the population's distribution of items p. On the other hand, the loss in accuracy of recommendations is measured as the percentages of ratings ρ and σ that the user would be willing to forge and suppress, respectively. Accordingly, the optimal trade-off between privacy and utility is defined as

$$\mathcal{R}(\rho,\sigma) = \min_{\substack{r,s \\ r_i\geqslant 0,\, s_i\geqslant 0, \\ q_i - s_i + r_i \geqslant 0, \\ \sum s_i = \sigma,\, \sum r_i = \rho}} \mathrm{D}\left(\frac{q-s+r}{1-\sigma+\rho}\,\middle\|\, p\right), \quad (3)$$

where the optimization variables are a *forgery strategy* r and a *suppression strategy* s.

In light of this formulation, it is straightforward to check, by virtue of (1), that

$$\mathcal{P}(\varphi) = \log n - \mathcal{R}(\varphi,\varphi)|_{p=u}.$$

In words, the function (2) characterizing the trade-off between privacy and message-deferral rate is a special case of the optimization problem (3), when the rates of forgery and suppression are equal to φ and the population's distribution is the uniform distribution. In the context of our formulation, the forgery and suppression strategies clearly correspond to the forwarding and storing strategies, respectively.

Having shown then that (2) is a particular case of (3), next we review a couple of results presented in [46] to be used in the coming sections.

The most relevant result is the intuitive principle that the optimal storing and forwarding strategies follow. Specifically, the former strategy lowers the highest values of q_i until these values are equal. This is done in such a way that the values lowered amount to φ. In a completely analogous manner, the latter strategy raises the lowest values of q_i until they match, for a total probability mass increment of φ. Finally, intermediate values of q_i remain unperturbed. Simply put, the effect of the optimal strategies on the actual user profile may be regarded as a combination of the well-known water-filling and reverse water-filling problems [45, §5.5].

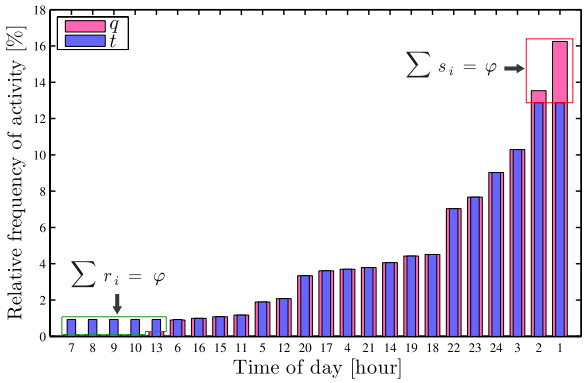


Fig. 5: This figure illustrates the intuition behind the optimal storing and forwarding strategies. Here, we have represented the actual user profile q depicted in Fig. 3(a). The optimal apparent profile t is obtained by applying the strategies shown in Fig. 3(b), which correspond to a deferral rate $\varphi = 0.04$.

The aforementioned principle was already anticipated in Fig. 3. In Fig. 5 we illustrate this more clearly. Particularly, this figure depicts the actual user profile shown in Fig.3(a) and its optimal apparent profile, resulting from the application of the optimal storing and forwarding strategies represented in Fig. 3(b). In Fig. 5, however, the components of those two profiles are sorted in increasing order of activity to emphasize the way these strategies operate.

Another interesting result from [46] confirms the intuition that there must exist a pair (ρ,σ) such that the privacy risk vanishes. In the context of our formulation, this implies that there is a deferral rate φ beyond which the maximum level of privacy or *critical privacy* is attained[11]. We refer to this rate as the *critical message-deferral rate* φ_{crit}.

Recall [44] that the *variational distance* between two PMFs p and q is defined as

$$\mathrm{TV}(p\,\|\,q) = \tfrac{1}{2}\sum_i |p_i - q_i|\,.$$

It can be shown [46] that the critical rate yields

$$\varphi_{\text{crit}} = \mathrm{TV}(u\,\|\,q). \quad (4)$$

From this expression, it is easy to verify that $\varphi_{\text{crit}} \geqslant 0$, with equality if, and only if, $q = u$. Later on, in Sec. 6, we shall determine the average critical rate within a population of Twitter, Facebook and Instagram users, as well as the PMF of this crucial parameter.

The last result is related to the orthogonality of the components of s and r. Specifically, it follows from [46] that, for any $\varphi \leqslant \varphi_{\text{crit}}$, the optimal storing and forwarding strategies satisfy

$$s_k\, r_k = 0,$$

10. Here users' profiles do not capture their interests, but their online activity.

11. Recall from Sec. 3.5 that Shannon's entropy is regarded here as a measure of privacy *gain*, whereas the KL divergence is interpreted as a measure of privacy *risk*.

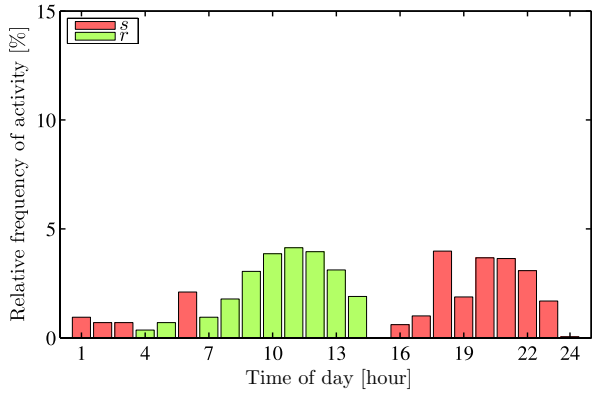


Fig. 6: Example of optimal storing and forwarding strategies which do not satisfy the principle of causality.

for $k = 1, \ldots, n$. The orthogonality of both strategies, in the sense indicated above, conforms to intuition—it would not make any sense to store messages in a given time period and, at the same time period, forward messages to the social networking server. This result is implicitly assumed throughout next subsection, Sec. 5.2.

5.2 Theoretical Analysis

Denote by $s = (s_1, \ldots, s_n)$ and $r = (r_1, \ldots, r_n)$ the solutions to the problem (2), conceptually, a storing strategy and a forwarding strategy that maximize the Shannon entropy of the apparent profile, $\mathrm{H}(t)$. Recall that these two tuples must satisfy

$$\sum s_i = \sum r_i = \varphi. \tag{5}$$

In Fig. 3(b) we depicted an example of these tuples. In that figure, the time instants when messages were stored were preceding the time instants when these messages were forwarded. That is, the figure showed the logical sequence in which messages are first kept in the buffer and then they are flushed out.

However, the solutions s and r do not need to satisfy this principle of causality; this was not specified as a constraint in the optimization problem (2). In fact, regardless of whether causality is satisfied or not, these two tuples must be interpreted as cyclic sequences, which are repeated continuously, e.g., every day or week, depending on the time frame covered by the user profile. This is how the storing and the forwarding strategies must be construed then in Fig. 6. Here, although no messages are forwarded at the time instants 1, 2 and 3 of the first cycle (day), in subsequent cycles these time instants will be used to output messages.

In the remainder of this section, we shall mathematically model the buffer. Specifically, we shall find a time instant such that, if the tuples s, r are moved to start at this instant, then every message to be forwarded during the next n consecutive time periods will actually be forwarded. This is time period 7 in the example shown in Fig. 6. With this time index, we shall be able to proceed to find an expression for the expected delay.

Denote by a_i the n consecutive permutations of the tuple $s_i - r_i$,

$$\begin{aligned} a_1 &= (\, s_1 - r_1\,, s_2 - r_2\,, \quad \ldots \quad , \quad s_n - r_n \quad), \\ a_2 &= (\, s_2 - r_2\,, \quad \ldots \quad , s_n - r_n\,, \quad s_1 - r_1 \quad), \\ &\vdots \\ a_n &= (s_n - r_n\,, s_1 - r_1\,, \quad \ldots \quad , s_{n-1} - r_{n-1}). \end{aligned}$$

The j-th element of the tuple a_i is denoted by $a_{i,j}$. Associated with each tuple a_i, define the sequence $(b_{i,j})_{j=1}^{\infty}$ as

$$b_{i,j} = \begin{cases} \max\{a_{i,j}, 0\}, & j = 1 \\ \max\{b_{i,j-1} + a_{i,k}, 0\}, & j = 2, 3, \ldots \end{cases},$$

where $k = j - n\lfloor \frac{j-1}{n} \rfloor$ indexes cyclically the tuple a_i. Conceptually, each sequence b_i models the ratio of messages to total number of messages that are stored in the buffer over the time index j, when the optimal storing and forwarding strategies are applied cyclically starting from the time index i.

Recall that the Heaviside step function [47] of a discrete variable x is defined as

$$\theta(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geqslant 0 \end{cases}.$$

Our first result demonstrates that, after a transient state and regardless of the starting index i, these sequences converge to a common, repeated pattern. As we show next, this is a consequence of (5).

Lemma 1:

(i) There exists some index $i = 1, \ldots, n$ such that $b_{i,n} = 0$.

(ii) Let i be an index satisfying $b_{i+1,n} = 0$. Then, for any $j = 1, \ldots, n$, there exists an index $l = i + 1 - j + n\,\theta(j - i - 1)$ such that

$$b_{j,k+l} = b_{i+1,k}, \text{ for } k = 1, 2, \ldots$$

Proof: Define the n cumulative sums $w_{i,j} = \sum_{k=1}^{j} a_{i+1,k}$ for $i = 1, \ldots, n-1$, and $w_{i,j} = \sum_{k=1}^{j} a_{i-n+1,k}$ for $i = n$, where the index j ranges from 1 to n. Note that $w_{i,j} = w_{n,i+j} - w_{n,i}$, for $i + j \leqslant n$; when $i + j > n$, substitute $i + j$ for $i + j - n$. Let $i \in \{1, \ldots, n\}$ be an index such that $w_{n,i}$ is minimal. Then, it immediately follows that $w_{i,j} \geqslant 0$ for $j = 1, \ldots, n$, which implies, by virtue of (5), that $b_{i+1,n} = 0$. Note that this holds also for the index $i = n$, for which $b_{1,n} = 0$. This proves statement (i).

We have showed that the index i that minimizes $w_{n,k}$ for all k satisfies $b_{i+1,n} = 0$. To prove (ii), first we shall show that $b_{j,l} = 0$ for all j. To this end, replace the index j with $j + 1$ in statement (ii), so that now $l = i - j + n\,\theta(j - i)$ and j goes from 0 to $n - 1$. Recall also that $w_{i,j} = w_{n,i+j} - w_{n,i}$ for $i + j \leqslant n$, and $w_{i,j} = w_{n,i+j-n} - w_{n,i}$ for $i + j > n$. With the previous

change of variable, note that $w_{j,l} = w_{n,i} - w_{n,j}$. Here, for consistency with the indexes of $w_{i,j}$, we substitute $j = 0$ for $j = n$. Having said this, observe that, for a given j

$$\min_k w_{j,k} = w_{n,i} - w_{n,j} = w_{j,l},$$

which clearly is nonpositive. Then, fix j and note that the set of possible values that $b_{j+1,k}$ may take on are $w_{j,k}$ plus the k terms $w_{j,k} - w_{j,m}$ for $m = 1, \ldots, k$. Since $\min_k w_{j,k} = w_{j,l} \leqslant 0$, it follows that $b_{j+1,l} = 0$. To conclude the proof, simply observe that $a_{j+1,l+1} = a_{i+1,1}$. ■

Hereafter we shall refer to the *starting index* i as the index satisfying $b_{i,n} = 0$. Since Lemma 1 shows that all sequences converge to a steady state where a pattern is repeated continuously, our analysis is restricted to the finite sequence $(b_{i,j})_{j=1}^{n}$ modeling this pattern and its corresponding tuple a_i.

Let C be the capacity of the buffer and α the total number of messages generated by the user throughout the considered time frame (a day, week, month, etc.). The next result gives a straightforward expression for C when the steady state is achieved.

Corollary 2: Let i be the starting index. In the steady state, the buffer capacity is

$$C = \alpha \max_{j \in \{1,\ldots,n\}} b_{i,j}.$$

Proof: It is immediate from the definition of $b_{i,j}$ and Lemma 1. ■

Next, we shall reorder the tuples s, r so that they begin at the starting index. Denote by s',r' the tuples starting with this index i, formally

$$s' = (s'_1, s'_2, \ldots, s'_n),$$
$$r' = (r'_1, r'_2, \ldots, r'_n),$$

where $s'_k = s_l$ and $r'_k = r_l$ with $l = i+k-1-n\lfloor \frac{i+k-2}{n} \rfloor$. Note that, when we reorder the storing and forwarding tuples this way, for every $r'_j > 0$ we can forward exactly r'_j messages at time period j.

In the following we define some notation that will be used in Theorem 3. Let Δ be an r.v. representing the number of time periods a message is delayed. Note that, on account of Lemma 1, the buffer does not retain any message for more than n time units. Consequently, the alphabet of Δ is the set $\{1, \ldots, n\}$. Denote by $\bar{\delta}$ its expected value, $\mathrm{E}\,\Delta$. Let D be a Bernoulli r.v. of parameter φ, modeling whether a message is delayed or not. Namely, $\mathrm{P}\{D = 1\} = \varphi$ is the probability that a message is delayed and $\mathrm{P}\{D = 0\} = 1 - \varphi$ is the probability it is not. Finally, define the set

$$\omega(j, k) = \{l : r'_l > 0, k < l < j\}.$$

Our next result, Theorem 3, provides a closed-form expression to calculate the expected delay in the steady state.

Theorem 3: Let i be an index satisfying $b_{i,n} = 0$. Then,

$$\bar{\delta} = \sum_{\delta=1}^{n} \delta \sum_{\substack{j-k=\delta, \\ r'_j, s'_j > 0}} \frac{r'_j\, s'_k}{b_{i,j-1}} \prod_{l \in \omega(j,k)} \left(1 - \frac{r'_l}{b_{i,l-1}}\right).$$

Proof: From Lemma 1, we know that all sequences b_k for $k = 1, \ldots, n$ converge to the finite sequence $(b_{i,j})_{j=1}^{n}$. Note that $\mathrm{E}\,\Delta = \mathrm{E}\,\mathrm{E}[\Delta|D] = \varphi\,\mathrm{E}_{\Delta|D}[\Delta|D = 1]$. Next, we proceed to calculate the conditional PMF $p_{\Delta|D}(\delta|1)$. Let A be an r.v. representing the time instant when a message arrives at the buffer, and L, the time instant when this message leaves the buffer. Accordingly,

$$\mathrm{P}\{\Delta = \delta | D = 1\} = \sum_{j-k=\delta} \mathrm{P}\{L = j | A = k\}\, \mathrm{P}\{A = k | D = 1\}.$$

Observe that $\mathrm{P}\{A = k|D = 1\} = s'_k/\varphi$. Further, note that $\mathrm{P}\{L = j|A = k\}$ is the probability that a message is not forwarded at the time instants $\omega(j, k)$, that is, $\prod_{l \in \omega(j,k)} \left(1 - \frac{r'_l}{b_{i,l-1}}\right)$, multiplied by the probability that this message is forwarded at time j, that is, $\frac{r'_j}{b_{i,j-1}}$. From this, it is immediate to derive the expression given in the statement of the theorem. ■

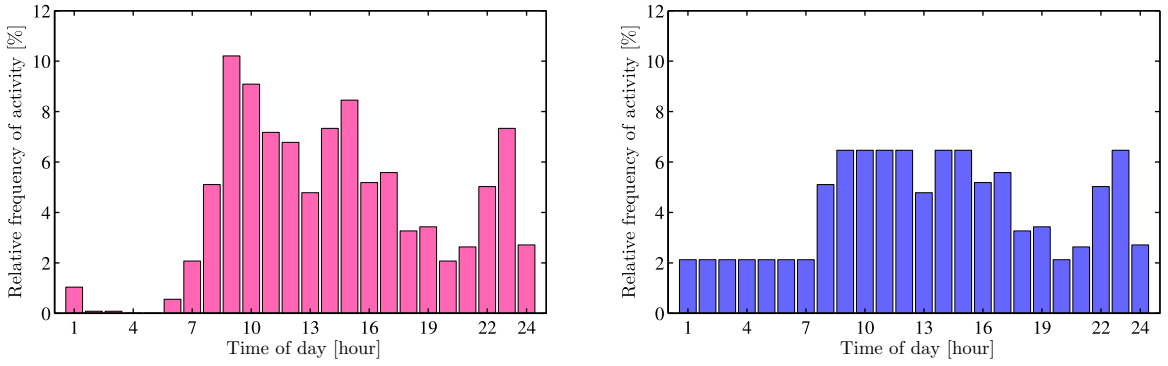
The expression obtained in Theorem 3 allows us therefore to estimate the expected delay that messages will experience for a given deferral rate. Although at first sight it may seem there is not a direct dependence on the parameter φ, recall that s and r are related to this parameter through (5).

In conclusion, the results provided in this subsection enable us to establish a connection between the message-deferral rate, i.e., our simplified, but mathematically tractable measure of utility, and more elaborate and informative utility metrics such as the expected delay and the message-storage capacity.

5.3 Numerical Example

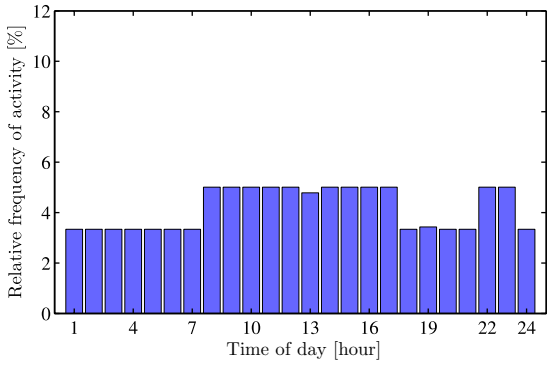
This subsection presents a numerical example that illustrates the analysis conducted in the previous subsection, and shows the privacy level achieved by a user who adheres to the proposed message-deferral mechanism. Throughout this subsection, all results correspond to the same user.

In Fig. 7 we represent the apparent profile of this user for different values of the message-deferral rate φ. When $\varphi = 0$, no perturbation takes place and the apparent profile t represented in Fig. 7(a) actually corresponds to the genuine user profile q. According to the reasoning behind the optimal storing and forwarding strategies described in Sec. 5.1, the higher φ, the more uniform is the resulting apparent profile. The maximum level of privacy is attained precisely for $\varphi = \varphi_{\text{crit}} \simeq 0.3206$, when the apparent profile is

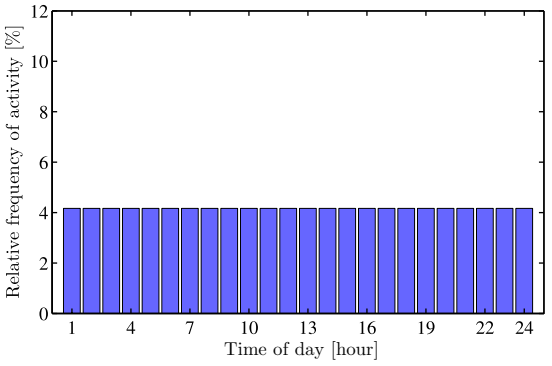


(a) $\varphi = 0$, $\mathrm{H}(t) \simeq 4.1060$ bits.

(b) $\varphi \simeq 0.1111$, $\mathrm{H}(t) \simeq 4.4402$ bits.



(c) $\varphi \simeq 0.2222$, $\mathrm{H}(t) \simeq 4.5567$ bits.



(d) $\varphi \simeq 0.3206$, $\mathrm{H}(t) \simeq 4.5850$ bits.

Fig. 7: Apparent profiles for different values of φ.

completely uniform and therefore $\mathrm{H}(t) = \log 24 \simeq 4.5850$ bits. All this information is also captured in Fig. 8, where we plot the privacy-deferral function (2), that is, the function modeling the optimal trade-off between privacy and utility, the latter being measured as the percentage of messages delayed.

In Fig. 9(a) we depicted the expected delay $\bar{\delta}$ for different values of φ. In particular, the results shown in this figure were computed theoretically, by applying Theorem 3, and experimentally. These latter experimental results were obtained by simulating the storing and forwarding processes as specified by the blocks *storage selector* and *forwarding selector* of the proposed architecture (see Sec. 4). Fig. 9(a) tells us, for example, that for $\varphi = 0.10$, the messages delayed were kept on the buffer for around 1.5 hours on average. As expected, for $\varphi \leqslant \varphi_{\text{crit}}$, we observe that $\bar{\delta}$ exhibits an increasing, nonlinear behavior with φ. The case when $\varphi \geqslant \varphi_{\text{crit}}$ is of no interest as, in practice, a user would not delay more messages than those strictly necessary to achieve the maximum level of privacy.

Finally, Fig. 9(b) shows, for different values of φ, the ratio between the number of messages stored in the buffer and the total number of messages generated by the user. For instance, when the user specifies $\varphi = 0.10$, the buffer must be designed to keep around 10% of all messages sent over a day. Clearly, we note that the buffer capacity is nonlinear with the deferral rate. Also, we observe that the user would need to store 28.1% of their messages for the apparent profile to become the uniform distribution.

6 Experimental Analysis

In this section we evaluate the extent to which the deferral of messages could enhance user privacy in a real-world scenario. The social networks chosen to conduct this evaluation are Twitter, Facebook and Instagram. With this experimental evaluation, we aim at demonstrating the technical feasibility of our scheme, and the benefits it would bring to both users and social networking sites.

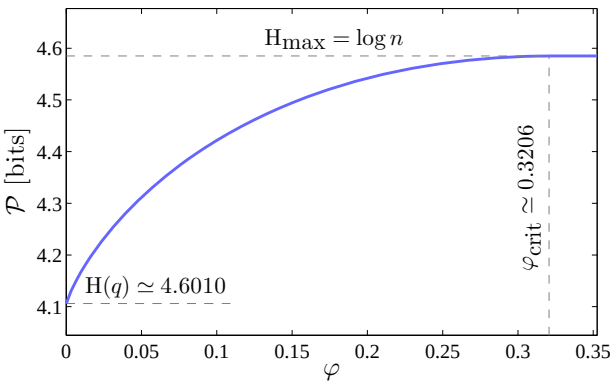


Fig. 8: Optimal trade-off between privacy and utility, the latter being measured as the message-deferral rate.

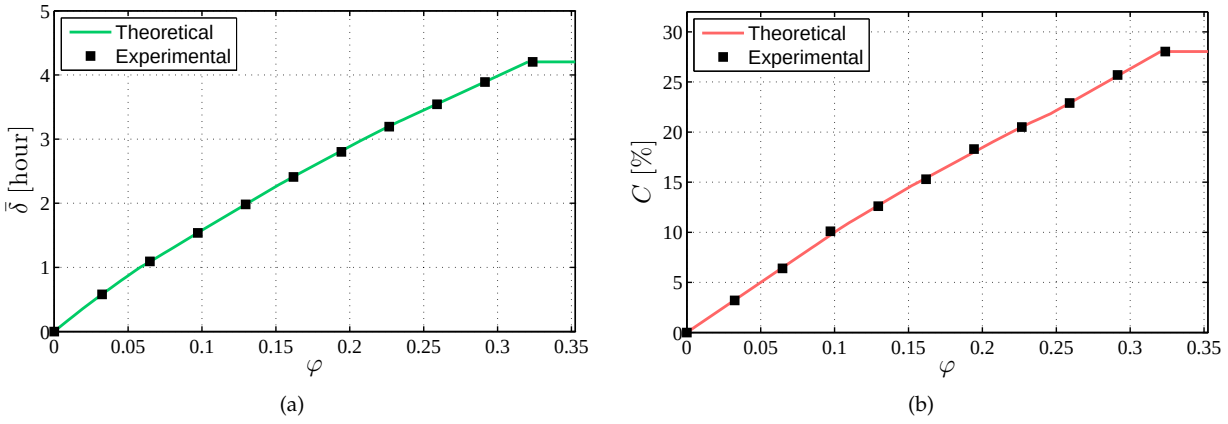


Fig. 9: Expected delay (a) and buffer capacity (b) for different values of the message-deferral rate. The buffer requirements are expressed in relative terms, compared to the user's activity.

TABLE 1: Summary of the data sets used in our experimental evaluation.

	Number of users	Average activity (messages)
Twitter	144	1 879.42
Facebook	529	208.40
Instagram	610	397.27

6.1 Data Sets

Our analysis has been conducted on the basis of three different data sets. In the case of Twitter, we employed 144 users, whose profiles were retrieved by using the Twitter API[12]. In particular, we gathered the timestamps of all public messages generated by those users before Oct. 25, 2013. From this information, we built their profiles as normalized histograms of tweets across 24 uniformly distributed time slots within one day. Previously, we filtered out those users with an activity level lower than 50 posts, since it would have been difficult to calculate a reliable estimate of their profiles with such a few messages. On average, users posted 1 879.42 messages each.

In the case of Facebook and Instagram, on the other hand, we relied on publicly available data sets. In the former social network, we experimented with the data set retrieved by the Software Systems group at the Max Planck Institute [48]. The data set in question contains wall posts from the Facebook New Orleans networks. As with the Twitter data, we eliminated those users with a low activity profile. The number of users and posts reduced to 529 and 110 243 respectively, yielding an average of 208.40 messages per user. Finally, as for Instagram, we used the data set available at [49] which included the comments posted to media by more than 2 100 users. After applying the preprocessing described above, the number of users became 610, posting 397.27 messages on average. Table 1 provides a summary of the data sets utilized in this experimental section.

12. https://dev.twitter.com

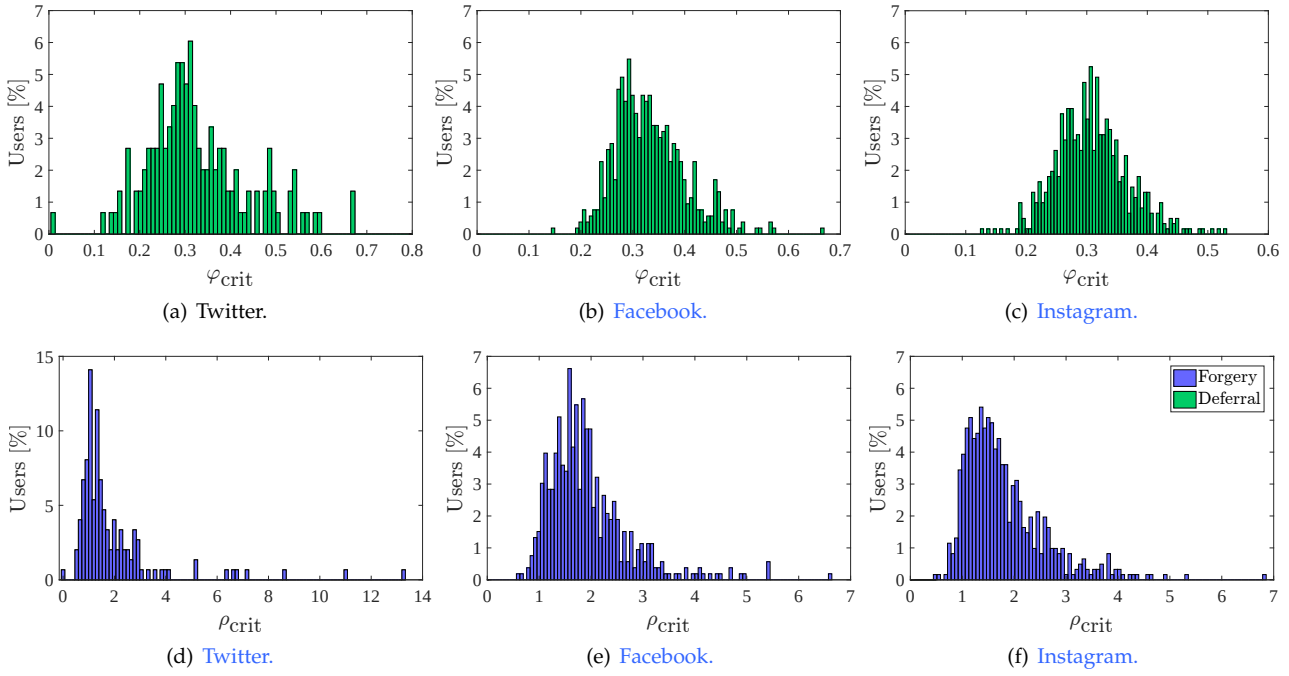
6.2 Privacy Technologies

In these experiments, we shall compare the proposed message-deferral strategy with two privacy mechanisms that pertain to the category of data perturbation. These mechanisms are *message forgery* and *uniform deferral*.

Information forgery is a rather common strategy for privacy protection, which has been studied in the context of several applications, from information retrieval [24], to anonymous-communication systems [15], to Web browsing [25], and to recommendation systems [46]. In this experimental evaluation, we shall consider the submission of false messages to counter time-based profiling attacks in social networks, although, as commented in Sec. 2.2, this strategy might not be adopted in real practice: users of these online services might not be disposed to post fake comments, since this might have a significant negative effect on the functionality provided by such services.

We shall use the notation introduced in Sec. 5.1 for rating forgery in recommendation systems, and accordingly denote by $\rho \in [0, \infty)$ the ratio of false messages to total posted messages. For the sake of a fair comparison, we shall consider an *optimized* version of this message-forgery mechanism, in the sense of maximizing the same privacy objective (Shannon's entropy) for a given ρ. It can be straightforwardly shown that the problem of optimal message forgery is equivalent to the problem (3) for $\sigma = 0$. Finally, as with message deferral, ρ_{crit} will denote the forgery rate beyond which the maximum level of privacy is achieved.

On the other hand, our comparative analysis will include a naive deferral mechanism that will serve as a baseline assessment. In particular, this mechanism will rely on the same message-deferral rate φ, but the delay experienced by each message will be drawn from a uniform distribution on the set $\{1, \ldots, 24\}$. It can be shown that this naive technique is equivalent to assuming $r = \varphi u$ and $s = \varphi q$, or considering the



(a) Twitter. (b) Facebook. (c) Instagram.

(d) Twitter. (e) Facebook. (f) Instagram.

Fig. 10: Probability distribution of the critical rates of optimized deferral and forgery for the three data sets considered in Sec. 6.

convex combination $t = (1-\varphi)\,q + \varphi\,u$. By virtue of these equivalences, it is also straightforward to verify that uniform deferral attains critical privacy if, and only if, $\varphi = 1$. This is obviously in the non-trivial case when $q \neq u$. Throughout this section, we shall occasionally refer this strategy as "random" deferral.

6.3 Results

In our first series of experiments, we computed the probability distribution of φ_{crit} and ρ_{crit}, that is, the deferral and forgery rates beyond which the maximum privacy level is attained[13]. The PMFs of such critical rates are shown in Fig. 10. In the case of optimized message deferral and the Twitter data set, we observe that the minimum, maximum and average values of φ_{crit} are approximately 0.01, 0.67 and 0.33. Also, we spot that a significant mass of probability is concentrated between $\varphi \simeq 0.2$ and $\varphi \simeq 0.4$, in particular, a 74% of users. This means that most users will not require delaying a large percentage of their tweets for their apparent profiles to become the uniform distribution.

Similar results are observed for the other two data sets. In the case of Facebook, for example, the critical deferral rate has an average value of 0.33, slightly smaller than for Twitter, but the minimum value is 0.14. As for Instagram, the results are a bit better: the maximum and average values of φ_{crit} are 0.53 and 0.30, respectively.

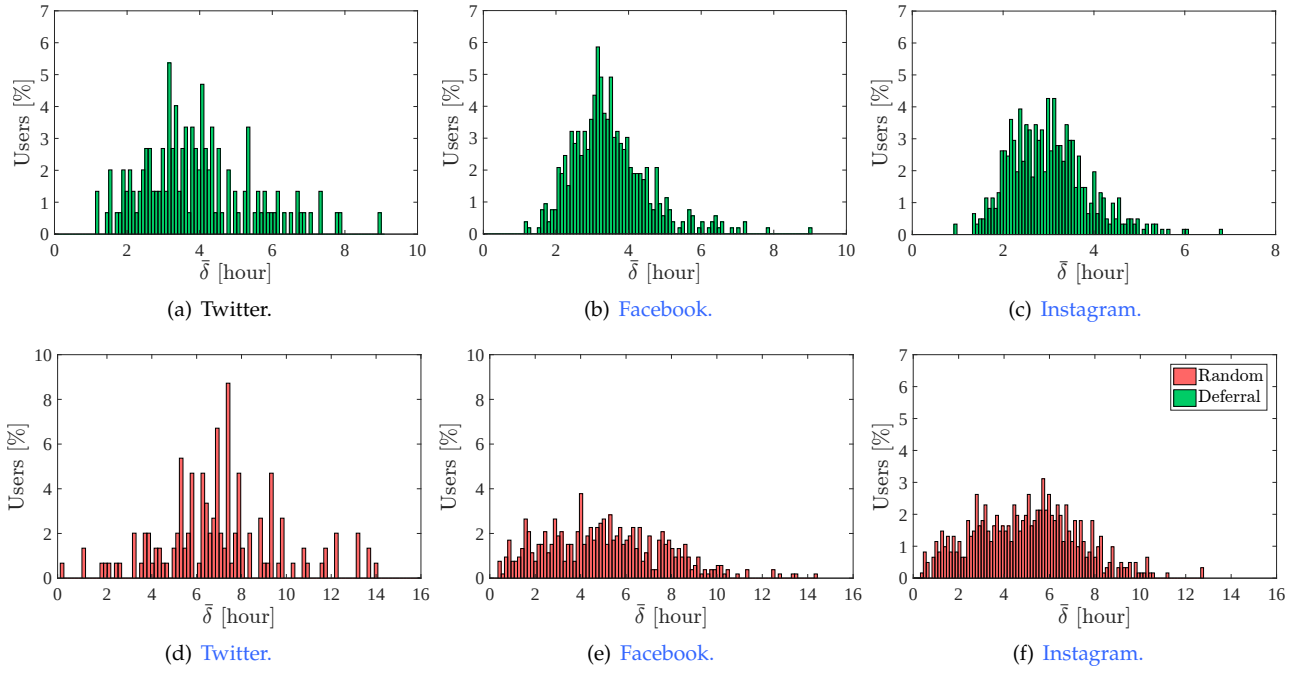
13. We omit the distribution of the critical-deferral rate for the uniform strategy since, as commented in Sec. 6.2, this strategy achieves critical privacy only when $\varphi = 1$. Consequently, the PMF of the critical rate is the trivial Dirac delta function centered at 1.

The distributions of the critical forgery rate are plotted in Figs. 10(d-f). The main conclusion that can be drawn from these figures is that users will need large values of ρ, in most cases above 100%, for their privacy to attain the maximum level. The results for Twitter, Facebook and Instagram yield an average rate of 1.921, 1.932 and 1.794, respectively. This is in contrast to the critical rate of the proposed deferral mechanism, which by definition cannot exceed 1.

The following figure, Fig. 11, shows the PMFs of the expected delay for our optimized deferral mechanism and for the uniform deferral strategy set forth in Sec. 6.2. The results are plotted in the case when all users apply these two mechanisms with the critical deferral rate of our optimized technology, given by (4). The results provided for optimized deferral has been obtained analytically by using the expressions derived in Sec. 5.2.

From Figs. 11(a-c), we check that the values of $\bar{\delta}|_{\varphi=\varphi_{\text{crit}}}$ are roughly concentrated between 1 and 8 hours. In the case of Twitter, however, the probability distribution seems to be more dispersed. The average values for this social network, Facebook and Instagram are 3.898, 3.474 and 2.996, respectively, which means that Instagram users will experience smaller average delays for the same level of (maximum) privacy protection.

In the case of "random" deferral (Figs. 11(d-f)), not entirely unexpectedly we spot more scattered distributions of $\bar{\delta}|_{\varphi=\varphi_{\text{crit}}}$. For example, in the three data sets considered in these experiments, we notice expected delays of up to 14 hours, whereas the maximum value provided by our optimized deferral strategy was 9.05 hours. In addition, we observe that the mean values



(a) Twitter. (b) Facebook. (c) Instagram.

(d) Twitter. (e) Facebook. (f) Instagram.

Fig. 11: PMFs of the expected delay when all users apply a deferral rate $\varphi = \varphi_{\text{crit}}$, for the optimized deferral strategy proposed in this work and for the naive "random" delay mechanism described in Sec. 6.2.

for the Twitter, Facebook and Instagram data sets are significantly greater than those exhibited by our mechanism. In particular, these mean values show an increase of 67.8% (Twitter), 24.5% (Facebook) and 66.3% (Instagram) with respect to optimized deferral.

Fig. 12 shows the buffer capacity for the optimized deferral mechanism and for the uniform strategy described in Sec. 6.2. Analogously to Fig. 11, these results have been obtained under the assumption that users choose a deferral rate $\varphi = \varphi_{\text{crit}}$ as given by (4).

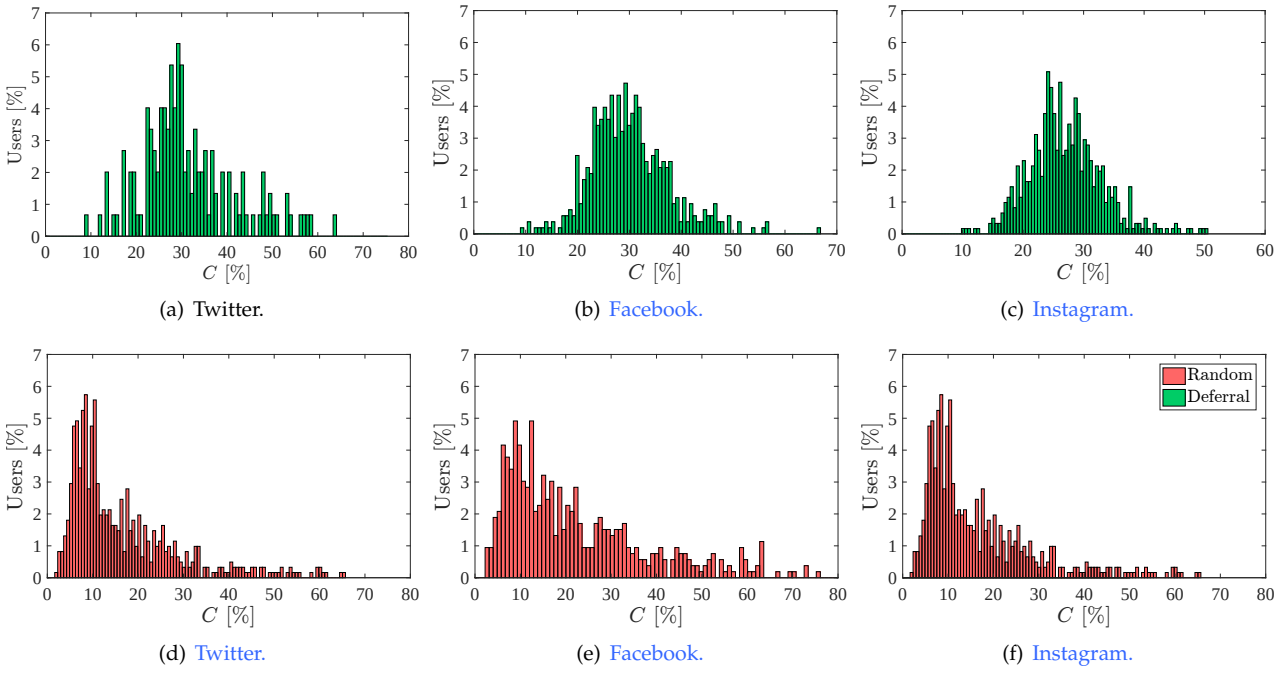
From Fig. 12(a), we notice that the minimum, mean and maximum values of $C|_{\varphi=\varphi_{\text{crit}}}$ are 8.92, 31.24 and 63.52% of Twitter users' messages. Similar results are observed for the other two data sets. In the case of Facebook and Instagram, though, we notice slightly smaller mean values of capacity. In particular, Figs. 12(b-c) show an expected buffer size of 28.31% and 27.75% of users' messages, respectively.

In the case of a uniform delay strategy, we observe users with buffer capacities around 1% for Twitter, and approximately 3% and 2% for Facebook and Instagram. This is in stark contrast to the deferral mechanism investigated in this work, which, according to these experiments, requires a minimum of 10% of message-storage capacity to attain the critical privacy. We note, however, that these smaller values of capacity (observed for uniform deferral) do not imply that users will achieve the maximum level of privacy. In fact, as we commented in Sec. 6.2, the naive deferral strategy achieves critical privacy if, and only if, $\varphi = 1$. Finally, we notice that the mean values of capacity for the Twitter, Facebook and Instagram data sets are 17.7%, 18.6% and 11.6% of user messages.

The second set of experiments contemplates a scenario where all users apply the three privacy-enhancing mechanisms under study, by using a common message deferral and forgery rate. Note that, in practice, each user would configure this rate independently, according to their specific privacy and utility requirements. Under the assumption of a common rate, Fig. 13 shows the privacy protection achieved by those users in terms of percentile curves (10^{th}, 50^{th} and 90^{th}) of relative privacy gain. In the case of optimized deferral and forgery, these results have been obtained by applying the closed-form expression for the optimal storing and forwarding strategies derived in [46]. Specifically, we computed the optimal strategies of each user for 100 uniformly distributed values of $\varphi, \rho \in [0, 0.999]$.

We start our analysis of this figure with optimized deferral and the Twitter data set. In Fig. 13(a), we observe how the percentile curves of relative privacy gain increase with φ until a certain rate, beyond which these curves are constant. This is consistent with the fact that users attain the maximum level of privacy, $\log n$, for $\varphi \geqslant \varphi_{\text{crit}}$. An interesting conclusion that can be drawn from this figure is that Twitter will require relatively small margins of privacy gain to achieve the critical-privacy level. This may be observed, for example, for $\varphi = 0.60$, i.e., when almost all users get their maximum level of privacy, according to Fig. 10(a). Concretely, for this value of φ, the 10^{th}, 50^{th} and 90^{th} percentile curves show privacy gains of only 4.59%, 10.78% and 27.60%, respectively.

When the strategy is to post false messages, rather than delaying them, we observe percentile curves with a lower rate of increase than for optimized



(a) Twitter. (b) Facebook. (c) Instagram.

(d) Twitter. (e) Facebook. (f) Instagram.

Fig. 12: Buffer capacity for different values of the message-deferral rate, and for the optimized and uniform deferral strategies. The buffer requirements are expressed in relative terms, compared to users' activity.

deferral. For example, while the 90$^{\text{th}}$ percentile curve attains its maximum value, 28.41%, for $\varphi \simeq 0.49$, message forgery does not provide this level of protection even for $\rho = 0.999$ (see Fig. 13(b). A similar behavior is observed for the 10$^{\text{th}}$ and 50$^{\text{th}}$ percentile curves.

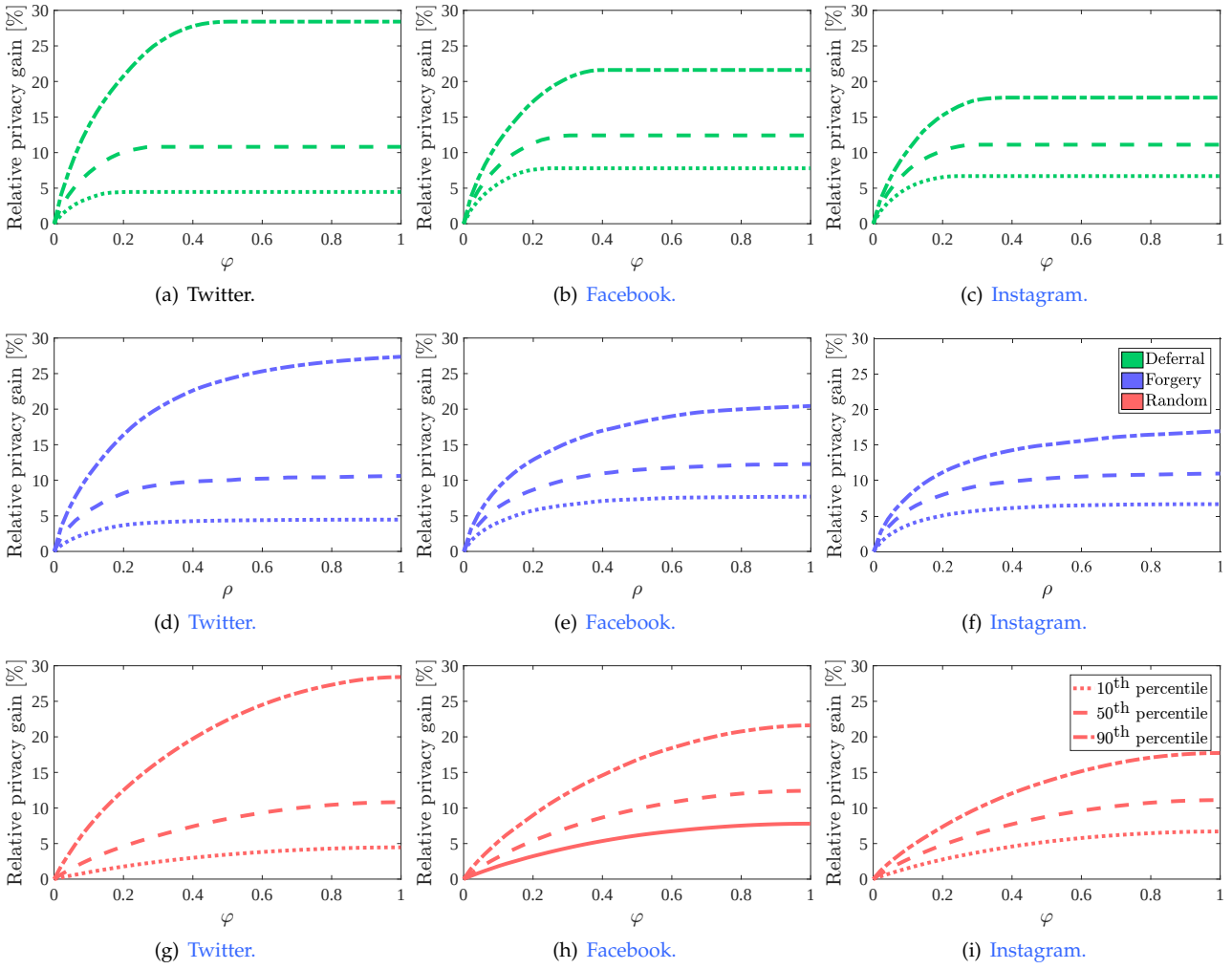
In the special case of uniform deferral, we notice that for values of φ smaller than 0.69 approximately, the 90$^{\text{th}}$ percentile curve is lower than that of message forgery. However, for $\varphi > 0.69$, the trend is reversed and users safeguard their privacy more efficiently by applying uniformly distributed delays. This, though, should come as no surprise, as according to Figs. 10(d-f) message forgery exhibits an average $\rho_{\text{crit}} > 1.794$ in the three data sets. In other words, users applying uniform deferral attain higher values of privacy gain for large perturbation rates, when compared to forgery.

Similar conclusions can be derived from the Facebook and Instagram data sets, with the main result being that optimized deferral again outperforms forgery and uniform delay. From Fig. 13(b,e,h), we observe that for $\varphi = 0.39$, 90% of Facebook users obtain a relative privacy gain greater than 21.6%. For an identical value of forgery rate, the submission of false messages by that same fraction of users would increase their privacy by at least 16.8%, almost 5 percentage points below optimized deferral. In the case of uniform deferral, this difference is accentuated for that deferral rate; we see 2.5 percentage points below forgery. However, for $\varphi > 0.71$, the 90$^{\text{th}}$ percentile curve surpasses that of message forgery.

On the other hand, the differences in relative privacy gain between the three data sets can be explained on the basis of the initial privacy values. The fact that we have smaller values of privacy gain for Instagram users is solely because these users have more flattened profiles than those of Facebook and Twitter. In particular, the average initial privacy (i.e., when $\varphi = \rho = 0$) is 4.0230, 4.0410 and 4.1067 bits for the users of Twitter, Facebook and Instagram, respectively.

The upshot of this analysis of the three technologies in terms of critical rate, delay, capacity and privacy gain, is that our PET strategy offers better privacy guarantees for any of the utility metrics considered in these experiments. For a given φ, the uniform strategy may lead to smaller expected delays and buffer capacities, but obviously the level of privacy attained is not comparable to that of optimized deferral and forgery. This result is true only for forgery rates roughly on the interval [0,0.7]. For larger values of φ, uniform deferral is more effective in protecting user privacy than forgery. As mentioned above, this is due to the fact that the forgery mechanism requires large rates of false messages, compared to uniform deferral ($\varphi_{\text{crit}} = 1$) and optimized deferral ($\varphi_{\text{crit}} \in [0.01, 0.67]$ from Figs. 10(a-c)).

Having examined the impact of our privacy mechanism on message delay, capacity and user privacy, now we look at the effect it might have from the point of view of traffic load. Recall that the objective of message deferral is to maximize the Shannon entropy of the apparent profile and thus to spread user activity uniformly over time. This is obviously beneficial from the standpoint of user privacy, as we have observed in our previous series of experiments. But at the same time, entropy maximization may help social networking sites manage their networking resources more



(a) Twitter. (b) Facebook. (c) Instagram.

(d) Twitter. (e) Facebook. (f) Instagram.

(g) Twitter. (h) Facebook. (i) Instagram.

Fig. 13: Percentile curves of relative privacy gain for different values of φ, for the three privacy technologies examined in these experiments, and for our Twitter, Facebook and Instagram data sets.

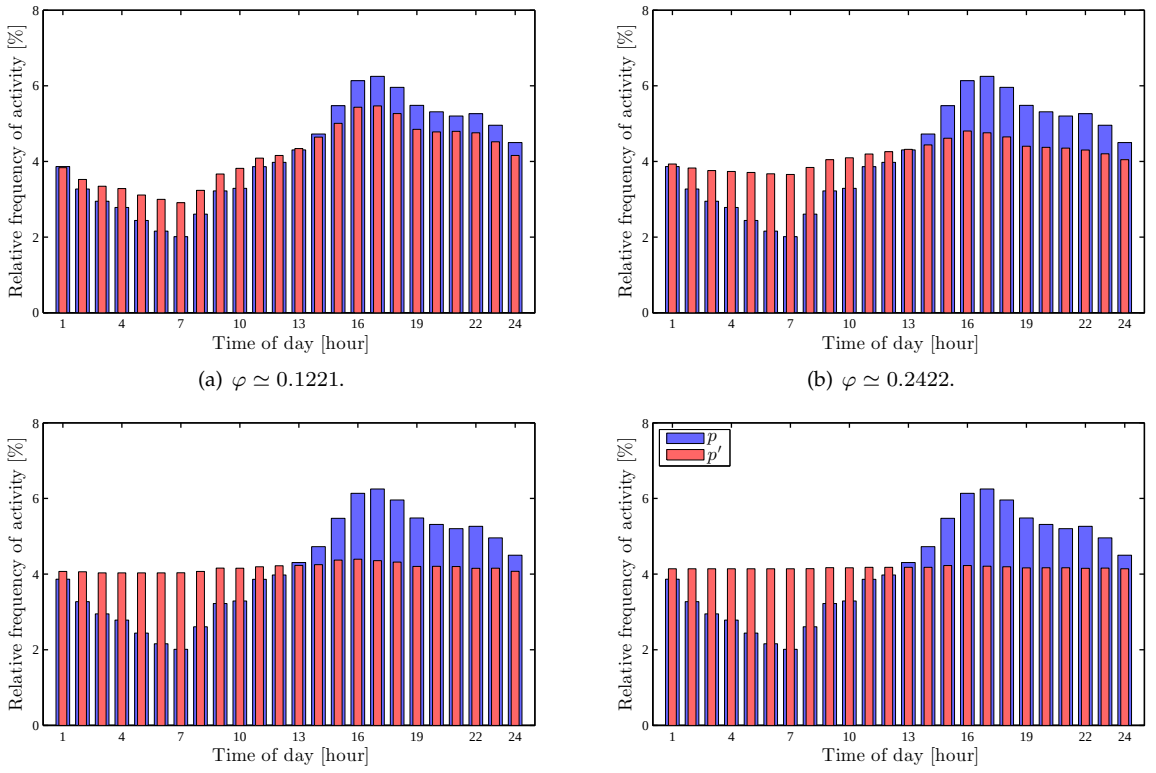
efficiently, as our mechanism contributes to distribute the message traffic load evenly.

Fig. 14 illustrates this point. In particular, it shows the percentage of messages posted to Twitter by our set of users within a day. Since we computed this as the aggregated profile of all users, we refer to it as the *population*'s profile p. The modified version of this relative histogram due to our mechanism is denoted by p'. We have represented this profile by assuming that all users apply a common message-deferral rate.

Not entirely unexpectedly, Fig. 14(a) shows that the time slots most affected by our PET are those with the lowest and highest activity. This is the case of the intervals 5, 6, 7 and 8 on the one hand, and 15, 16, 17, 18 and 19 on the other. For this relatively small value of deferral rate, the number of messages posted between 6 a.m. and 7 a.m. is increased by 44.68%, whereas the amount of messages sent between 16 p.m. and 17 p.m. is reduced by 12.50%. In Fig. 14(d), $\varphi \simeq 0.4844$ and the overall profile of activity p' becomes nearly uniform. In this last case, the largest increase in the number of tweets is observed for the time slot 7, while the largest reduction in the number of tweets is spotted for the time period 17. In particular, in those time intervals we observe an increase and a reduction of 106.03% and 32.65%, respectively. In summary, should our data set be representative of the whole population of Twitter users, the extensive application of the proposed PET could reduce substantially the number of networking resources and maximize the efficiency of such resources.

7 Conclusions and Future Work

Motivated by the lack of previous works specifically addressing the threat of time profiling in social networks, as well as the danger that such type of attack entails, the paper at hand presents an optimized, delay-based mechanism. This approach consists in an intelligent delay of a given number of messages posted by users in social networks in a manner that the observed profiles generated by the attacker do not break the privacy of those users. In other words, the attacker is unable to infer any time-based sensitive information by just observing and logging the timestamp of each interaction of the end users with the social networking sites.



(a) $\varphi \simeq 0.1221$. (b) $\varphi \simeq 0.2422$.

(c) $\varphi \simeq 0.3633$. (d) $\varphi \simeq 0.4844$.

Fig. 14: Relative histogram of the tweets in our data set within one day. We denote this histogram as p. As a consequence of the optimized deferral of those tweets, the profile p results in the modified profile p'.

Moreover, a detailed architecture implementing this mechanism has been described and analyzed, showing the feasibility of our proposal. Yet, any PET comes at the cost of certain utility loss. Hence, we have studied two meaningful utility metrics specific for our smart deferral mechanism (both in terms of the message deferral rate), namely: expected message delay and messages storage capacity. As shown, both metrics exhibit an increasing, nonlinear behavior with regards to the deferral rate. When the critical deferral rate (beyond which the maximum level of privacy is attained) is known, those outcomes become remarkably helpful to assess the optimal capacity for the messages buffer, as well as the average expected delay of each message in the system.

Finally, a comprehensive set of experiments has been conducted on three of the most popular social networks, Facebook, Instagram and Twitter, analyzing the behavior of 1 283 users, demonstrating the suitability of our solution and comparing it with two data-perturbative privacy technologies. In particular, it has been proved that most of the studied users will not require delaying a large percentage of their tweets for their apparent profiles to become the uniform distribution. Likewise, users in our data set will require relatively small margins of privacy gain to achieve the critical-privacy level. Another interesting conclusion states that our approach may help social networking sites manage their networking resources more efficiently, as it contributes to distribute the traffic load evenly. Furthermore, the mean values for the messages expected delay and messages storage capacity in our experiments, respectively, was 3.89 hours and 31.24% of users' messages.

As for the future research lines derived from this work, we are investigating some of the assumptions made in this work. Thus for instance, since we acknowledge that the user activity may vary significantly over time, we need to consider this fact in order to periodically update users' profiles. In the same direction, we want to study the bootstrapping problem, i.e., how to define users' profiles when the system is launched for the first time, or while the system is learning the actual users' profiles. Last but not least, we also aim at investigating the challenges derived from deploying and implementing our solution over a real environment, such as those related to the fact that users may in fact exhibit activity profiles with specific active time periods.

References

[1] D. Rosenblum, "What anyone can know: The privacy risks of social networking sites," *IEEE Secur., Priv.*, vol. 5, no. 3, pp. 40–49, 2007.

[2] R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, "Preventing private information inference attacks on social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 8, pp. 1849–1862, 2013.

[3] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, "Inferring private information using social network data," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 1145–1146.

[4] F. Gómez Mármol, M. Gil Pérez, and G. Martínez Pérez, "Reporting Offensive Content in Social Networks: Toward a Reputation-based Assessment Approach," *IEEE Internet Computing*, vol. 18, no. 2, pp. 32–40, 2014. [Online]. Available: http://dx.doi.org/10.1109/MIC.2013.132

[5] S. Pina Ros, A. Pina Canelles, M. Gil Pérez, F. Gómez Mármol, and G. Martínez Pérez, "Chasing offensive conducts in social networks: A reputation-based practical approach for Frisber," *ACM Transactions on Internet Technology*, vol. 15, no. 4, pp. 1–20, 2015. [Online]. Available: http://dx.doi.org/10.1145/2797139

[6] Z. Younis and R. A. Khatib, "Trending in ramadan — What do people tweet about during the holy month?" The Online Project, Tech. Rep., 2014. [Online]. Available: http://www.theonlineproject.me/files/reports/Trending_in_Ramadan_-_English1.pdf

[7] "Social media in ramadan — Exploring arab user habits on Facebook and Twitter," The Online Project, Tech. Rep., 2013. [Online]. Available: http://theonlineproject.me/files/newsletters/Social-Media-in-Ramadan-Report-English.pdf

[8] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, "All your contacts belong to us: Automated identity theft attacks on social networks," in *Proc. ACM Int. WWW Conf.*, Sanibel Island, FL, May 2009, pp. 551–560.

[9] J. R. Douceur, "The sybil attack," in *Proc. Int. Workshop Peer-to-Peer Syst. (IPTPS)*. London, UK: Springer-Verlag, 2002, pp. 251–260.

[10] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, "SybilGuard: Defending against Sybil attacks via social networks," in *Proc. ACM Conf. Special Interest Group Data Commun. (SIGCOMM)*, Pisa, Italy, Sep. 2006, pp. 267–278.

[11] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao, "Sybillimit: A near-optimal social network defense against sybil attacks," *IEEE/ACM Trans. Netw.*, vol. 18, no. 3, pp. 885–898, Jun. 2010.

[12] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Cancún, Mexico, Apr. 2008, pp. 506–515.

[13] ——, "The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks," *Knowl., Inform. Syst.*, vol. 28, no. 1, pp. 47–77, 2011.

[14] X. Shen, B. Tan, and C. Zhai, "Privacy protection in personalized search," *ACM Spec. Interest Group Inform. Retrieval (SIGIR) Forum*, vol. 41, no. 1, pp. 4–17, Jun. 2007. [Online]. Available: http://doi.acm.org/10.1145/1273221.1273222

[15] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Commun. ACM*, vol. 24, no. 2, pp. 84–88, 1981.

[16] L. Cottrell, "Mixmaster and remailer attacks," 1994. [Online]. Available: http://obscura.com/~loki/remailer/remailer-essay.html

[17] G. Danezis, "Mix-networks with restricted routes," in *Proc. Int. Symp. Priv. Enhanc. Technol. (PETS)*. Lecture Notes Comput. Sci. (LNCS), 2003, pp. 1–17.

[18] D. Kesdogan, J. Egner, and R. Büschkes, "Stop-and-go mixes: Providing probabilistic anonymity in an open system," in *Proc. Inform. Hiding Workshop (IH)*. Springer-Verlag, Apr. 1998, pp. 83–98.

[19] O. Berthold, A. Pfitzmann, and R. Standtke, "The disadvantages of free MIX routes and how to overcome them," in *Proc. Design. Priv. Enhanc. Technol.: Workshop Design Issues Anon., Unobser.*, ser. Lecture Notes Comput. Sci. (LNCS). Berkeley, CA: Springer-Verlag, Jul. 2000, pp. 30–45.

[20] C. Díaz, S. Seys, J. Claessens, and B. Preneel, "Towards measuring anonymity," in *Proc. Int. Symp. Priv. Enhanc. Technol. (PETS)*, ser. Lecture Notes Comput. Sci. (LNCS), vol. 2482. Springer-Verlag, Apr. 2002, pp. 54–68.

[21] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *Proc. Int. Symp. Priv. Enhanc. Technol. (PETS)*, vol. 2482. Springer-Verlag, 2002, pp. 41–53.

[22] S. Steinbrecher and S. Kopsell, "Modelling unlinkability," in *Proc. Int. Symp. Priv. Enhanc. Technol. (PETS)*. Springer-Verlag, 2003, pp. 32–47.

[23] C. Díaz, "Anonymity and privacy in electronic services," Ph.D. dissertation, Katholieke Univ. Leuven, Dec. 2005.

[24] D. Rebollo-Monedero and J. Forné, "Optimal query forgery for private information retrieval," *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4631–4642, 2010.

[25] D. C. Howe and H. Nissenbaum, *Lessons from the Identity Trail: Privacy, Anonymity and Identity in a Networked Society*. NY: Oxford Univ. Press, 2009, ch. TrackMeNot: Resisting surveillance in Web search, pp. 417–436. [Online]. Available: http://mrl.nyu.edu/~dhowe/trackmenot

[26] J. Parra-Arnau, A. Perego, E. Ferrari, J. Forné, and D. Rebollo-Monedero, "Privacy-preserving enhanced collaborative tagging," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 180–193, Jan. 2014. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2012.248

[27] J. Parra-Arnau, D. Rebollo-Monedero, J. Forné, J. L. Muñoz, and O. Esparza, "Optimal tag suppression for privacy protection in the semantic Web," *Data, Knowl. Eng.*, vol. 81–82, pp. 46–66, Nov. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2012.07.004

[28] M. Deng, "Privacy preserving content protection," Ph.D. dissertation, Katholieke Univ. Leuven, Jun. 2010.

[29] B. N. Levine, M. K. Reiter, C. Wang, and M. Wright, "Timing attacks in low-latency mix systems," in *Proc. Int. Financial Cryptogr. Conf.* Springer-Verlag, Feb. 2004, pp. 251–265.

[30] K. Bauer, D. McCoy, D. Grunwald, T. Kohno, and D. Sicker, "Low-resource routing attacks against anonymous systems," University of Colorado, Tech. Rep., 2007.

[31] S. J. Murdoch and G. Danezis, "Low-cost traffic analysis of tor," in *Proc. IEEE Symp. Secur., Priv. (SP)*, May 2005, pp. 183–195.

[32] B. Pfitzmann and A. Pfitzmann, "How to break the direct RSA implementation of mixes," in *Proc. Annual Int. Conf. Theory, Appl. of Cryptogr. Techniques (EUROCRYPT)*. Springer-Verlag, May 1990, pp. 373–381.

[33] W. M. Grossman, "alt.scientology.war," 1996. [Online]. Available: www.wired.com/wired/archive/3.12/alt.scientology.war_pr.html

[34] "AOL search data scandal," Aug. 2006, accessed on 2013-11-15. [Online]. Available: http://en.wikipedia.org/wiki/AOL_search_data_scandal

[35] "European data protection supervisor," May 2013. [Online]. Available: http://www.edps.europa.eu

[36] "Twitter charts - xefer." [Online]. Available: http://xefer.com/twitter/

[37] Y. Xu, K. Wang, B. Zhang, and Z. Chen, "Privacy-enhancing personalized Web search," in *Proc. Int. WWW Conf.* ACM, 2007, pp. 591–600.

[38] S. Ye, F. Wu, R. Pandey, and H. Chen, "Noise injection for search privacy protection," in *Proc. Int. Conf. Comput. Sci., Eng.* IEEE Comput. Soc., 2009, pp. 1–8.

[39] A. Erola, J. Castellà-Roca, A. Viejo, and J. M. Mateo-Sanz, "Exploiting social networks to provide privacy in personalized Web search," *J. Syst., Softw.*, vol. 84, no. 10, pp. 1734–745, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0164121211001117

[40] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "Measuring the privacy of user profiles in personalized information systems," *Future Gen. Comput. Syst. (FGCS), Special Issue Data, Knowl. Eng.*, vol. 33, pp. 53–63, Apr. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.future.2013.01.001

[41] M. Hildebrandt, J. Backhouse, V. Andronikou, E. Benoist, A. Canhoto, C. Diaz, M. Gasson, Z. Geradts, M. Meints, T. Nabeth, J. P. V. Bendegem, S. V. der Hof, A. Vedder, and A. Yannopoulos, "Descriptive analysis and inventory of profiling practices – deliverable 7.2," Future Identity Inform. Soc. (FIDIS), Tech. Rep., 2005.

[42] M. Hildebrandt and S. Gutwirth, Eds., *Profiling the European Citizen: Cross-Disciplinary Perspectives*. Springer-Verlag, 2008.

[43] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, no. 9, pp. 939–952, Sep. 1982.

[44] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[45] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

[46] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "Optimal forgery and suppression of ratings for privacy enhancement in recommendation systems," *Entropy*, vol. 16, no. 3, pp.

1586–1631, Mar. 2014. [Online]. Available: http://www.mdpi.com/1099-4300/16/3/1586

[47] T. M. Apostol, *Mathematical Analysis. A Modern Approach to Advanced Calculus*, 2nd ed. Addison Wesley, 1974.

[48] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*, August 2009.

[49] E. Ferrara, R. Interdonato, and A. Tagarelli, "Online popularity and topical interests through the lens of instagram," in *Proc. ACM Conf. Hypertext, Soc. Media (HT)*, 2014, pp. 24–34.