A Metric to Evaluate Interaction Obfuscation in Online Social Networks

Ero Balsa and Carmela Troncoso and Claudia Diaz
KU Leuven ESAT/COSIC, IBBT
Kasteelpark Arenberg 10
Leuven, Belgium
firstname.secondname@esat.kuleuven.be

Received (6 July 2012)
Revised (20 August 2012)
Accepted [a] (8 October 2012)

Online social networks (OSNs) have become one of the main communication channels in today's information society, and their emergence has raised new privacy concerns. The content uploaded to OSNs (such as pictures, status updates, comments) is by default available to the OSN provider, and often to other people to whom the user who uploaded the content did not intend to give access. A different class of concerns relates to sensitive information that can be inferred from the behavior of users. For example, the analysis of user interactions augments social network graphs with potentially privacy-sensitive details on the nature of social relations, such as the strength of user relationships. A solution to prevent such inferences is to automatically generate dummy interactions that obfuscate the real interactions between OSN users. Given an adversary that observes the obfuscated interactions, the goal is to prevent the adversary from recovering parameters of interest (e.g., relationships strength) that accurately describe the real user interactions. The design and evaluation of obfuscation strategies requires metrics that express the level of protection they would offer when deployed in a particular OSN with its underlying user interaction patterns. In this paper we propose mutual information as obfuscation metric. It measures the amount of information leaked by the (observable) obfuscated interactions in the system on the (concealed) real interactions between users. We show that the metric is suitable for comparing different obfuscation strategies, and flexible to accommodate different network topologies and user communication patterns. Obfuscation comes at the cost of network overhead, and the proposed metric contributes to enabling the optimization of strategies to achieve good levels of privacy protection at minimum overhead. We provide a detailed methodology to compute the metric and perform experiments that illustrate its suitability.

Keywords: online social networks, privacy, obfuscation, traffic analysis, metric, mutual information.

1. Introduction

Hundreds of millions of people use online social networks (OSNs) to share information and interact with their friends. This increasing reliance on OSNs to communicate has given rise to a host of privacy risks. The most prominent concern

[a]The published article is available at `http://www.worldscientific.com/toc/ijufks/20/06`

relates to the fact that the OSN provider can see all the information uploaded by the users. A number of prior works [1;2;4;3] have proposed solutions to protect content confidentiality. The main idea is to have a proxy (a browser plugin or OSN application) encrypting the uploaded data so that it is only visible to users who have the corresponding decryption keys, and hence not accessible to the OSN provider.

Even if content is kept confidential, interactions between users (i.e., their *communication profiles*) may disclose potentially sensitive information. [?] [5;6;7] Most existing work on protecting communication profiles against traffic analysis focuses on providing anonymity properties. [?] Anonymity is however not a viable option in OSNs, as users typically have an account and a personal page and thus can at best be pseudonymous. Moreover, often users must explicitly *befriend* other users to be able to interact with them and thus friendship relationships are explicit and known to the OSN provider. At the same time, social network users establish a large number of friendships of which only a small fraction corresponds to close relationships. Hence, concealing which relationships are meaningful in contrast to non-important acquaintances provides protection against inferences that exploit relationship strength information.

A common approach to conceal the "importance" (or *weight*) of user relationships is to generate *fake* or *dummy* interactions indistinguishable from actual user interactions. While this is a promising approach to hide meaningful relationships, the generation of dummies imposes communication and storage overheads on the network. In this work we propose to use mutual information to measure the degree of profile obfuscation provided by a dummy generation strategy (DGS), which in turn enables the search for optimal strategies that maximize protection with minimal overhead. We evaluate the suitability of the metric in OSNs empirically testing different DGSs, user behaviors, and network topologies. To the best of our knowledge, this is the first work that addresses the protection of communication profiles against traffic analysis in the specific context of OSN.

The rest of this paper is organized as follows. Section 2 describes the system and adversary models, and Section 3 introduces the mutual information as metric to evaluate the degree of obfuscation provided by a DGS, and describes our methodology to compute it. In Section 4 we empirically validate the suitability of the metric for OSN scenarios. We discuss some open questions in Section 5, and finally conclude in Section 6.

2. Traffic Analysis Resistant Online Social Networks

2.1. *An Abstract Model for Traffic Analysis Resistant OSNs*

Content protection. We consider an OSN in which the content uploaded to the network is protected through encryption, such that each piece of information can only be decrypted by the designated recipients (i.e., the data is kept confidential towards the OSN provider). We make abstraction of the concrete key management

and cryptographic protocols, and assume that a software tool (e.g., a browser plug-in[1] or an OSN application[4]) is available to the users. Figures 1 and 2 illustrate the operation of the plug-in "Scramble!" [1] when a user Bob browses the personal page of his friend Alice. The page that Bob downloads from the server contains five posts of Alice's conversations with Bob and Charlie, encrypted with keys she shares with them (hence the posts are not readable by the OSN provider). Bob's plug-in only has keys to decrypt the last three posts, as illustrated in Fig. 2b, and thus it presents him with a version of Alice's page in which only these three messages are included (in clear), as shown in Fig. 2c.

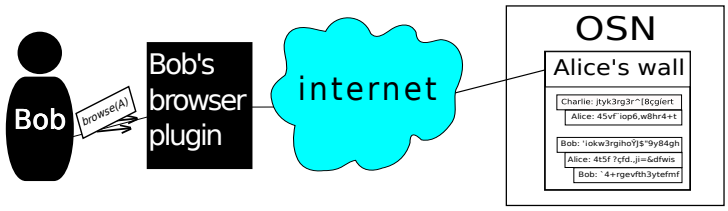
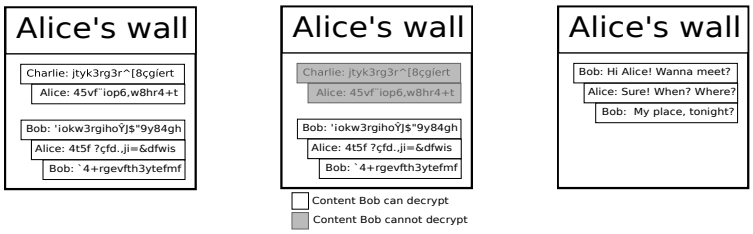


Fig. 1: Bob using a content protection plugin to browse Alice's page



(a) What Bob downloads

(b) What the plugin interprets

(c) What Bob sees

Fig. 2: How Alice's page looks like

Network model. We consider an OSN constituted by a set $\mathcal{U}$ of N users that establish friend relationships with each other. These relationships are modeled as *links* in the network graph. We make the following assumptions:

- Each user $u_i \in \mathcal{U}$ has a set of *friends* $\mathcal{F}_i \subseteq \mathcal{U}$ with whom she has a relationship (link) in the OSN. We consider that friend relationships are symmetric, i.e., $\forall i, j : u_j \in \mathcal{F}_i \Leftrightarrow u_i \in \mathcal{F}_j$. Users may communicate with these friends, or just have them as contacts without interacting with them.[19;20] If a user u_j is not a friend of user u_i (i.e., $u_j \notin \mathcal{F}_i$), then u_i never communicates with u_j.

In line with prior work,[9;12;14;11] we define the communication profile ψ_i of user u_i as the probability distribution describing the relative frequency of interaction w_{ij} of u_i with each of her friends u_j. Formally, $\psi_i = [w_{i1} \cdots w_{iN}]$. We refer to w_{ij} as the *weight* of the edge between u_i with u_j in the social graph. The weights w_{ij} satisfy the following conditions:

$$w_{ij} \geq 0, \quad \forall j \,|\, u_j \in \mathcal{F}_i,$$
$$w_{ij} = 0, \quad \forall j \,|\, u_j \notin \mathcal{F}_i,$$

$$\sum_{w_{ij} \in \psi_i} w_{ij} = 1$$

Note that u_i may initiate interactions with u_j with more relative frequency than u_j with u_i, and thus w_{ij} is not necessarily equal to w_{ji}.

Dummy interactions. We consider that the content protection software tool can also generate dummy interactions transparently to the user, and that it filters out dummy content in the same way as real content for which no decryption keys are available. These interactions are indistinguishable from user generated interactions towards the service provider, who sees both real and dummy messages as encrypted data.

We consider that the tool implements a dummy generation strategy (DGS) defining how dummy interactions are distributed amongst friends. The DGS assigns a dummy weight d_{ij} to each friend u_j that determines the amount of dummy interactions generated involving u_j. Concrete examples of DGSs are given in Section 4.

2.2. *Adversary model*

We consider a global passive adversary who knows the list of friends of every user in the OSN and monitors all the interactions between them (e.g., the OSN provider in a centralized social network). The goal of the adversary is to recover the weights w of the friendship links in the network.

Given a history of interactions, the adversary computes c_{ij} as follows. c_{ij} increases by one whenever user u_i initiates an interaction that involves exclusively user u_j, and by $\frac{1}{|G|}$ when it involves a subset G of her friends, $u_j \in G \subseteq \mathcal{F}_i$. Then, the weights w_{ij} can be estimated as:

$$w_{ij} = \frac{c_{ij}}{\sum_k c_{ik}}$$

In the presence of dummy interactions however, the adversary can only recover an *obfuscated* version of the genuine weights w. Let us call the weights recovered by the adversary *observed weights* and denote them as o, computed analogously to w_{ij}, but including dummy interactions in the computation of c_{ij}.

While we assume that the adversary cannot distinguish real from dummy interactions, we note that our model is also suitable for an adversary that is able to probabilistically (mis)classify dummy interactions. In this case, the observed weights o would be computed filtering out interactions classified as dummy, or alternatively, considering their contribution proportional to the likelihood of being real user interactions. In addition to this, the adversary we consider does not care about the different possible forms of interaction in OSNs (e.g., private messages, posts, page browsings). The model can however be trivially extended to consider that certain types of interactions contribute more than others to the weight of a friendship link.

3. Mutual Information as a Measure of Interaction Obfuscation

The mutual information is an information-theoretic quantity that captures the amount of information that is obtained about one random variable by observing another. Several prior works use mutual information as metric for privacy, in particular to measure anonymity properties. Moskowitz et al.[16] use the mutual information to study the relationship between quasi-anonymity and covert channels. Zhu and Bettati[17] propose to measure anonymity computing the mutual information between two random variables X and Y denoting, respectively, the actual and the suspected sender-receiver pairs communicating through a mix.[18] There are important differences between prior works and the metric proposed in this paper: i) we consider OSNs rather than anonymous communication networks, ii) we consider the generation of dummy traffic, and iii) the property of interest is obfuscation rather than anonymity.

We use the mutual information to measure the amount of information that the adversary obtains from the observed weights o about the real weights w. In order to do this, we model the real and observed weights as random variables W and O, respectively; then compute the mutual information between both variables as:

$$I(W;O) = \sum_{w \in W} \sum_{o \in O} p(w,o) \log \left(\frac{p(w,o)}{p(w) \cdot p(o)} \right) \tag{1}$$

The obfuscation provided by a DGS can range from perfect, i.e., $I(W,O) = 0$; to no obfuscation at all, i.e., $I(W,O) = H(W)$. When a DGS provides perfect obfuscation, the observed weights o are *independent* from the real weights w. Hence, $p(w,o) = p(w) \cdot p(o)$ and $I(W,O) = 0$, indicating that the observed weights o carry no information about w. On the other extreme when a DGS provides no protection at all, the observed weights o uniquely determine the value of w. In this case the mutual information between both random variables is maximal; i.e., $I(W;O) = H(W)$, where $H(W)$ denotes the Shannon entropy[21] of W:

$$H(W) = -\sum_{w \in W} p(w) \log(p(w))$$

A normalized version of the metric can be obtained by dividing the mutual information $I(W;O)$ by the entropy of the real weights random variable, $H(W)$.

This would result in a bounded metric, namely, its possible values would be limited to the interval $[0, 1]$. A normalized metric may however be less expressive. Whereas the proposed *non-normalized* metric reveals the amount of bits that the adversary gains from the observation of o, the normalized version represents the performance of a DGS in relative terms – and it is less intuitive to interpret.

The mutual information metric can be generalized to compare DGSs from perspectives other than particular relationships between two users. For instance, previous work [10;11] has shown that considering full communication profiles ψ_i (as opposed to individual weights w_{ij}) increases the accuracy with which the adversary can estimate w_{ij}. To measure the amount of information that the adversary obtains about the real profiles ψ_i from the observed profiles θ_i, it suffices to redefine the metric to consider a random variable Ψ describing the real profiles ψ and a random variable Θ describing the observed profiles θ. Another option is to redefine the metric to evaluate the system as a whole by considering random variables that describe matrices containing one real (respectively observed) profile per row, instead of just one profile or individual weight at a time. Note however that considering more information at once (e.g., profiles versus weights, or matrices instead of profiles) may not be feasible due to heavy computational and memory requirements.

Finally, the mutual information metric is flexible enough to evaluate other types of leakage different from the actual relationship weights. We illustrate this by evaluating the effectiveness with which a DGS conceals the best friend of a user in Sect. 4.

3.1. *Computing the Metric: Practical issues.*

We note that, although the relative frequency with which users communicate should be modeled as a continuous random variable, our metric (Eq. 1) is based on the mutual information defined for discrete random variables. This is because we consider that it is unlikely to have access to the continuous versions of the random variables W, O, and the joint variable (W, O). O and the joint (W, O) cannot be computed analytically due to the unpredictability of interactions between users and must be estimated from the observations. W can be obtained sampling observations of different OSNs. In both cases it is infeasible to explore a continuous state space, hence the random variables must be discretized.

3.1.1. *Quantization.*

Random variables can be quantized to reduce their outcome to a discrete series of values. The step of quantization, Δ, defines the length of the intervals where continuous values are mapped to a single discrete value. The influence of the step of quantization is twofold. On the positive side, increasing the quantization step "diminishes" the state space, reducing the number of samples required to compute the metric. However, with a large quantization step many values of w are grouped, hence providing coarser information to the system designer.

In addition, note that the quantization step does not need to be uniform. For example, we can define an arbitrary threshold t so that a user's friends are grouped into "close friends" (i.e., $u_j \in \mathcal{F}_i$ such that $w_{ij} \geq t$), and "acquaintainces" (i.e., $u_j \in \mathcal{F}_i$ such that $w_{ij} < t$).

3.1.2. *Sampling an OSN to estimate W, O and (W, O)*

We now explain how a DGS simulator can be used to estimate the random variables W, O and (W, O). Such simulator intertwines dummy interactions with user interactions that can be simulated or taken from existing social network data.

In each simulation, we obtain a sample of these variables by choosing two users u_i and u_j uniformly at random and storing (w_{ij}, o_{ij}) computed as described in Sect. 2.2. The process is repeated to obtain an arbitrary number s of samples.

Once samples are available, we compute the probability $p((W, O) = (w_x, o_y))$ by counting the number of occurrences $C_{(w_x,o_y)}$ of each pair of values (w_x, o_y) and dividing it by the total number of samples s. (The subscripts x and y belong to the set of quantized values that the weights w and o may take.) However, using a finite number of samples introduces an error in the estimation. We model $p(W, O)$ as a multinomial distribution and use Bayesian Inference to obtain a bound on this error.

The Dirichlet distribution is a conjugate prior for the multinomial distribution. Its probability density function represents the belief that the probability of occurrence of (w, o) is $p(w, o)$ given it has been observed $C_{(w,o)}$ times. We obtain δ samples $p(w, o)$ using the Dirichlet distribution with $C_{(w_x,o_y)}$ as input parameters:

$$p(W, O) \sim \text{Dirichlet}(C_{(w_1,o_1)} + 1, ..., C_{(w_1,o_m)} + 1, ..., C_{(w_m,o_1)} + 1, ..., C_{(w_m,o_m)} + 1)$$

The "+1" in the formula above indicate that we assume very limited prior knowledge on the real probability values, i.e., we ignore any characteristic from the actual distribution $p(W, O)$, other than assuming that all pairs (w, o) have a probability greater than zero.

For each sample drawn from the Dirichlet, we calculate the mutual information as follows:

$$I(W, O) = \sum_{w \in W} \sum_{o \in O} p(w, o) \log \left(\frac{p(w, o)}{\sum_{o \in O} p(w, o) \cdot \sum_{w \in W} p(w, o)} \right)$$

We take the median value of $I(W, O)$ as the estimated value of the mutual information, and consider the lowest value in the first quartile and the highest value in the third quartile as error bounds. This means that we consider the interval containing 50% of the values around the median. One can be more or less conservative about the error choosing a looser or tighter bound.

Table 1 offers a summary of the notation we have introduced throughout this section.

Table 1: Summary of notation

Symbol	Meaning
N	Size of the social network (in number of users)
$\mathcal{U}$	Set of users in the online social network
u_i	A user in the social network
$\mathcal{F}_i$	Set of friends of user u_i
w_{ij}	Weight of the link between two users u_i and u_j
ψ_i	Communication profile of u_i
W	Random variable of real weights w
Ψ	Random variable of real profiles ψ
d_{ij}	Dummy weight used by the DGS to decide which dummies from u_i must involve u_j
o_{ij}	Observed weight of the link between two users u_i and u_j
θ_i	Observed profile of u_i
O	Random variable of observed weights o
Θ	Random variable of observed profiles θ
c_{ij}	Amount of real interactions of u_i involving u_j
H	Shannon entropy
I	Mutual information
Δ	Step of quantization

4. Evaluation

In this section we illustrate the suitability of the mutual information as metric to evaluate the information leaked by a dummy generation strategy. For this purpose we apply the metric in different scenarios varying the DGS, the network topology, and the user behavior. We generate synthetic traces of OSN interactions using a Python OSN simulator,[b] which simulates both real and dummy interactions.[22]

4.1. *Experimental Setup*

The social graph. For the purpose of our evaluation, we use two toy-example social networks. The first is a regular network of size $N = 20$ users, each of which has 6 friends. In particular, $\mathcal{F}_i = \{u_j\}, j = \{i-3, i-2, i-1, i+1, i+2, i+3\}$ mod 20. The second is a fully connected network of size $N = 4$ users, i.e., all users are friends with each other. These networks are orders of magnitude smaller than typical OSNs. Nevertheless, they are sufficient to evaluate the effectiveness and flexibility of our metric.

User behavior. To illustrate how the mutual information captures differences in the performance of a DGS when the user behavior changes, we consider two types of communication profiles. *Worst case* profiles model scenarios in which users

[b]The code is available upon request.

communicate in pairs, i.e., each user interacts exclusively with one of her friends and never with the rest of her contacts. In other words, each user profile has a single weight with value 1 corresponding to the communication partner and 0 for the other friends. We consider this profile to be a "worst case" for a DGS, as the strategy must conceal one very strong relationship that concentrates all the user interactions. On the other hand, *Skewed* profiles model a more realistic case in which users communicate with all their friends, but some friends receive significantly more traffic than others. We generate skewed profiles following the approach described by Diaz et al. [?]

Dummy traffic generation strategies (DGS). We consider two DGS to illustrate how the mutual information captures the difference in the protection they provide. Both strategies generate a set of dummy weights d_{ij} for each user u_i by drawing samples from a uniform distribution, and normalizing the resulting vector. Note that dummy weights d_{ij} are independent from their corresponding real weight w_{ij}.

The *non-adaptive* strategy simply generates dummy actions from u_i to u_j according to d_{ij} without taking into account the interaction history nor the real weights. The *adaptive* strategy, on the other hand, monitors the interactions generated by u_i, and generates dummy traffic so that the observed weights o_{ij} recovered by the adversary are as close as possible to the dummy weights d_{ij}. For this purpose, whenever the observed weights o_{ij} deviate from the target weights d_{ij}, the strategy dynamically changes the recipients of the next dummy actions so that the value of o_{ij} is brought back to d_{ij}.

Quantization. Varying the quantization scheme makes possible different analyses that relate to diverse adversarial goals. We study the effect of the step and type of quantization in the performance of the metric. We consider a uniform quantization with the number of steps varying between two and five; i.e., $\Delta = \{1/2, 1/3, 1/4, 1/5\}$. We expect coarser intervals of quantization (e.g., $\Delta = 1/2$) to lead to smaller values of mutual information, showing that the adversary loses information by reducing the state space. Conversely, a powerful adversary with large computation resources will most likely obtain more information when sampling the probability space with a smaller step (e.g., $\Delta = 1/5$).

4.2. *Results*

In this section we present the experimental results of our evaluation. In all figures, the vertical axis represents the mutual information, and the symbols represent different quantization steps. To better illustrate the influence of other parameters in our experiments we keep the quantization step uniform, and fix the user profiles to be skewed, unless stated differently. The horizontal axis represents the dummy rate, which is the number of dummy interactions generated per real interaction. For example, a dummy rate of 3 means that for every real user interaction the plug-in generates 3 dummy interactions. When the rate is zero, no dummies are gener-

ated and the weights observed by the adversary correspond to the real weights (i.e., $o_{ij} = w_{ij}$). This represents the maximal information leakage (the adversary recovers all the meaningful information), and is reflected by the mutual information taking value $H(W)$ in the case of individual weights and $H(\Psi)$ in the case of communication profiles.

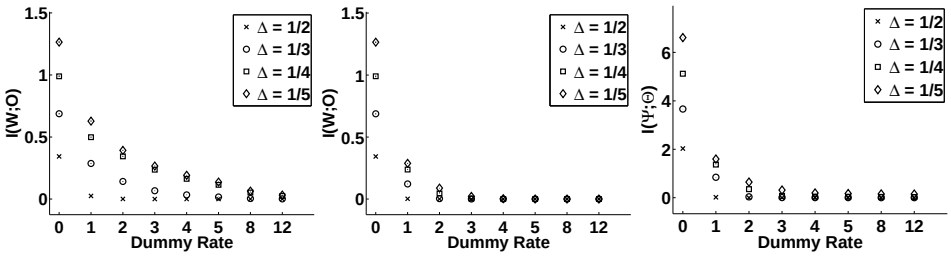
In order to minimize the estimation error, in each experiment we draw $s = 500\,000$ samples of (w, o) and (ψ, θ) to compute $p(w, o)$ and $p(\psi, \theta)$, respectively. For each quantization step Δ we represent the median value of mutual information ($\delta = 1000$ samples). The values of the first and third quartiles are not visible in the figures, as they are almost identical to the median. This tiny estimation error guarantees that enough samples of (w, o) and (ψ, θ) have been drawn.

Dummy Generation Strategy. Our first experiment is dedicated to show how the mutual information captures the difference in the protection provided by the non-adaptive and the adaptive dummy generation strategies. We consider an OSN with $N = 20$ users. Figures 3 (a) and (b) show our results. As expected, the metric reflects how the adaptive DGS, that takes into account the history of interactions when choosing the recipients of dummy actions, performs consistently better than the non-adaptive version that generates dummies independently from previous actions.

Dummy Rate. Figure3 illustrates how the mutual information captures the influence of the dummy rate. More dummies decrease the dependence of the observed variables on the real ones, hence increasing the obfuscation and reducing the information gain. Moreover, we see that increasing the dummy rate brings diminishing returns. Introducing a small amount of dummies reduces considerably the amount of information that can be extracted from the observation. Nevertheless, further increasing the overhead in the network does not bring benefits as the profiles are already obfuscated and there is no gain in adding dummy interactions.

Quantization step. Note that the quantization step has a great influence on the information gain. Not surprisingly, the amount of information available to the adversary about the real weights is larger for smaller quantization steps. However, we observe that the decay function is steeper for small steps, and soon the results converge. This indicates that performing computationally inexpensive analyses using coarse information are sufficient to evaluate the performance of a given DGS, or to compare two different DGSs.

Individual weights vs. profiles. The mutual information also captures the increase in information gained by the adversary when she considers user profiles as opposed to individual weights (see Figures 3 b and c). This is because by considering profiles as a whole, the adversary considers interdependencies between interactions (e.g., when Alice sends a message to Bob, she is not sending a message to Charlie) improving the accuracy of the estimation of individual weights. However, we can observe that the mutual information decrease with the dummy rate is very similar for both individual weights and profiles. This suggests that performing the evalua-



(a) Non adaptive DGS. Individual weights

(b) Adaptive DGS. Individual weights

(c) Adaptive DGS. Profiles

Fig. 3: Mutual information for adaptive and non-adaptive DGSs and individual weights vs. profiles for different quantization steps. $N = 20$ users, skewed profiles.

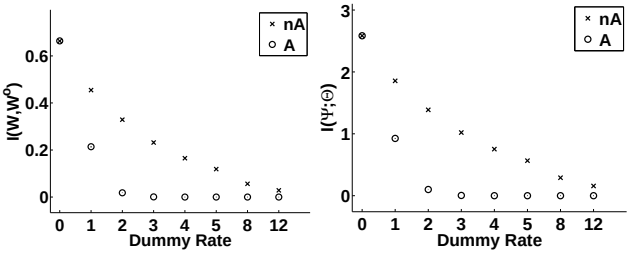
tion at the level of individual weights may serve to predict a DGS' performance at the profile level, hence reducing the computational complexity of the analysis.

Non-uniform quantization. As mentioned in Sect. 3.1.1 our metric does not require a uniform quantization step, but the quantization can be adjusted to accommodate specific adversarial goals. For example, the adversary can modify the quantization scheme to identify the "best" friend of a user, defined as the one with which the user interacts the most. To this end, a per-user threshold is defined, $t = \max_{\psi_i}(w_{ij})$ such that two quantization intervals are created: one containing the maximum weight, and another grouping the remaining weights.

Figure 4 displays the mutual information for individual weights and profiles for both adaptive (A) and non-adaptive (nA) dummy generation strategies, demonstrating that the metric is flexible enough to capture the change in the adversarial goal. We observe that in this case study the mutual information decays slower than when uniform quantization is used. This is because concealing the best friends requires more dummy interactions than obfuscating the profile as a whole. These results also confirm that adaptively generating dummies works better than generating them independently; and that the adversary gains more information when considering full profiles instead of individual weights.

Influence of network topology and user behaviour. Lastly, we analyze how the mutual information captures the effect of the topology and connectivity of the social graph as well as user behaviour on the performance of the adaptive DGS. We ran simulations in the full-meshed network of size $N = 4$ users described in Sect. 4.1 for both skewed and worst-case profiles.

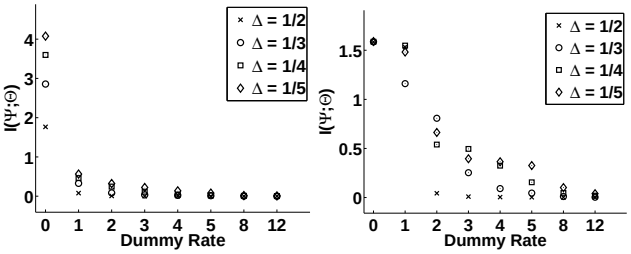
The results shown in Fig. 5 demonstrate that the mutual information reflects variations driven by changes in user behavior, capturing the expected negative effect that *worst-case* profiles have in the performance of the DGS. There are two features of the worst-case profiles worthy to discuss. First, we observe (Fig. 5b) that at



(a) Individual weights (b) Profiles

Fig. 4: Mutual information for non-uniform quantization for adaptive (A) and non-adaptive (nA) strategies. $N = 20$ users, skewed profiles.

dummy rate 0 there is no gain in diminishing the quantization step. Recall that users only communicate with one of their friends; hence when no dummies are generated no matter the quantification step used only two quantization intervals have samples (the ones corresponding to $w = 0$ and $w = 1$). Second, unintituively, for some dummy rates Fig. 5b shows mutual information values larger for coarse quantization than for fine quantization. This error is due to the fact that the real weights in worst-case profiles are far from uniformly distributed whereas the quantization used is uniform. Nevertheless, the mutual information converges for all quantization steps as the dummy rate increases.



(a) *Skewed* profiles (b) *Worst-case* profiles

Fig. 5: Effect of topology and user behaviour. $N = 4$ users, fully connected topology.

The figure also shows that the mutual information decreases much faster for *skewed* profiles than in the previous example (cfr. Fig. 3c), where the network was not fully connected and there were $N = 20$ users. This demonstrates that the metric correctly captures the fact that the performance of the DGS decreases as the number of friends increases.

5. Discussion

Simulated traffic data vs. real traffic data from deployed systems. In this work we have used simulated OSN user interactions to assess the suitability of the mutual information as a metric for interaction obfuscation evaluation. Many recent works use data collected from real networks for their security and privacy analyses.[23;24;25;26] Our experiments show that the metric presented in this paper is not tailored to a specific OSN topology, dummy strategy, or user behavior; and hence should be able to accommodate any system and usage pattern, including those existing in deployed social networks. We note however that even though simulated data is sufficient to assess the suitability of the metric, our results suggest that the effectiveness of a DGS strongly depends on parameters such as the users' behavior or their number of friends. Hence, designing a DGS for a particular OSN requires an evaluation with real data (i.e., actual OSN structure and user behavior) in order to assess the practical effectiveness of the DGS.

Scalability of the metric. We have considered in our experiments networks that are orders of magnitude smaller than practical OSNs. The reason to do this was to obtain results in reasonable time, as our implementation to compute the mutual information is not optimized, hence computationally intensive. Nevertheless, the size of the network is orthogonal to the fundamental question we want to answer in this work, namely, how to quantify the amount of information leaked on real user interactions when dummy interactions are generated. Our experiments show that the mutual information is a promising candidate metric for this purpose.

Besides, as pointed out in Sect. 4.2, the result of our experiments hints that the computation of the metric can be made efficient hence suitable for large networks such as the ones deployed in the real world. First, the computation time of the metric is heavily dependent on the quantification step. Our results show however that as the dummy rate increases, the benefits of decreasing the quantization step are very limited, as considering coarser intervals provides very similar information to more fine-grained intervals. Hence, networks can be efficiently analyzed using a coarse quantization scheme. Also if the considered DGS is devised to protect particular goals (e.g., best friends), the designer can consider an asymmetric and/or dynamic quantification step to efficiently evaluate the system.

Further, our results indicate that the decrease of mutual information as a function of the dummy rate is very similar for individual weights and for profiles. Hence, it may be possible to perform the analysis uniquely on individual weights to evaluate a DGS. This analysis is significantly more efficient and hence permits the analyst to evaluate large networks.

In addition to the performance gain that can be obtained by adjusting the parameters of the analysis as described in Sect. 4.2, future work should try to find more efficient ways to compute the mutual information. A possible approach is to use advanced Bayesian inference techniques that use sampling to reduce the complexity

of the problem, as proposed in prior work for the analysis of mix networks.[10;27]

Network overhead. In this work we have abstracted practical issues related to the deployment of dummy generation strategies in order to focus on the quantification of the information leaked about real user interactions when dummy interactions are generated. We note that prior to deployment, a feasibility analysis must be performed besides the security evaluation to ensure that the computational and storage requirements of a DGS are compatible with the OSN capabilities.

6. Conclusion

Previous work aimed at mitigating the privacy risks arising from the use of OSNs focuses on providing users with means to protect the uploaded content.[1;2;4;3] These solutions disregard the fact that even when content confidentiality is guaranteed, user interactions can still be used to infer sensitive private information.

In this paper we have proposed mutual information as a metric to measure the information leaked on the users' communication profiles when they are obfuscated using dummy interactions. We have provided a methodology to compute the metric and empirically evaluated its suitability for the OSN scenario. We have showed that the metric correctly captures changes in the parameters of the system, and that varying the quantization step allows to efficiently evaluate dummy generation strategies given limited computational resources. Further work is required to develop more efficient methods for computing the metric for large networks, as well as to extend it to account for additional information that may be available to the adversary.

Acknowledgements.

This work was supported in part by the projects: GOA TENSE (GOA/11/007), IAP Programme P6/26 BCRYPT, EC ICT-2007-216676 ECRYPT II, IWT SBO SPION, FWO G.0360.11N, and FWO G.068611N. C. Troncoso and C. Diaz are funded by the Fund for Scientific Research in Flanders (FWO). The authors wish to thank David Rebollo for valuable discussions and suggestions.

References

1. Filipe Beato, Markulf Kohlweiss, and Karel Wouters. Scramble! your social network data. In *PETS*, pages 211–225, 2011.
2. Saikat Guha, Kevin Tang, and Paul Francis. Noyb: privacy in online social networks. In *Workshop on Online social networks (WOSN)*, pages 49–54. ACM, 2008. ISBN 978-1-60558-182-8. doi: http://doi.acm.org/10.1145/1397735.GTF08.
3. Wanying Luo, Qi Xie, and Urs Hengartner. Facecloak: An architecture for user privacy on social networking sites. In *International Conference on Computational Science and Engineering (CSE)*, pages 26–33. IEEE Computer Society, 2009. ISBN 978-0-7695-3823-5. doi: http://dx.doi.org/10.1109/CSE.2009.387.
4. Matthew M. Lucas and Nikita Borisov. Flybynight: mitigating the privacy risks of social networking. In *ACM Workshop on Privacy in the Electronic Society (WPES)*,

pages 1–8. ACM, 2008. ISBN 978-1-60558-289-4. doi: http://doi.acm.org/10.1145/1456403.LB08.

5. Carter Jernigan and Behram Mistree. Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10), Sep 2009.
6. Ieng-Fat Lam, Kuan-Ta Chen, and Ling-Jyh Chen. Involuntary information leakage in social network services. In Kanta Matsuura and Eiichiro Fujisaki, editors, *3rd International Workshop on Security (IWSEC 2008)*, volume 5312 of *Lecture Notes in Computer Science*, pages 167–183. Springer, 2008.
7. Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. User interactions in social networks and their implications. In *European conference on Computer Systems (EuroSys)*, pages 205–218. ACM, 2009. ISBN 978-1-60558-482-9. doi: http://doi.acm.org/10.1145/1519065.WBSPZ09.
8. George Danezis. Statistical disclosure attacks: Traffic confirmation in open environments. In Gritzalis, Vimercati, Samarati, and Katsikas, editors, *Proceedings of Security and Privacy in the Age of Uncertainty, (SEC2003)*, pages 421–426, Athens, May 2003. IFIP TC11, Kluwer.
9. George Danezis, Claudia Diaz, and Carmela Troncoso. Two-sided statistical disclosure attack. In Nikita Borisov and Philippe Golle, editors, *7th International Symposium on Privacy Enhancing Technologies (PETS 2007)*, volume 4776 of *Lecture Notes in Computer Science*, pages 30–44. Springer-Verlag, 2007.
10. George Danezis and Carmela Troncoso. Vida: How to use bayesian inference to de-anonymize persistent communications. In Ian Goldberg and Mikhail J. Atallah, editors, *9th Privacy Enhancing Technologies Symposium (PETS 2009)*, volume 5672 of *Lecture Notes in Computer Science*, pages 56–72. Springer, 2009.
11. Carmela Troncoso, Benedikt Gierlichs, Bart Preneel, and Ingrid Verbauwhede. Perfect matching disclosure attacks. In Nikita Borisov and Ian Goldberg, editors, *8th International Symposium on Privacy Enhancing Technologies (PETS 2008)*, volume 5134 of *Lecture Notes in Computer Science*, pages 2–23. Springer-Verlag, 2008.
12. Claudia Diaz, Stefaan Seys, Joris Claessens, and Bart Preneel. Towards measuring anonymity. In Roger Dingledine and Paul F. Syverson, editors, *2nd International Workshop on Privacy Enhancing Technologies (PET 2002)*, volume 2482 of *Lecture Notes in Computer Science*, pages 54–68. Springer, 2002.
13. Matthew Edman, Fikret Sivrikaya, and Bülent Yener. A combinatorial approach to measuring anonymity. In *IEEE International Conference on Intelligence and Security Informatics (ISI 2007)*, pages 356–363. IEEE, 2007.
14. Andrei Serjantov and George Danezis. Towards an information theoretic metric for anonymity. In Roger Dingledine and Paul F. Syverson, editors, *2nd International Workshop on Privacy Enhancing Technologies (PET 2002)*, volume 2482 of *Lecture Notes in Computer Science*, pages 41–53. Springer, 2002.
15. Gergely Tóth and Zoltán Hornák. Measuring anonymity in a non-adaptive, real-time system. In David Martin and Andrei Serjantov, editors, *4th International Workshop on Privacy Enhancing Technologies (PET 2004)*, volume 3424 of *Lecture Notes in Computer Science*, pages 226–241. Springer, 2004.
16. Ira S. Moskowitz, Richard E. Newman, and Paul F. Syverson. Quasi-anonymous channels. In *International Conference on Communication, Network, and Information Security (CNIS 2003)*, pages 126–131, 2003.
17. Ye Zhu and Riccardo Bettati. Anonymity vs. information leakage in anonymity systems. In *25th International Conference on Distributed Computing Systems (ICDCS 2005)*, pages 514–524. IEEE Computer Society, 2005.
18. David Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms.

Commun. ACM, 24(2):84–90, 1981.

19. Hyunwoo Chun, Haewoon Kwak, Young-Ho Eom, Yong-Yeol Ahn, Sue B. Moon, and Hawoong Jeong. Comparison of online social relations in volume vs interaction: a case study of Cyworld. In Konstantina Papagiannaki and Zhi-Li Zhang, editors, *8th ACM SIGCOMM Conference on Internet Measurement 2008*, pages 57–70. ACM, 2008.
20. Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), 2009.
21. Claude Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423:623–656, 1948.
22. Ero Balsa. Privacy-preserving social networks. design and evaluation. Master's thesis, K.U. Leuven, 2010.
23. Marco Balduzzi, Christian Platzer, Thorsten Holz, Engin Kirda, Davide Balzarotti, and Christopher Kruegel. Abusing social networks for automated user profiling. In Somesh Jha, Robin Sommer, and Christian Kreibich, editors, *13th International Symposium on Recent Advances in Intrusion Detection (RAID 2010)*, volume 6307 of *Lecture Notes in Computer Science*, pages 422–441. Springer, 2010.
24. Joseph Bonneau, Jonathan Anderson, and George Danezis. Prying data out of a social network. In Nasrullah Memon and Reda Alhajj, editors, *2009 International Conference on Advances in Social Network Analysis and Mining (ASONAM 2009)*, pages 249–254. IEEE Computer Society, 2009. ISBN 978-0-7695-3689-7.
25. Dieudonné Tchuente, C. Marie-Françoise Canut, Nadine Baptiste-Jessel, André Péninou, and Anass El Haddadi. Visualizing the evolution of users' profiles from online social networks. In Nasrullah Memon and Reda Alhajj, editors, *International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2010)*, pages 370–374. IEEE Computer Society, 2010.
26. Christo Wilson, Alessandra Sala, Joseph Bonneau, Robert Zablit, and Ben Zhao. Don't Tread on Me: Moderating Access to OSN Data with SpikeStrip . *WOSN 2010: The Third Workshop on Online Social Networks*, 2010. URL `http://www.cs.ucsb.edu/~ravenben/publications/pdf/spikestrip-wosn10.pdf`.
27. Carmela Troncoso and George Danezis. The bayesian traffic analysis of mix networks. In Ehab Al-Shaer, Somesh Jha, and Angelos D. Keromytis, editors, *Proceedings of the 2009 ACM Conference on Computer and Communications Securityv (CCS 2009)*, pages 369–379. ACM, 2009.
28. Claudia Diaz and Bart Preneel. Reasoning about the anonymity provided by pool mixes that generate dummy traffic. In *Information Hiding Workshop (IH)*, LNCS, May 2004.