On the Privacy of Anonymized Networks

Pedram Pedarsani Matthias Grossglauser *
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland

ABSTRACT

The proliferation of online social networks, and the concomitant accumulation of user data, give rise to hotly debated issues of privacy, security, and control. One specific challenge is the sharing or public release of *anonymized data* without accidentally leaking personally identifiable information (PII). Unfortunately, it is often difficult to ascertain that sophisticated statistical techniques, potentially employing additional external data sources, are unable to break anonymity.

In this paper, we consider an instance of this problem, where the object of interest is the structure of a social network, i.e., a graph describing users and their links. Recent work demonstrates that anonymizing node identities may not be sufficient to keep the network private: the availability of node and link data from another domain, which is correlated with the anonymized network, has been used to re-identify the anonymized nodes. This paper is about conditions under which such a de-anonymization process is possible.

We attempt to shed light on the following question: can we assume that a sufficiently sparse network is inherently anonymous, in the sense that even with unlimited computational power, de-anonymization is impossible? Our approach is to introduce a random graph model for a version of the de-anonymization problem, which is parameterized by the expected node degree and a similarity parameter that controls the correlation between two graphs over the same vertex set. We find simple conditions on these parameters delineating the boundary of privacy, and show that the mean node degree need only grow slightly faster than $\log n$ with network size n for nodes to be identifiable. Our results have policy implications for sharing of anonymized network information.

Categories and Subject Descriptors

H.1 [**Models and Principles**]: General; G.3 [**Mathematics of Computing**]: Probability and Statistics; K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*

*This work has been supported by the Swiss National Science Foundation (SNSF) under grant 200021-116510.

KDD'11, August 21–24, 2011, San Diego, California, USA.

General Terms

Theory, Measurement

Keywords

network privacy, de-anonymization, social networks, random graphs, graph sampling

1. INTRODUCTION

The emergence of online social networks such as Facebook, Twitter, MySpace, Linkedin, etc. with hundreds of millions of users implies that an unprecedented amount of user data is now in the hands of the providers of such services. Not surprisingly, the fair use of this information, the appropriate notions of privacy and security, and the technical and legal tools to control its sharing and dissemination, have become controversial and hotly debated problems in the scientific community and beyond.

There are many reasons why social network data might be shared between organizations, or even released into the public domain. First, this information is very valuable for scientific purposes: the modest number of publicly available datasets has led to a broad variety of research projects and results. For example (and without any claim to exhaustiveness), promising research directions in this area include probabilistic modeling of network properties and dynamics [19], real data measurement and analysis [25, 24, 16, 6], and developing scalable algorithms to navigate and infer data from large-scale networks [18, 20]. Obviously, this requires great care in order to avoid the accidental release of sensitive information about individual users. As AOL's public relations disaster a few years ago [1] illustrates, simply anonymizing user identities may not be sufficient to prevent an attacker from identifying individual users through other means. Second, online social networks are increasingly integrated with other services on the web, which requires a certain amount of sharing between organizations (e.g., facebook third-party applications and the facebook connect function on third-party websites). Third, it has been recognized that social network information has strong potential for marketing purposes (e.g., for churn prediction [32] or for targeted advertisement [12]). In all of these areas lurks the risk of accidental or deliberate violations of user privacy. Several works address the privacy issue in social networks [8, 17], and propose mechanisms to preserve users' privacy [39, 9], or suggest the vulnerability of online social networks to different attacks [15].

In order to protect users' privacy, an established method is replacing their identities with random unique IDs, a process known as *anonymization.* However, recent works have shown methods that are able to infer the true user identities under certain conditions, by relying on side information [28, 3, 38]. We present a more detailed

summary of the related work in this field in Section 2. Most of the works in this area focus on proposing *algorithms* and *methods* for de-anonymizing networks, tested on various real datasets of online networks [23, 27]. Papers in this category propose algorithms for either attacking *specific* users in publicly available network data and revealing their identity [3], or de-anonymizing a fraction of all users in a network [38]. A major challenge in all these scenarios is scalability. Recently, algorithms that are applicable for de-anonymizing large networks have also been introduced [28, 27].

What is still lacking in the literature is a thorough understanding of the conditions under which de-anonymization is feasible. We would like to be able to ascertain when a network's anonymity can be guaranteed, given the side information and computational resources available to an attacker. In this paper, we attempt to make a step in this direction. We study a challenging version (from the perspective of an attacker) of the de-anonymization problem, where the attacker has no side information about nodes in the network to be attacked other than network structure, specifically a correlated version of the edge set of that network obtained from other sources.

Contribution. To the best of our knowledge, ours is the first paper to pursue a theoretical treatment of network de-anonymization problem, and in particular, considering its feasibility for *large* networks. Our contributions are three-fold: We explore fundamental limits for de-anonymization regardless of the specific algorithm employed, and investigate the relationship between network parameters and the possibility of guaranteeing anonymity in such networks. Moreover, we introduce a mathematically tractable model that captures the notion of correlated networks, and uses the idea of graph sampling to control the structural similarity of two graphs. This model is based on random graphs, and can be viewed as a generalization of the classical automorphism group problem for random graphs [5, chapter 9]. Finally, we prove that a surprisingly simple and mild condition on the scaling of the expected degree with the number of nodes is sufficient for de-anonymization to be feasible, with strong implications on privacy.

The following important observation is behind our modeling approach: In most real cases, although nodes are anonymized in the released data of social networks, the structure of the graph is preserved, i.e., this is equivalent to having access to an unlabeled graph. We assume that an attacker has access to an auxiliary labeled network, in which user identities are known. Such a network could be obtained for example from public data, or inferred from other sources. This type of attack is also considered in [28].

To give a concrete example, we ask whether it would be safe for an academic institution to release a database of anonymized email or call logs, if an attacker has available to him a correlated but highly incomplete set of likely social links between the staff and students of that institution (e.g., by mining the public web site of groups, departments, and so on)? Could an attacker use this incomplete side information to reverse-engineer the anonymized identities in the database, and therefore the communication pattern of this university? More generally, most of us have many different online identities that are in different hands, and the social links in these different domains are likely not completely identical, but correlated.

It is clear that the availability of additional side information (e.g., class labels for users such as from demographic information, or richer link information such as directed interactions, time stamps) can only further benefit the attacker. Here, we assume that the attacker only has the graph structure for re-identification of nodes.

In this paper we explore the problem of approximate graph matching introduced above. We use the notion of *graph sampling* to develop a model of *similar* or *correlated* graphs. Graph sampling has been used in other contexts, e.g., as a way to estimate node and edge features from the network [35, 11], or to generate different snapshots of an observed network as samples from a hidden underlying graph [31]. The structural similarity we seek is achieved by sampling the two graphs from an underlying *generator* graph. Our key result is that under surprisingly mild conditions on the model parameters, depending on the extent of the overlap between the two graphs, it is possible to establish a perfect mapping between the nodes of the two graphs as the number of nodes grows large.

Our results for approximate graph matching not only exhibit the risk of a privacy breach in the release of even the most basic information about real networks (i.e., only anonymized users and their links), but can have useful applications as well. If matching is feasible, one can combine several "noisy" public (anonymized) versions of social networks obtained from different sources into a more precise, combined network. In another scenario, suppose we have the call graph between all the phone numbers in an organization, and the graph of email exchanges between email addresses in this same organization. One could then establish the correspondence between phone numbers and email addresses solely through the structure of the two social networks (which we expect to be similar but not exactly equal).

We should emphasize that this paper only addresses the *feasibility* of de-anonymization. This amounts to establishing that there exists a cost function over the two graphs, such that minimizing this function finds the correct matching with high probability. We do not address the computational complexity of this process. The recent work of Schmatikov and Narayanan [28] report success in de-anonymizing fairly large networks. However, their work focuses on heuristics for matching, which they evaluate over samples of real social networks, while our focus is to understand the boundaries of anonymity in terms of fundamental network properties.

The remainder of this paper is organized as follows. Section 2 briefly discusses related work. In Section 3, we formally define the de-anonymization problem, and introduce a mathematical model for approximate graph matching of large networks. Section 4 is the core of this paper where we prove that in our model, perfect matching is feasible under mild conditions on the expected degree of the graphs and on their similarity. Section 5 discusses numerical experiments using social network data to justify the assumptions in our model. Finally, Section 6 concludes the paper with a discussion of the implications of the result.

2. RELATED WORK

We briefly summarize related work in network de-anonymization and approximate graph matching. This can be categorized as follows: 1) papers relevant to network modeling with direct application for de-anonymizing users in social networks, 2) papers in the area of graph isomorphism and approximate graph matching, mostly from applications in machine learning and pattern recognition problems.

In the first category, in their recent work [28], Narayanan and Shmatikov propose a novel algorithm for de-anonymizing social networks, based purely on network topology. Their algorithm uses the structural similarity of a target and an auxiliary network. Although the goal and problem definition of our contribution is similar to theirs, we seek insights into the fundamental conditions for de-anonymization to be feasible, while they demonstrate the effectiveness of de-anonymization of a real social network using heuristics.

Backstrom et al. introduce active and passive attacks for de-anonymization of social networks [3]. They show how a target users can be identified in a very large network by identifying a

neighborhood subgraph around the user using only network structure. They investigate the effectiveness of these attacks both theoretically and empirically. A limitation of active attacks is the necessity of creating fake (dummy) nodes in the social network before its release (which is of course a strong limitation in practice), while passive attacks are capable of re-identifying only a limited number of users, but without the need for fake nodes. Thus, the method works best for de-anonymization of *specific* users within the network, or a small fraction of all users. A similar attack model is analyzed in [10], where an attacker is allowed to issue queries that reveal a k-hop subgraph around a target node; they analyze the privacy risk to the identity of the target node and to the presence of specific links, both using random graph models and real data.

Finally, a novel de-anonymization attack is introduced by Wondracek et al. [38] that exploits group membership information available on social networking sites. They show that information about the group memberships of a user is often sufficient to uniquely identify this user, or at least to significantly reduce the set of possible candidates, and assess the feasibility of the attack both theoretically and empirically.

In the second category, several works propose different techniques for exact and approximate graph matching, mostly in image processing and pattern recognition. In [30], Cordella et al. propose a so called VF algorithm as a solution for *exact* subgraph matching, or subgraph isomorphism, exhibiting less complexity compared to the famous Ullmann backtracking algorithm [37]. In [36], Tian and Patel suggest an approximate graph matching tool (TALE) through a novel indexing method that incorporates graph structural information in a hybrid index structure. Although the structural information for matching graphs is used, the approximate matching problem in such cases is generally defined as node mismatches or inconsistencies in node attributes, rather than structural difference (in edges) as in our case.

Other works in this area propose different methods such as random walks on graphs [7], using EM algorithm and singular value decomposition [22], and the edit-distance criterion for approximate matching different types of graphs [34, 26]. Because of the complexity of matrix manipulation and computation of probability distributions, such methods are not feasible for application to very large networks.

Our contribution to this existing body of work is to introduce a mathematically tractable, parsimonious model for the problem of matching two similar graphs, and to derive asymptotic bounds in terms of fundamental parameters for network anonymity, independently of specific algorithms.

3. PROBLEM DEFINITION AND MODEL

We define the problem of matching the vertex sets of two graphs, and introduce the $G(n, p; s)$ random graph model, which generates two similar graphs $G_{1,2}$ over the same vertex set. As mentioned before, the goal is to match the vertices of two unlabeled graphs whose edge sets are correlated but not necessarily equal. The motivation for our model is its parsimony and symmetric structure, ingredients for its mathematical tractability.

The model assumes that the observed networks $G_{1,2}$ are incomplete manifestations of a true underlying network G of relationships. For example, the edges of G might represent the true relationships between a set of people, while $G_{1,2}$ capture the observable interactions between these people, such as communications (email, phone calls, proximity, and so on), or "friend"-relationships in a social network. $G_{1,2}$ might alternatively represent observations of the same network at different points in time.

To elaborate on this, let $G = (V, E)$ be a generator graph with vertex set V and edge set E. We assume here that G is an Erdös-Rényi random graph $G(n, p)$ with n nodes, where every edge exists with identical probability p, independently of all the other edges. For a fixed realization of $G = G(n, p)$, we generate two graphs $G_{1,2} = (V, E_{1,2})$ by sampling the vertex set E twice. More precisely, each edge $e \in E$ is in the edge set of $E_{1,2}$ with probability s, independently of everything else. As a result, the sample graphs $G_{1,2}$ are themselves Erdös-Rényi random graph $G(n, ps)$, but their edge sets are correlated, in that the existence of an edge in E_1 implies that the existence of this edge in E_2 is more likely than unconditionally (provided $p < 1$ and $s > 0$) (see Fig. 1). The $G(n, p)$ model has been widely used in the study of complex and social networks [4, 14, 29, 2], which makes it a plausible candidate for the study of the approximate matching problem.

Our goal is to determine whether it is possible to find the correct mapping between the nodes of G_1 and G_2, assuming we only see unlabeled versions of these two graphs (and without access to the generator G). This is equivalent to the assumption that the two graphs have different vertex label sets that contain no information about the graphs, such as random labels allocated in an anonymization procedure. Using this model, our problem can be viewed as the generalization of the classical automorphism group problem in random graphs. We discuss this and also the effect of the choice of other graph models at the end of Section 4 and also in Section 6.

We formally define the graph matching problem as follows. We assume that $G_{1,2}$ are only available in unlabeled form (or equivalently, with two arbitrary and unknown sets of labels). Let π denote a permutation on V, i.e., one way of mapping vertices from G_1 onto G_2. The number of such permutations is $n!$. The identity permutation, denoted by π_0, is the correct mapping between the nodes of G_1 and G_2. We seek an error function over the set of permutations, which succeeds if it is uniquely minimized by π_0.

Therefore, to solve the matching problem, we are interested to show the following:

Among all possible permutations between the two vertex sets, the identity permutation π_0 is the permutation that minimizes an error function, giving the node matching between the two graphs.

The error function should measure to what extent the structures of graphs G_1 and G_2 resemble each other under a given permutation. The structural difference can be viewed as the difference between the corresponding edge sets. This idea has also been investigated in the field of pattern recognition where the *edge-consistency* of two graph patterns (in matching a data graph to a model graph) is used to obtain the correspondence errors [26, 22].

We introduce the error measure for edge-inconsistency, considering only the structures of two graphs $G_1(V, E_1)$ and $G_2(V, E_2)$. The matching error Δ can be generally defined as

$$\Delta_\pi \quad = \quad \sum_{e \in E_1} \mathbf{1}_{\{\pi(e) \notin E_2\}} + \sum_{e \in E_2} \mathbf{1}_{\{\pi^{-1}(e) \notin E_1\}}, \qquad (1)$$

where $\mathbf{1}_{\{A\}}$ denotes the indicator function. In other words, permutation π defines a mapping between the nodes of G_1 and G_2, and Δ counts the number of edges that exist in one graph with the corresponding edges not existing in the other graph under matching π. This is the simplest error function that can be assumed for such a setting when comparing the structures of two graphs. While this cost function is not necessarily optimal (depending on the graph model) nor computationally efficient, it lends itself to probablistic analysis. Specifically, we prove below that if the sampling probability s is beyond some threshold, as n grows large, the identity permutation π_0 minimizes the error function (1).

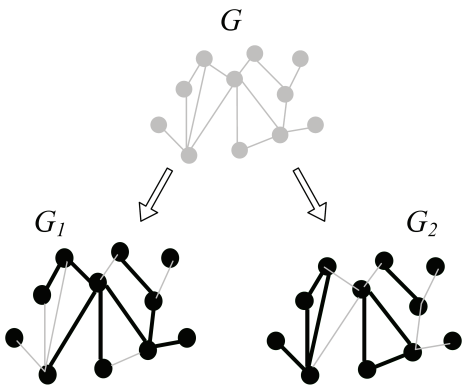

Figure 1: Sampling process applied to the underlying graph G, resulting in the two sampled graphs G_1 and G_2 to be matched.

We reiterate that we do not address the algorithmic aspects of de-anonymization, including the computational complexity of enumerating all mappings and computing their error. Instead, we next show conditions on the model parameters such that minimizing the error function is almost surely equivalent to identifying the correct mapping using only the structures of the two sampled graphs, i.e., we show that *de-anonymization is feasible*, and it is not possible to guarantee anonymity.

4. CONDITIONS FOR PERFECT MATCHING

Following the model introduced in Section 3, we state the main theorem of this paper, followed by its proof.

THEOREM 4.1. *For the $G(n,p;s)$ matching problem with $s = \omega(1/n)$ and $p \to 0$, if*

$$ps\,\frac{s^2}{2-s} = 8\frac{\log n}{n} + \omega(n^{-1}), \tag{2}$$

then the identity permutation π_0 minimizes the error criterion (1) a.a.s[1], yielding perfect matching of the vertex sets of G_1 and G_2. [2]

PROOF. We denote by Δ_0 the error induced by the identity permutation and Δ_π the error induced by the permutation π. Figure 2 depicts two possible mappings between the same G_1 and G_2 shown in Figure 1 corresponding to the identity mapping π_0 and a permutation π_2 (in which all nodes are fixed except two) respectively, together with their error.

To show the result, we define Π_k on V as the set of all permutations that fix $n-k$ nodes and permute k nodes, calling them an order-k permutation. The number of such permutations, referred to as "rencontre numbers", is as follows [33]:

$$|\Pi_k| = R(n, n-k) = \binom{n}{k} \cdot (!k), \tag{3}$$

where $!k$ is the subfactorial of k, denoting the number of permutations of k objects in which no object appears in its natural place. It is easily verified that $R(n,n-k)$ can be upper-bounded as follows:

$$|\Pi_k| = \binom{n}{k} \cdot (!k) \le \binom{n}{k} \cdot \left(\frac{k!}{2}\right) \le n^k. \tag{4}$$

The random variables introduced below are indexed by n, which we omit unless required by the context. We define

$$S_k = \sum_{\pi \in \Pi_k} \mathbf{1}_{\{\Delta_\pi \le \Delta_0\}}.$$

S_k counts the number of order-k permutations for which the number of matching errors is at most that of the identity permutation. Thus, $S = \sum_{k=2}^{n} S_k$ is the total number of false matches. The expected number of errors can be computed as:

$$\begin{aligned}\mathsf{E}[S] &= \sum_{k=2}^{n} \mathsf{E}[S_k] = \sum_{k=2}^{n} \sum_{\pi \in \Pi_k} \mathsf{E}\left[\mathbf{1}_{\{\Delta_\pi \le \Delta_0\}}\right] \\ &= \sum_{k=2}^{n} \sum_{\pi \in \Pi_k} \mathsf{P}\{\Delta_\pi - \Delta_0 \le 0\},\end{aligned}$$

where the expectation is over $G(n,p;s)$.

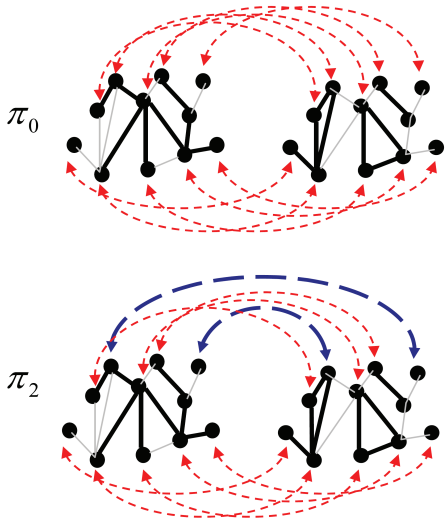

Figure 2: The identity permutation π_0 versus a permutation $\pi_2 \in \Pi_2$ that mismatches $k=2$ vertices for mapping G_1 to G_2. The error in each case corresponds to the number of edges in one graph with the mapped edge not existing in the other graph. Thus, $\Delta_0 = 8$ and $\Delta_{\pi_2} = 10$, where Δ_0 is the edge difference as a result of the sampling process, and Δ_{π_2} is induced by both the sampling process and the wrong mapping of two nodes in π_2.

S counts the total number of non-identity permutations that minimize the error, and we need to show that with high probability no such permutations exist. By the First Moment Method (following Markov's inequality), since S is a non-negative integer-valued random variable, to show that $\mathsf{P}\{S=0\} \to 1$, it suffices to show that $E[S] \to 0$.

Using this method and substituting (4) in the above, it is then sufficient to show that

$$\mathsf{E}[S] \le \sum_{k=2}^{n} n^k \max_{\pi \in \Pi_k} \mathsf{P}\{\Delta_\pi - \Delta_0 \le 0\} \to 0. \tag{5}$$

[1] a.a.s: asymptotically almost surely, i.e., with probability going to 1 as the number of nodes n goes to infinity. In general, *asymptotic* refers to the behavior for $n \to \infty$.

[2] We use the standard asymptotic notation (o, O, ω, Ω, and θ).

We bound the error probability for a fixed order-k permutation π, i.e., we bound the probability term in (5). For permutation π, let V_π be the set of vertices for which $v \neq \pi(v)$, and let $E_\pi = V_\pi \times V$, i.e., the set of possible edges between one or two vertices mismatched under π. Note that every edge satisfying $e \neq \pi(e)$ is in E_π. The inverse is not true, because transpositions in π (a pair (u,v) such that $\pi(u) = v$ and $\pi(v) = u$) induce invariant edges. The cardinality e_k of E_π is

$$e_k = |E_\pi| = \left(\begin{array}{c} k \\ 2 \end{array} \right) + k(n-k),$$

where the first term is the number of unordered node pairs both in V_π, and the second term is the number of unordered node pairs with one node in V_π.

As every edge e in the complement of E_π (i.e., in $(V \times V) - E_\pi$) is by definition invariant under π, they contribute equally to Δ_0 and Δ_π. Therefore, we can write $\Delta_\pi - \Delta_0 = X_\pi - Y_\pi$, where

$$\begin{aligned} X_\pi &= \sum_{e \in E_\pi} |\mathbf{1}_{\{e \in E_\pi^1\}} - \mathbf{1}_{\{\pi(e) \in E_\pi^2\}}|, \\ Y_\pi &= \sum_{e \in E_\pi} |\mathbf{1}_{\{e \in E_\pi^1\}} - \mathbf{1}_{\{e \in E_\pi^2\}}|, \end{aligned} \tag{6}$$

with $E_\pi^{1,2} = E_\pi \cap E(G_{1,2})$, i.e., the set of edges in $G_{1,2}$ incident to at least one mismatched vertex. Here, Y_π is the number of errors for the identity permutation within the set E_π, i.e., the number of sampling errors within E_π. Note that X_π and Y_π are not independent, because they are functions of the same random sets $E_\pi^{1,2}$.

Y_π counts the number of edges in E_π that are sampled in only one of $G_{1,2}$, i.e., the number of sampling errors under the identity permutation. The probability for each possible edge to be in $E(G)$ and exactly one of $G_{1,2}$ is $2ps(1-s)$. Thus Y_π is binomial with probability $2ps(1-s)$.

For X_π, we need to proceed more carefully. Assume π has $\phi \geq 0$ transpositions. First, note that each transposition in π induces one invariant edge $e = \pi(e) = \pi^{-1}(e)$ in E_π (such an edge contributes to X_π with probability $2ps(1-s)$).

The remaining $e_k - \phi$ edges are not invariant under π. Each pair of such edges $(e, \pi(e))$ contributes 1 to X_π if $e \in G_1$ and $\pi(e) \notin G_2$ or vice versa (cf. (6)). The probability for exactly one of two *different* edges in E_π to be sampled is $2ps(1-ps)$. Note that the terms in (6) are dependent, because conditional on $|\mathbf{1}_{\{e \in E_\pi^1\}} - \mathbf{1}_{\{\pi(e) \in E_\pi^2\}}| = 1$, at least one of e or $\pi(e)$ is present in the generator G. Thus, the conditional probability of an adjacent pair (either $(\pi^{-1}(e), e)$ or $(\pi(e), \pi(\pi(e)))$ contributing 1 to (6) is $s(1-ps)$. We conservatively ignore this positive correlation and stochastically lower-bound X_π by assuming that each pair of edges $(e, \pi(e))$ contributes an i.i.d. Bernoulli with parameter $2ps(1-ps)$ to (6).

Thus, X_π is stochastically lower-bounded by the sum of two independent binomials $\mathsf{Bi}(e_k - \phi, 2ps(1-ps)) + \mathsf{Bi}(\phi, 2ps(1-s))$, where ϕ is the number of transpositions in π. By definition, a transposition can occur only between two vertices that are both in V_π. Hence, $\phi \leq \lfloor k/2 \rfloor \leq k/2$.

Thus, we have

$$X_\pi \overset{(stoch.)}{\geq} \mathsf{Bi}(e_k - \lfloor k/2 \rfloor, 2ps(1-ps)) \tag{7}$$

$$Y_\pi \sim \mathsf{Bi}(e_k, 2ps(1-s)). \tag{8}$$

We upper-bound the probability of the event $\{X_\pi - Y_\pi \leq 0\}$ using the following lemma, which holds regardless of dependence between X_π and Y_π:

LEMMA 4.1. *Let X_1 and X_2 be two binomial random variables with means λ_1 and λ_2, where $\lambda_2 > \lambda_1$. Then,*

$$\mathsf{P}\{X_2 - X_1 \leq 0\} \leq 2 \exp\left(-\frac{1}{8} \frac{(\lambda_2 - \lambda_1)^2}{\lambda_2 + \lambda_1} \right). \tag{9}$$

PROOF OF LEMMA. Let X_1 and X_2 be two binomial random variables with means λ_1 and λ_2. The probability of the event $\{X_2 - X_1 \leq 0\}$ can be upper-bounded as follows:

$$\mathsf{P}\{X_2 - X_1 \leq 0\} \leq \mathsf{P}\{X_1 \geq x\} + \mathsf{P}\{X_2 \leq x\}, \tag{10}$$

for any x.

We now find an upper-bound for the right-hand side of (10). We use the Chernoff bounds for the binomial random variables X_1 and X_2 using the following theorem [13]:

If $X \in \mathsf{Bi}(n,p)$ and $\lambda = np$, then,

$$\mathsf{P}\{X > \lambda + t\} \leq \exp\left(-\frac{t^2}{2(\lambda + t/3)} \right), \quad t \geq 0; \tag{11}$$

$$\mathsf{P}\{X < \lambda - t\} \leq \exp\left(-\frac{t^2}{2\lambda} \right), \quad t \geq 0. \tag{12}$$

We upper-bound $\mathsf{P}\{X_1 \geq x\}$ and $\mathsf{P}\{X_2 \leq x\}$ using (11) and (12) (for two arbitrary positive values of t_1 and t_2 respectively). We set $x = (\lambda_1 + \lambda_2)/2$, and thus $t_1 = t_2 = (\lambda_2 - \lambda_1)/2$. Using $\lambda_2 > \lambda_1$ allows to bound the two exponents as follows:

$$\begin{aligned} \mathsf{P}\{X_1 \geq x\} &\leq \exp\left(-\frac{1}{8} \frac{(\lambda_2 - \lambda_1)^2}{\lambda_1 + (\lambda_2 - \lambda_1)/6} \right) \\ &\leq \exp\left(-\frac{1}{8} \frac{(\lambda_2 - \lambda_1)^2}{\lambda_1 + \lambda_2} \right), \end{aligned} \tag{13}$$

and

$$\begin{aligned} \mathsf{P}\{X_2 \leq x\} &\leq \exp\left(-\frac{1}{8} \frac{(\lambda_2 - \lambda_1)^2}{\lambda_2} \right) \\ &\leq \exp\left(-\frac{1}{8} \frac{(\lambda_2 - \lambda_1)^2}{\lambda_1 + \lambda_2} \right). \end{aligned} \tag{14}$$

This completes the proof. □

Now let λ_π and λ_0 denote the means of X_π and Y_π respectively, with values,

$$\lambda_\pi = 2ps(1-ps)(e_k - k/2) \tag{15}$$

$$\lambda_0 = 2pse_k(1-s). \tag{16}$$

Since $0 \leq s \leq 1$, $2 \leq k \leq n$, and $e_k \simeq k(n - k/2)$, to satisfy $\lambda_\pi > \lambda_0$ we need to have,

$$\begin{aligned} &2ps(1-ps)(k(n-k/2) - k/2) > 2psk(n-k/2)(1-s) \\ &\qquad \Longrightarrow s > \left(\tfrac{1-ps}{1-p} \right) \frac{1}{2n-k}, \end{aligned} \tag{17}$$

which will be satisfied for $s = \omega(1/n)$ and $p \to 0$.

Thus, using the above lemma, we obtain,

$$\mathsf{P}\{X_\pi - Y_\pi \leq 0\} \leq 2 \exp\left(-\underbrace{\frac{1}{8} \frac{(\lambda_\pi - \lambda_0)^2}{\lambda_\pi + \lambda_0}}_{f(n,p,k)} \right). \tag{18}$$

Substituting (15) and (16) in (18) yields:

$$f(n,p,k) = \frac{1}{8}\frac{(2ps\,((e_k - k/2) - (e_k - e_k s)))^2}{2ps\,((e_k - k/2) + (e_k - e_k s))}$$
$$= \frac{ps}{4}\frac{((k/2)\,((2n-k)s-1))^2}{(k/2)\,((2n-k)(2-s)-1)}$$

For $s = \omega(1/n)$ we have $(2n-k)s = \omega(1)$. Thus,

$$f(n,p,k) \simeq \frac{ps}{4}\frac{(k/2)\,((2n-k)s)^2}{(2n-k)(2-s)}$$
$$\simeq \frac{ps}{4}\frac{s^2}{2-s}\,k\,(n-k/2)\,. \quad (19)$$

Using (4), (5), and (19), we have,

$$\mathsf{E}[S] \leq 2\sum_{k=2}^{n} n^k \cdot \exp(-f(n,p,k))$$
$$\overset{(a)}{\simeq} 2\sum_{k=2}^{n} n^k \exp\left(-k\left(n-\frac{k}{2}\right)\frac{ps}{4}\cdot\frac{s^2}{2-s}\right)$$
$$\overset{(b)}{\leq} 2\sum_{k=2}^{\infty} \exp\left(k\left(\log n - \frac{nps}{8}\cdot\frac{s^2}{2-s}\right)\right), \quad (20)$$

where (a) is derived using (19), and (b) uses $k \leq n$. The geometric series goes to zero if the first term goes to zero, which is implied by the condition in the statement of the theorem. This completes the proof. □

A more direct approach to prove the result would be to try to condition on a property of the underlying graph G and/or of $G_{1,2}$ that is both asymptotically almost sure, and for which one could show that uniformly over all permutations π, the number of errors is higher than for the identity π_0. It is difficult to identify such a property that would make the second part of the problem tractable. Instead, we show the result using a method commonly employed in the random graph literature [5, 13], which allows us to analyze a fixed permutation π over the full probability space $G(n,p;s)$.

A remarkable aspect of our result is that for fixed similarity parameter s, the condition is $ps = 8c\log n/n$ for some $c(s) > 1$. As expected, $c(s) = (2-s)/s^2$ is monotonically decreasing in $(0,1)$, and $c(1) = 1$. Thus, for an overall edge sampling probability ps of a bit larger than $8\log n/n$, with high probability the identity permutation minimizes the error function and yields the correct mapping. Note that the threshold for connectivity of $G_{1,2} = G(n,ps)$ (and for the disappearance of isolated vertices) is $ps = \log n/n$ [5, 13]. It is obvious that it is impossible to perfectly match a pair of graphs $G_{1,2}$ when at least one of them possesses more than one isolated vertex (as these necessarily give rise to multiple permutations with equal error counts). Therefore, $ps = \log n/n$ is a lower bound for zero-error graph matching using any technique (i.e., any cost function). Our bound for $G(n,p;s)$ matching is therefore tight, up to a constant function of s.

For the case of $s = 1$, the approximate graph matching problem is equivalent to the classical automorphism group problem for random graphs [5]. Specifically, it is known that $G(n,p)$ is asymmetric (has an automorphism group of size one) for $p = \log n/n + \omega(1)$. This suggests that the constant $c(s)$ in our result can be improved upon through more refined bounding techniques. Indeed, we use relatively loose bounds in several places: in particular, we underestimate the mean of X_π quite significantly by ignoring the positive correlation (within each cycle of π) in the terms of (6); also, we assume the worst-case dependence between X_π and Y_π in (10), even though they are in reality positively correlated through the generator G. These bounds are sufficient to show the asymptotic result to within a constant, but more precise techniques akin to those used to show the classical automorphism result may allow to go further. Another obvious extension of our work would employ other generator graph structures such as random regular graphs, small world models, or scale free graphs.

5. NUMERICAL EXPERIMENTS

To be mathematically tractable and parsimonious, our model inevitably embodies several strong assumptions: (i) the underlying graph is a $G = G(n,p)$ random graph, and the edge sets of the "visible" graphs $G_{1,2}$ are sampled (ii) independently and (iii) with identical probability s from G. Despite these assumptions, we believe that our model and our result on anonymity conditions have implications for real networks and scenarios. Although we are unable to explore the validity of our assumptions in full generality, we wish at least to provide some evidence to justify them. First, it is fairly clear that the underlying graph of a social network would possess a structure very different from a $G(n,p)$, as demonstrated in many studies illustrating fascinating properties such as skewed degree distributions and the small-world effect. However, we conjecture that de-anonymizing two networks sampled from a random graph is harder than more "structured" networks. A random graph is in some sense "maximally uniform", and we therefore believe that for other, more realistic hidden graphs G, de-anonymization might in fact be possible under even weaker conditions. This is of course a promising and fascinating area for further research. Second, we consider de-anonymization successful if the error function $\Delta(.)$ has a unique minimum at π_0. We argue that this function is not too sensitive to a non-uniform sampling process over the edge set (i.e., assuming each edge e is sampled with its own sampling rate $s(e)$), provided the sampling process is similar in both graphs, and uncorrelated across edges. This is because the impact of this non-uniformity on X_0 and X_π above would cancel out to a certain extent. On the other hand, if the sampling process to obtain the two samples G_1 and G_2 were very different, then this could make de-anonymization much harder. For example, if the sampling rate over some subset of vertices were atypically large in G_1 compared to the rest, but atypically large for a *different* subset in G_2, then these two high-rate subsets would be likely to be falsely matched. Therefore, it appears that de-anonymization would be quite sensitive to such differences in the sampling process for G_1 and G_2.

While we have not quantified the above argument, it did lead us to explore the stability of the sampling process through some numerical experiments. In order to motivate and illustrate the concept of *similarity* between networks, and also verify the assumption of *independence* in sampling the edges, we present an example of a real social network: an email graph, in which nodes represent email addresses and edges represent message exchanges. The network evolves in time through the observation of new messages that are exchanged.

We consider a dataset of email messages collected at the mail server of EPFL. The dataset includes logs of email exchanges among users on a weekly basis for a period of 75 weeks. In our dataset, the email exchanges *among EPFL users* is considered (i.e., internal EPFL network). The dataset includes snapshots of the network aggregated by week, such that timestamps are in the timescale of weeks (i.e., all messages sent in a particular week have the same timestamp). Using such dataset, we construct the email network of each week for the internal EPFL network.

Having introduced the above, we investigate the similarity between different snapshots of the network, each being a sample of an underlying hidden email network. Note that in order to map real

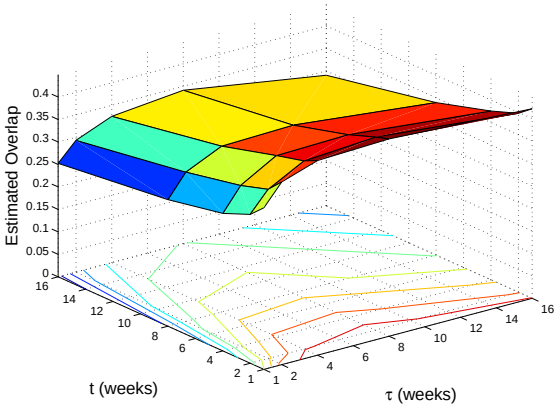

Figure 3: Estimated average edge overlap among overlapped nodes for EPFL internal network, as a function of window size and distance.

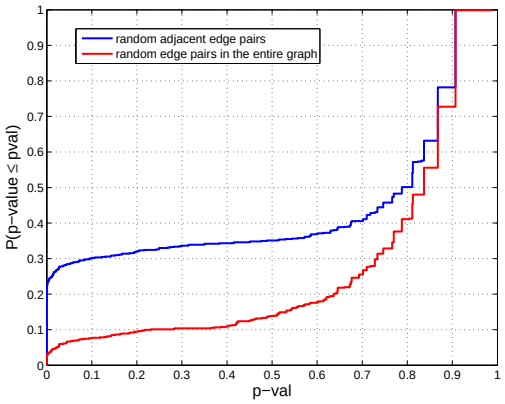

Figure 4: The CDF of the p-values of the Pearson's Chi-Square test for independence, for 1) random adjacent edge pairs (top curve), 2) random edge pairs in the entire graph (bottom curve). Using $\alpha = 0.05$, the test verifies the statistical equivalence of edge pairs in the EPFL dataset.

data to our sampling model, the existence of a hidden underlying graph (including all possible email exchanges over all times) is inevitable - to which we do not have access. However, measuring the amount of *edge overlap* between different snapshots gives us an estimation of the similarity degree between different network samples, or whether the graphs are the outcome of similar sampling processes. Also, since two network snapshots do not contain the same number of nodes necessarily, we estimate the edge overlap as *the proportion of edges among overlapped nodes that exist in both graphs*.

To accomplish the above, we need to pick two networks to be compared. We randomly choose a starting timestamp t_s (week number) in the entire dataset, and construct the first graph starting from t_s accumulated over a window size of τ weeks. For the second graph, we build it starting from timestamp $t_s+\tau+t-1$, again accumulated over a window size of τ, where t denotes the time distance between the two graphs (in weeks). In other words, τ corresponds to the density of the graph (the larger it is, the denser the graph will be), and t implies the time distance between different samples. As an example, $\tau = 1, t = 1$ corresponds to the email network of two consecutive snapshots (each consisting of email exchanges over a one-week period), whereas $\tau = 2, t = 3$ corresponds to two graphs, each consisting of email exchanges over a period of $\tau = 2$ weeks, with a time distance of $t = 3$ weeks. Finally, for each value of τ and t, we repeat the random choice of the networks 30 times and compute the average.

Figure 3 depicts the estimated average edge overlap as a function of the windows size and time distance. It can be observed that the estimated edge overlap is quite significant, and it also exhibits a small increase as τ increases and t decreases, which matches intuition since it is expected that two larger and denser networks have more overlap, and as the samples are farther apart the overlap decreases. However, this change is small over a wide span of the density and distance values. Thus, the graph similarity is fairly robust over different densities and distances. The experiment shows that two graphs sampled from a hidden underlying graph (a hidden overall email network in this case) are similar in structure (even if the sampling processes are non-uniform), with the sampling process being quite stable over different intervals.

Finally, we verify the assumption of independent edge sampling in our model, through looking at the correlation among the edges. In general, the emergence of an edge might correlate with the existence of other edges. In order to investigate how far the independence assumption is from reality, we examine edge correlation in the EPFL internal network. To do so, we choose a random pair of edges from the final accumulated graph (i.e., $\tau = 75$), and examine their joint appearance in 75 weekly snapshots. We use the Chi-Square test for independence to determine whether there is a significant relationship between the appearance of the two edges. We assume a null hypothesis that two randomly chosen edges e_1 and e_2 appear independently, and use the Pearson χ^2 test to decide whether we should reject the null hypothesis, separately for each set of 75 edge pair appearances. We compute the χ^2 test statistics of each sample set of 75 weeks as $X^2 = \sum \frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$, $i, j = 0, 1$, where i denotes the existence (1) or non-existence (0) mode of e_1 (similarly $j = 0, 1$ for e_2), $O_{i,j}$ is the observed frequency count of e_1 at mode i and e_2 at mode j, and $E_{i,j} = n_i * n_j/n$, n_i being the total number of sampled observations of e_1 at mode i (similarly n_j for e_2 at mode j) and n being the total number of samples (75). The p-value is calculated as $P\left\{X^2 \leq \chi^2(1)\right\}$, where $\chi^2(1)$ is a Chi-Square random variable with one degree of freedom, as the number of bins for each categorial variable equals 2. We derive the p-value of the test and reject the independence hypothesis if the p-value is smaller than the significance level ($\alpha = 0.05$ in our tests). Repeating this for a large number of random edge pairs (752 in our experiments), we find that 93% of the edge pairs are statistically indistinguishable (p-value $> \alpha = 0.05$).

To strengthen our test even further, we do the same experiment above, by choosing random pairs of *adjacent* edges - i.e., edges incident to the same node - thinking that such edges might express a high correlation. We find that even in this case, most edge pairs (72%) are statistically independent. Figure 4 depicts the CDF of p-values found for each selected pair over 75 weeks, for both experiments. The plot clearly shows that in most cases, p-value is greater than α, as mentioned above.

Finally, we repeat the above experiments for triple edges , i.e. choosing three random edges in the accumulated graph, the null hypothesis being that three randomly chosen edges e_1, e_2 and e_3 appear independently. Again, we consider two cases, one where the edges are chosen randomly in the entire graph, and the other further correlated version where the edges are sampled from the set of 3-chains in the graph, i.e. paths of length 3. Figure 5 depicts the CDF of the p-values for each selected triple over 75 weeks, for

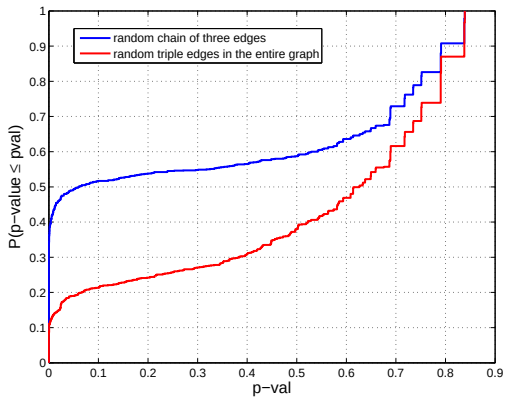

Figure 5: The CDF of the p-values of the Pearson's Chi-Square test for independence, for 1) random chain of three edges (top curve), 2) random triple edges in the entire graph (bottom curve). Using $\alpha = 0.05$, the test verifies the statistical equivalence of triple edges in the EPFL dataset.

both experiments, repeated 1000 times. It is observed that 81% of the random triple edges are independent. Further, for the correlated chain of edges, we observe that 50% of the random 3-chains are statistically indistinguishable. Further experiments show that as the number of randomly chosen edges increases, there will be a higher dependence for their joint appearance, as expected.

Our results suggest that the independence assumption clearly does not hold generally, but many small sets of edges do behave independently. To what extent the i.i.d. assumption built into our model is realistic, in the sense that it would correctly predict the boundary of privacy in real networks, is a subject of further investigation.

6. DISCUSSION AND CONCLUSION

In this paper, we considered the privacy issue in social networks and investigated the possibility of de-anonymization from a mathematical perspective. We defined the problem in the context of approximate graph matching, with the goal of finding the correct mapping between the node sets of two structurally similar graphs. Using ideas from graph sampling in modeling evolution of networks, we proposed a probabilistic model to derive two sampled versions of an underlying graph as "noisy" versions of the networks to be matched. Elaborating our model for the case of random graphs, we proved that using the simplest matching criterion based only on network topology, a perfect matching between the nodes can be established with high probability as the network size grows large, under simple conditions for the sampling process. More specifically, we proved that a surprisingly mild condition on the scaling of the expected degree with the number of nodes is sufficient for de-anonymization to be feasible. For this, we expressed lower bounds for the sampling probability, or more intuitively, the extent of overlap in the edges of two graphs, so that it yields perfect matching.

Two conditions in our theorem are $s = \omega(1/n)$ and $ps \to 0$. How these parameters relate to real networks is of course a crucial and interesting question. Social networks tend to be sparse ($p \to 0$), and a reasonable assumption may be to assume a fixed average node degree ($p = c/n$), as the number of contacts is usually the result of local interactions that should not be influenced by the rest of the network[3]. The scaling of s is more debatable, as it depends on the nature of the two networks. If G_1 and G_2 capture the social interactions between a set of people using different methods (e.g., email and phone calls), then it would make sense to postulate a constant s independent of the size of the network, as the choice of method (i.e., generating a link) would be a purely local one, and therefore not influenced by the rest of the network. However, more cross-domain data should be studied to verify this.

Our result shows that given a specific cost function $\Delta(.)$, a pair of correlated graphs can be perfectly matched under certain conditions. An interesting question would be the converse: can we find conditions such that no cost function could give a match? In the $G(n, p; s)$ model, it is straightforward to show such a converse of the form $ps = o(\log n/n)$, as alluded to before. In this case, G_1 and G_2 would have isolated vertices a.a.s., and obviously no method would be able to determine the correct matching among these. More precise converses, as well as variations of our model (e.g., assuming other generator graphs G) are the topic of future work.

Our work implies the feasibility of de-anonymization of a target network by using the structural similarity of a known auxiliary network, and raises privacy concerns about sharing the simplest topological information of users with partners and third-party applications. One consequence of our work might be guidelines on how to release or share only sampled versions of networks, by enforcing the sparsity constraint to guarantee anonymity. This would be promising provided such a thinned-out network would still provide enough information for the task at hand.

In future, we intend to generalize our approach to a broader class of graphs. As discussed above, we conjecture that in some sense, a random graph as the generator G may be more difficult than a more "structured" graph. On the other hand, the i.i.d. sampling process in our model is an idealistic assumption, and the impact of relaxing it should be explored. Finally, and perhaps most importantly, while this paper proves the *existence* of the perfect matching using the proposed error function, the algorithmic complexity of searching in such a vast space is still an open problem.

Acknowledgments

We are indebted to Mohamed Kafsi and Patrick Thiran for fruitful discussions and feedback on a draft of the manuscript. We gratefully acknowledge financial support for this project from the Swiss National Science Foundation (SNSF).

7. REFERENCES

[1] AOL Search Data Scandal. `http://en.wikipedia.org/wiki/AOL_search_data_scandal`.

[2] W. Aiello, F. Chung, and L. Lu. A Random Graph Model for Massive Graphs. In *STOC '00*, pages 171–180, 2000.

[3] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore Art Thou R3579X?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography. In *WWW '07*, pages 181–190, 2007.

[3]Note however that recent work on network densification [21] observes a dependence of the form $p = cn^{\alpha-2}$, for some $\alpha > 1$, i.e., an average degree that grows as $n^{\alpha-1}$ with the size of the network. While the underlying causes of densification are still being debated [31], it is still the case that $p \to 0$. Note that for fixed s, a network densifying in this way would always be non-anonymous, as $n^{\alpha-2} = \omega(\log n/n)$.

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex Networks: Structure and Dynamics. *Physics Reports*, 424(4-5):175–308, 2006.

[5] B. Bollobas. *Random Graphs (2nd edition)*. Cambridge University Press, 2001.

[6] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of Human Mobility on Opportunistic Forwarding Algorithms. *IEEE Transactions on Mobile Computing*, 6:606–620, 2007.

[7] M. Gori, M. Maggini, and L. Sarti. Exact and Approximate Graph Matching Using Random Walks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(7):1100–1111, 2005.

[8] R. Gross and A. Acquisti. Information Revelation and Privacy in Online Social Networks. In *WPES '05: Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80, 2005.

[9] S. Guha, K. Tang, and P. Francis. NOYB: Privacy in Online Social Networks. In *WOSN'08: Workshop on Online Social Networks*, 2008.

[10] M. Hay, G. Miklau, D. Jensen, D. Towsley, and C. Li. Resisting Structural Re-Identification in Anonymized Networks. *VLDB Journal*, 19(6), December 2010.

[11] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On Near-Uniform URL Sampling. In *Proceedings of the 9th international World Wide Web conference on Computer networks*, pages 295–308, 2000.

[12] S. Hill, F. Provost, and C. Volinsky. Network-based Marketing: Identifying Likely Adopters via Consumer Networks. *Statistical Science*, 21:256–276, 2006.

[13] S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*. Wiley, 2000.

[14] B. Karrer and M. E. J. Newman. Random Graph Models for Directed Acyclic Networks. *Physical review. E*, 2009.

[15] A. Korolova, R. Motwani, S. U. Nabar, and Y. Xu. Link Privacy in Social Networks. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 289–298, 2008.

[16] G. Kossinets, J. Kleinberg, and D. Watts. The Structure of Information Pathways in a Social Communication Network. In *KDD '08*, pages 435–443, 2008.

[17] B. Krishnamurthy and C. Willis. Characterizing Privacy in Online Social Networks. In *WOSN'08: Workshop on Online Social Networks*, 2008.

[18] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-Tracking and the Dynamics of the News Cycle. In *KDD '09*, pages 497–506, 2009.

[19] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker Graphs: An Approach to Modeling Networks. *J. Mach. Learn. Res.*, 11:985–1042, 2010.

[20] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting Positive and Negative Links in Online Social Networks. In *WWW '10*, pages 641–650, 2010.

[21] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In *KDD '05*, pages 177–187, 2005.

[22] B. Luo and E. R. Hancock. Structural Graph Matching Using the EM Algorithm and Singular Value Decomposition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(10):1120–1136, 2001.

[23] D. Martin and A. Schulman. Deanonymizing Users of the SafeWeb Anonymizing Service. In *Proceedings of the 11th USENIX Security Symposium*, pages 123–137, 2002.

[24] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the Flickr Social Network. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 25–30, 2008.

[25] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, 2007.

[26] R. Myers, R. C. Wilson, and E. R. Hancock. Bayesian Graph Edit Distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(6), 2000.

[27] A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. In *SP '08: Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pages 111–125, 2008.

[28] A. Narayanan and V. Shmatikov. De-anonymizing Social Networks. *Security and Privacy, IEEE Symposium on*, 0:173–187, 2009.

[29] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(Suppl 1):2566–2572, 2002.

[30] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. A (Sub)Graph Isomorphism Algorithm for Matching Large Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(10):1367–1372, 2004.

[31] P. Pedarsani, D. R. Figueiredo, and M. Grossglauser. Densification Arising from Sampling Fixed Graphs. In *SIGMETRICS '08*, pages 205–216, 2008.

[32] Y. Richter, E. Yom-Tov, and N. Slonim. Predicting Customer Churn in Mobile Networks through Analysis of Social Groups. In *Proc. SIAM Int'l Conference on Data Mining (SDM) 2010*, 2010.

[33] J. Riordan. *An Introduction to Combinatorial Analysis*. Wiley, 1958.

[34] A. Sanfeliu and K. Fu. A Distance Measure between Attributed Relational Graphs for Pattern Recognition. *IEEE Transactions On Systems, Man, and Cybernetics*, 13(3):353–362, 1983.

[35] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On Unbiased Sampling for Unstructured Peer-to-Peer Networks. *IEEE/ACM Trans. Netw.*, 17(2):377–390, 2009.

[36] Y. Tian and J. M. Patel. TALE: A Tool for Approximate Large Graph Matching. *Data Engineering, International Conference on*, 0:963–972, 2008.

[37] J. R. Ullmann. An Algorithm for Subgraph Isomorphism. *J. ACM*, 23(1):31–42, 1976.

[38] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel. A Practical Attack to De-Anonymize Social Network Users. In *IEEE Symposium on Security & Privacy*, 2010.

[39] B. Zhou and J. Pei. Preserving Privacy in Social Networks Against Neighborhood Attacks. In *ICDE '08: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 506–515, 2008.