We are IntechOpen,
the world's leading publisher of
Open Access books
Built by scientists, for scientists

4,500
Open access books available

119,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

A Review of Several Privacy Violation Measures for Large Networks under Active Attacks

Tanima Chatterjee, Nasim Mobasheri and Bhaskar DasGupta

Abstract

It is by now a standard practice to use the concepts and terminologies of network science to analyze social networks of interconnections between people such as Facebook, Twitter and LinkedIn. The powers and implications of such social network analysis are indeed indisputable; for example, such analysis may uncover previously unknown knowledge on community-based involvements, media usages and individual engagements. However, all these benefits are not necessarily cost-free since a malicious individual could compromise privacy of users of these social networks for harmful purposes that may result in the disclosure of sensitive data that may be linked to its users. A natural way to avoid this consists of an "anonymization process" of the relevant social network. However, since such anonymization processes may not always succeed, an important research goal is to quantify and measure how much privacy a given social network can achieve. Toward this goal, some recent research works have aimed at evaluating the resistance of a social network against active privacy-violating attacks by introducing and studying a new and meaningful privacy measure for social networks. In this chapter, we review both theoretical and empirical aspects of such privacy violation measures of large networks under active attacks.

Keywords: social networks, privacy measure, active attacks, (k, ℓ)-anonymity, algorithmic complexity

1. Introduction

In recent years, social networks have become an indisputable part of people's lives. The emergence of such networks has altered how we interact with the world. A given individual's day-to-day activities like media consumption, job hunting and social interaction have changed, along with how businesses and other beneficial entities interact with them through marketing, advertising, and information diffusion. This has led to an unstoppable race of collecting information and interaction from social networks by researchers, governments, and business entities for various purposes. From a research point of view, social networks and their interaction mechanisms provide valuable insight in many fields of study, such as sociology, psychology, advertising, and recommendation systems. It is only natural that the information contained in these networks and the value they hold have been and will be targeted by bad actors for malicious activities. The importance of these networks

and the value of information that can be retrieved from them have led social network researchers to take a closer look at methods to combat such bad actors as well as formulate network measures that can provide an insight to the privacy of these networks. In this survey, we will look at one such measure known as (k, ℓ)-anonymity [1] and will discuss some theoretical and empirical results regarding this measure.

1.1 Overview of the paper

Given the irrefutable importance of social networks in our daily lives and the ever increasing risk of compromising valuable personal data through privacy attacks against these networks, it is preferable to know how secure a given social network is against privacy attacks. This necessitates a deeper look into the types of privacy attacks and how to cope with them. There is an extensive literature on privacy preserving computational models in variety of application areas such as multi-party communications or distributed computing settings [2–6]. In this chapter, we focus on a specific type of attack known as *background-based active attack* and one measure that reflects the resistance of any given network against such attacks. The organization of the rest of the paper is as follows:

- In Section 2 we briefly discuss the notion of privacy in social networks and review some literature on privacy violating attacks on social networks. We also introduce the (k, ℓ)-anonymity privacy measure and some corresponding network measurement which are the basis for this measure.

- In Section 3 we review some basic terminologies and notations that will be used in formulation of the three problems introduced in Section 4.

- Section 4 contains three problems that arise from theoretical investigation of the (k, ℓ)-anonymity.

- Section 5 contains the results of an empirical study on the resistance of real-world social networks.

- Finally, we end this chapter with some concluding remarks in Section 6.

2. Privacy measures in social networks

We begin by discussing the mathematical structure that fit the most to represent social networks. A social network is often portrayed as a graph [7, 8] $G = (V, E)$ where V is a set of nodes representing the social members, and E is the set of edges portraying the relationship among these members. Both nodes and edges may have extra attributes, such as weights, that provide extra information about the nature of these social bonds (*e.g.*, trust or popularity); however, throughout this survey we will consider the simplest form of graphs, namely undirected and unweighted graphs, to model our social networks.

As we discussed in the previous section, the information that the social networks provide are invaluable. Due to the very nature of many social network applications, the identity of the members or the nature of relationship between members is quite sensitive and valuable. Thus, when releasing a social network we want to remove any attributes that may help identify these kinds of sensitive data. Assuming all members and their relationships are of high sensitivity, preventing *identity disclosure*

or *link disclosure* becomes an important task. One popular method to prevent such disclosures is *anonymization*. In an anonymization process, we publish the network without identifying the corresponding nodes or potentially identifiable attributes. Even after anonymizing the network, we will still be releasing many informative attributes encoded by the network structure; for example, attributes such as node degree, connectivity, or other similar graph properties can still help the adversaries in compromising the user privacies of a published network.

Adversaries usually rely on background knowledge to compromise the privacy of published anonymized social networks. For understanding the failure of current privacy preservation methods such as anonymization, we need to have a proper model for the adversary background knowledge. Although it's challenging to have a comprehensive model of all possible types of adversary background knowledge, it is very useful to model the background knowledge via structural properties of networks such as node degrees, embedded subgraphs, node neighbors, etc. [9]. Backstrom et al. [10] were the first to introduce a category of attacks on anonymized social graphs. The models introduced in [10] are background-based attacks and are *widely* used in privacy analysis of social networks. The two main types of attacks are as follows.

1. *Passive attacks* in which the adversary will *not* modify the network by injecting new nodes, but instead will use the structural knowledge to detect the location of a *known* node. In this type of attacks, the adversary can benefit from the fact that most nodes in real social networks often belong to a small uniquely identifiable subgraph [10]. An adversary can then build a coalition with members of such subgraphs and attempt to re-identify the subgraphs in the anonymized published network, thus compromising the privacy of neighboring nodes.

2. *Active attacks* in which the adversary will choose an arbitrary set of target users, create new nodes and insert them into a social network in a way that they are connected to the target set and they form a distinguishable subgraph. After the anonymized version of the social network is published, the adversary can then use the subgraph as a *fingerprint* to re-identify the targeted users and compromise their privacy.

The authors in [10] also showed that it *is* possible to compromise the privacy of any social network of n nodes with high probability using *only* $O(\sqrt{\log n})$ attacker nodes. In a *passive attack*, adversary's structural knowledge will give her/him a global view of the network depending on the global structure of the network. It could pose a high privacy risk if an adversary were to combine this global view with the local structural knowledge obtained using an active attack. As an example, consider the network in **Figure 1**. If we only have global structural knowledge, it is not possible to differentiate the nodes v_3 and v_4 (*e.g.* , same node degrees, *etc.*). However, controlling just one extra node in the graph, such as the node v_1, provides local structural knowledge such as distances between nodes, and using the knowledge of the distance of v_1 from v_3 and v_4 ($d_{v_1,v_3} = 1$ and $d_{v_1,v_4} = 2$) one can easily differentiate node v_3 from node v_4.

There are several well-studied strategies for coping with active attacks on a social network [9, 11, 12] via addressing the anonymization process of the social network. However, in this chapter we will focus on a measure that evaluates how resistant a social network is against this type of privacy attack. Introduced by Trujillo-Rasua et al. [1], (k, ℓ)-anonymity is a novel and, to the best of our knowledge, the *only* privacy measure examining the structural resistance of a given graph against active attacks. The (k, ℓ)-anonymity is a measure based on metric representation of nodes, where k is a privacy threshold and l is the maximum number of

attacker nodes that may be inserted in the network. It was shown in [1] that graphs satisfying (k, ℓ)-anonymity can successfully deter adversaries controlling at most l nodes in the graph from re-identifying nodes with probability higher than $\frac{1}{k}$.

2.1 (k, ℓ)-anonymity

The (k, ℓ)-anonymity measure is based on a concept known as k-metric anti-dimension of graphs. To facilitate further discussions about the measure, we first introduce some notations and terminologies. For a simple connected graph $G = (V, E)$, where V is set of nodes and E is set of edges, let $dist_{v_i, v_j}$ denote distance (*i.e.*, number of edges in a shortest path) between the nodes v_i and v_j. Given and ordered set of nodes $S = \{v_1, \ldots, v_t\}$ and a node u we define the metric representation of u with respect to S as a vector $\mathbf{d}_{u,-S} = (dist_{u,v_1}, \ldots, dist_{u,v_t})$. Metric representations of nodes are closely related to the concept of a *resolving set* of a graph. Inspired by the problem of identifying an intruder in a network and introduced separately by Slater [13] and by Harary and Melter [14], a resolving set of graph provides recognition of every pair of nodes in graph.

Definition 1 (resolving set). Given a graph $G = (V, E)$, a subset $S \subseteq V$ is called a resolving set for G if, for each pair of nodes $(u, v) \in G$, there exist a node $x \in S$ such that $dist_{x,u} \neq dist_{x,v}$. A smallest-cardinality resolving set is called the metric basis, and its cardinality is referred to as the metric dimension of G.

The concepts of metric representation and resolving set inspired the introduction of another network measure known as *k-antiresolving set* that will be used as the founding base for (k, ℓ)-anonymity.

Definition 2 (k-antiresolving set). Given a graph $G = (V, E)$, $S \subset V$ is called a k-antiresolving set of G if k is the largest integer such that, for every node $v \in V \backslash S$, there exist at least $k - 1$ nodes $u_1, u_2, \ldots, u_{k-1} \in V \backslash S$ with the same metric representation with respect to S as v.

A k-antiresolving set of *minimum* cardinality is called a *k-antiresolving basis*, and its cardinality denotes the *k-metric antidimension* $adim_k(G)$ of G. Note that the k-antiresolving set may not exist for every k in a graph.

The (k,l)-anonymity measure is built upon the k-antiresolving set concept. Assume the adversary has gained control of a subset S of nodes in the graph G, where S is a k-antiresolving set for G. Then the adversary *cannot* uniquely re-identify any node based on the background knowledge (namely, the knowledge of metric representation of a node v with respect to S) with probability higher than $\frac{1}{k}$. (k, ℓ)-anonymity is formally defined as [1].
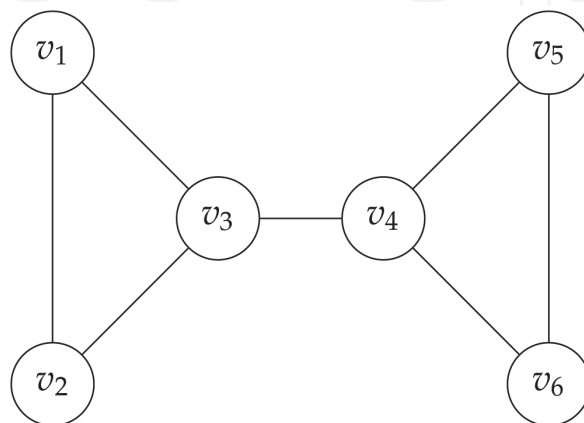

Figure 1.
A simple graph G *used in Section 2 to illustrate the high risk posed by combining knowledge gained by active and passive attacks.*

Definition 3 ((k, ℓ)-anonymity). A graph G under active attack satisfies (k, ℓ)-anonymity if k is the smallest positive integer so that the k-metric antidimension of G is less than or equal to l.

In the above definition, k is a parameter depicting the privacy threshold and l represents the maximum number of attacker nodes. It is safe to assume that number of attacker nodes l is significantly smaller than number of nodes present in the network as injecting attacker nodes or gaining control of existing nodes is difficult without being detected [15].

3. Basic terminologies and notations

For the exposition in the remainder of this chapter, we will need some notations and terminologies which we introduce here. Consider the (undirected unweighted) graph G in **Figure 2**. We will use this graph to illustrate the terminologies and notations that are introduced.

- The metric representation of node v_i is denoted by
  $\mathbf{d}_{v_i} = (dist_{v_1,v_i}, dist_{v_2,v_i}, \dots, dist_{v_n,v_i})$.

  ○ For example, in **Figure 2**, $\mathbf{d}_{v_1} = (0, 1, 2, 3, 3, 2)$

- The diameter of G is the length of the longest shortest path and is denoted by
  $diam(G) = \max_{v_i, v_j \in V} \left\{ dist_{v_i, v_j} \right\}$.

  ○ For example, in **Figure 2**, $diam(G) = 3$.

- The open neighborhood of node v_i is a subset of all nodes directly connected to v_i and denoted by $\mathsf{Nbr}(v_i) = \{v_j | \{v_i, v_j\} \in E\}$.

  ○ For example, in **Figure 2**, $\mathsf{Nbr}(v_2) = \{v_1, v_3, v_6\}$.

- The metric representation of a node v_i with respect to a subset such as $S \subset V$ is denoted by $\mathbf{d}_{v_i, -S}$.

  ○ For example, in **Figure 2**, $\mathbf{d}_{v_1, -\{v_3, v_4\}} = (2, 3)$.
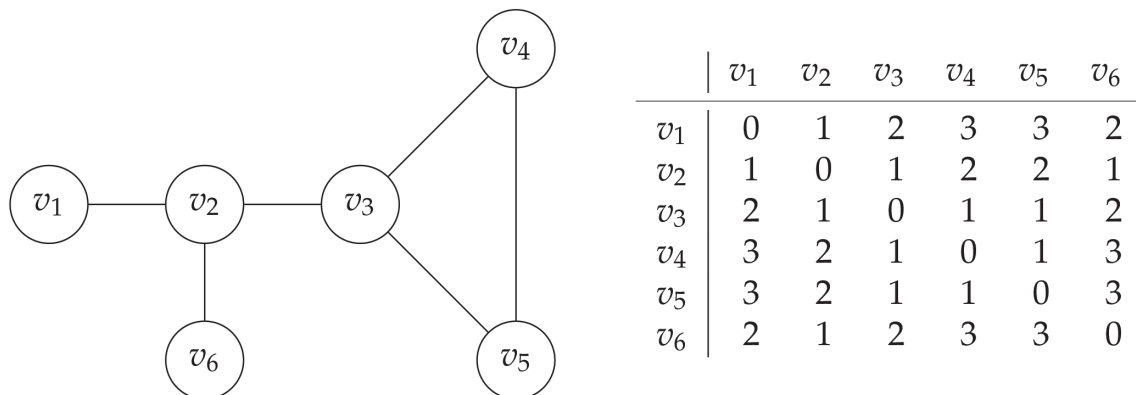

	v_1	v_2	v_3	v_4	v_5	v_6
v_1	0	1	2	3	3	2
v_2	1	0	1	2	2	1
v_3	2	1	0	1	1	2
v_4	3	2	1	0	1	3
v_5	3	2	1	1	0	3
v_6	2	1	2	3	3	0

Figure 2.
An example used in Section 3 for illustrating various notations.

- We can expand the previous notation to reflect the metric representation of a subset of nodes $V' \subset V$ S with respect to S as $\mathcal{D}_{V',-S} = \{ \mathbf{d}_{v_l,-S} \,|\, v_l \in V' \}$.

  - For example, in **Figure 2**, $\mathcal{D}_{\{v_1,v_2\},-\{v_3,v_4\}} = \{ (2,3), (1,2) \}$. Note that the first pair (2,3) corresponds to v_1 and the second pair (1,2) corresponds to v_2.

- We define a partition $\prod = \{V_1, \ V_2, \dots, V_t\}$ of $V' \subseteq V$ as one with the following properties:

  - $\bigcup_{i=1}^{t} V_i = V'$, and

  - for all $i \neq j$, $V_i \cap V_j = \emptyset$.

- We define a refinement $\prod' = \{V_1', V_2', \dots, V_\ell'\}$ of a partition $\prod$, denoted by $\prod' \prec_r \prod$, as one that can be obtained from $\prod$ using the following rules:

  - For every node $v_j \in \left(\bigcup_{i=1}^{t} V_i\right) \backslash \left(\bigcup_{i=1}^{\ell} V_i'\right)$, remove v_j from the set in $\prod$ that contains it.

  - Optionally, for every set V_ℓ in $\prod$, replace V_ℓ by a partition of V_ℓ.

  - If there exists an empty set, remove it.

    i. For example, in **Figure 2**, $\{\{v_1,v_2\}, \{v_3\}, \{v_5\}\} \prec_r \{\{v_1,v_2\}, \{v_3,v_4,v_5\}\}$.

- We define an *equivalence relation* (and related notations) over set of same-length vectors $\mathcal{D}_{V\backslash V',-V'}$ for some $\emptyset \subset V' \subset V$ as follows:

  - The set $_{\text{of}}$ equivalence classes, which forms a partition of $\mathcal{D}_{V\backslash V',-V'}$, is denoted by $\prod_{V\backslash V',-V'}^{=}$

    i. For example, in **Figure 2**, $\prod_{\{v_1,v_2,v_6\},-\{v_3,v_5\}}^{=} = \{(2,3), (1,2), (2,3)\}$.

  - We declare two nodes $v_i, v_j \in V\backslash V'$ to be in the same equivalence class if $\mathbf{d}_{v_i,-V'}$ and $\mathbf{d}_{v_j,-V'}$ belong to the same equivalence class in $\prod_{V\backslash V',-V'}^{=}$; thus $\prod_{V\backslash V',-V'}^{=}$ also defines a partition into equivalence classes of $V\backslash V'$.

  - The *measure* of the equivalence relation is defined as

    $$\mu\left(\mathcal{D}_{V\backslash V',-V'}\right) \overset{\text{def}}{=} \min_{y \in \prod_{V\backslash V',-V'}^{=}} \{|y|\}.$$

  - If a set S is a k-antiresolving set then $\mathcal{D}_{V\backslash S,-S}$ defines a partition into equivalence classes of measure k.

    i. For example, in **Figure 2**, $\mu\left(\mathcal{D}_{\{v_1,v_2,v_6\},-\{v_3,v_5\}}\right) = 1$ and $\{v_3,v_5\}$ is a 1-antiresolving set.

4. Theoretical results

To understand graph resistance against privacy attacks, one needs to study the (k, ℓ)-anonymity in greater details. Thus, we look into some computational problems related to this measure that were formalized and investigated in [16]. This section contains three problems from [16] and the respective algorithms to solve each problem efficiently. It is important to note that (k, ℓ)-anonymity in its basic definition sets no limitation for the adversary, which means that an adversary can take control of as many nodes as she/he can. However, in real world there are many mechanisms designed solely to prevent such attacks and thus the chances of being caught are significantly high. This notion is the motivation behind several problems with respect to measuring the (k, ℓ)-anonymity in a graph.

We now state the three problems for analyzing (k, ℓ)-anonymity. Problem 1 simply checks to find a k-antiresolving set for the largest possible value of k. Problem 2 sets a restriction for number of nodes the adversary can control and attempts to find the largest possible value of k while minimizing the number of nodes that are compromised. Problem 3 introduces a version of the problem that attempts to address the trade-off between privacy threshold and number of compromised nodes.

Problem 1 (metric antidimension ($ADIM$)). Find a k-antiresolving subset of nodes S that maximizes k.

Problem 1 assumes there are *no* limitations on the number of attacker nodes, thus finding an absolute bound for privacy violation. Note that solution to Problem 1, denoted by k_{opt}, shows that, given no bound on number of the nodes an adversary can control, it is feasible to uniquely re-identify k_{opt} nodes with probability $\frac{1}{k_{opt}}$. The assumptions in Problem 1 are rarely plausible in practice; due to mechanisms present to counter such attacks, the more nodes the adversary controls, the higher the risk of being exposed. Thus, a limit on number of attacker nodes is necessary, which leads us to Problem 2.

Problem 2 ($k_{\geq}$-metric antidimension ($ADIM_{\geq k}$)). Given k, find a k'-antiresolving set S such that (i) $k' >= k$ and, (ii) S is of minimum cardinality.

Problem 2 is an extension to Problem 1 that attempts to find the largest value of k while minimizing the number of attacker nodes. A solution to this problem asserts few interesting statements. For example, an adversary controlling l attacker nodes where $\ell < |\mathcal{L}_{opt}^{\geq k}|$ cannot uniquely re-identify any node in the network with a probability better than $\frac{1}{k}$. However, using enough number of nodes ($\geq |\mathcal{L}_{opt}^{\geq k}|$) one can re-establish such possibilities.

The third problem focuses on a trade-off between number of attacker nodes and the privacy violation probability. Given two measures (k, ℓ)-anonymity and (k', ℓ')-anonymity where $k' > k$ and $\ell' < \ell$, it is easy to observe that (k', ℓ')-anonymity measure provides a smaller privacy violation probability but also has lower tolerance for attacker nodes. The trade-off leads us to the third problem.

Problem 3 ($k_{=}$-metric antidimension ($ADIM_{=k}$)) Given a positive integer k, find a k antiresolving subset of nodes S with minimum cardinality if such a subset exists.

Chatterjee et al. [16] investigated Problems 1–3 from a computational complexity perspective. The following theorems summarizes their finding on Problems 1–3. The non-trivial mathematical proofs for these theorems are unfortunately outside of the scope of this chapter; we strongly recommend readers who are interested in the proofs to read the original paper [16].

Theorem 1. [16]

1. Both *ADIM* and $ADIM_{\geq k}$ can be solved in $O(n^4)$ time.

2. Both *ADIM* and $ADIM_{\geq k}$ can also be solved in $O\left(\frac{n^4 \log n}{k}\right)$ time with high probability.

Theorem 2. [16]

1. $ADIM_{=k}$ is NP-Complete for any k in the range $1 \leq k \leq n^\varepsilon$ where $0 \leq \varepsilon < \frac{1}{2}$ is any arbitrary constant, even if the diameter of the input graph is 2.

2. Assuming NP$\not\subseteq$DTIME$\left(n^{\log \log n}\right)$, there exists a universal constant $\delta > 0$ such that $ADIM_{=k}$ does not admit $\left(\frac{1}{\delta} \ln n\right)$ approximation for any integer k in the range $1 \leq k \leq n^\varepsilon$ for any constant $0 \leq \varepsilon < \frac{1}{2}$, even if the diameter of the input graph is 2.

3. If $k = n - c$ for some constant c then $ADIM_{=k}$ can be solved in polynomial time.

Theorem 3. [16]

1. $ADIM_{=1}$ admits $(1 + \ln(n - 1))$ approximation in $O(n^3)$ time.

2. If G has at least one node of degree 1 then $ADIM_{=1}$ can be solved in $O(n^3)$ time.

3. If G does not contain a cycle of 4 edges then $ADIM_{=1}$ can be solved in $O(n^3)$ time.

4.1 Algorithms

The following algorithms were devised in [16] to address Problems 1–3. It is important to note that *ADIM* can be solved in $O(n^5)$ time by repeatedly solving $ADIM_{\geq k}$ for $k = n - 1, n - 2, \ldots, 1$ to find the largest obtainable value for k such that $\mathcal{L}_{opt}^{\geq k} < \infty$. However, few modifications to Algorithm 1 directly result in $O(n^4)$ solution, which is shown in Algorithm 2.

5. Empirical results

In [18], DasGupta et al. investigated the resistance of 8 real-world network against active attacks with respect to the (k, ℓ)-anonymity. All the networks under investigation were unweighted graphs and the direction of edges (if the network was directed) was ignored during the analysis. **Table 1** contains the general information regarding these networks. Results for both *ADIM* and $ADIM_{\geq k}$ were obtained by running Algorithm 1 on the networks, the return statements from Algorithm 1 being an exact solution to Problem 2. On the other hand, the exact solution for Problem 1 can be achieved by combining Algorithm 1 and binary search on k to find the largest value of k such that $V_{opt}^{\geq k} \neq \emptyset$ [18].

Algorithm 1: $O(n^4)$ time deterministic algorithm for $ADIM_{\geq k}$ [16]

---

1   Compute $\mathbf{d}_i$ for all $i = 1, 2, \ldots, n$ in $O(n^3)$ time using Floyd-Warshall algorithm [17]

2   $\widehat{\mathcal{L}_{opt}^{\geq k}} \leftarrow \infty$ $\widehat{V_{opt}^{\geq k}} \leftarrow \emptyset$

3   **for** *each $v_i \in V$* **do**

4      $V' = \{v_i\}$; done $\leftarrow$ FALSE

5   **while** *($V \setminus V' \neq \emptyset$) $\wedge$ ($\neg$ done)* **do**

6   Compute $\mu(\mathcal{D}_{V \setminus V', -V'})$

7   **if** $\mu(\mathcal{D}_{V \setminus V', -V'}) \geq k$ **and** $|V'| < \widehat{\mathcal{L}_{opt}^{\geq k}}$ **then**

8            $\widehat{\mathcal{L}_{opt}^{\geq k}} \leftarrow V'$; $\widehat{V_{opt}^{\geq k}} \leftarrow V'$; done $\leftarrow$ TRUE

9   **else**

10   let $V_1, \ldots, V_l$ be the only $\ell > 0$ equivalence classes

11   in $\prod =_{V \setminus V', -V'}$ such that

12            $|V_1| = \ldots = |V_l| = \mu(\mathcal{D}_{V \setminus V', -V'})$

13            $V' \leftarrow V' \cup \left(\cup_{t=1}^{\ell} V_t \right)$

14   **end**

15   **end**

16   **end**

17   **return** $\widehat{\mathcal{L}_{opt}^{\geq k}}$ *and* $\widehat{V_{opt}^{\geq k}}$ *as our solution*

Algorithm 2: $O(n^4)$ time deterministic algorithm for *ADIM* [16]

---

1   Compute $\mathbf{d}_i$ for all $i = 1, 2, \ldots, n$ in $O(n^3)$ time using Floyd-Warshall algorithm [17]

2   $\widehat{V_{opt}^{\geq k}} \leftarrow \emptyset$; $\widehat{k_{opt}} \leftarrow 0$

3   **for** *each $v_i \in V$* **do**

4      $V' = \{v_i\}$

5   **while** $V \setminus V' \neq \emptyset$ **do**

6   compute $\mu(\mathcal{D}_{V \setminus V', -V'})$

7   **if** $\mu(\mathcal{D}_{V \setminus V', -V'}) > \widehat{k_{opt}}$ **then**

8            $\widehat{k_{opt}} \leftarrow \mu(\mathcal{D}_{V \setminus V', -V'})$

9            $\widehat{V_{opt}^{\geq k}} \leftarrow V'$

10   **else**

11   let $V_1, \ldots, V_\ell$ be the only $\ell > 0$ equivalence classes

12   in $\prod_{V \setminus V', -V'}^{=}$ such that

13            $|V_1| = \ldots = |V_l| = \mu(\mathcal{D}_{V \setminus V', -V'})$

14            $V' \leftarrow V' \cup \left(\cup_{t=1}^{\ell} V_t \right)$

15   **end**

16   **end**

17   **end**

18   **return** $\widehat{k_{opt}}$ *and* $\widehat{V_{opt}^{\geq k}}$ *as our solution*

Algorithm 3: (*resp.* **Algorithm 4**) $O(\frac{n^4 log n}{k})$ time randomized algorithm for $ADIM_{\geq k}$ (*resp.* ADIM) [16]

---

1   Compute $\mathbf{d}_i$ for all $i = 1, 2, \ldots, n$ in $O(n^3)$ time using Floyd-Warshall algorithm [17]

2   $\widehat{\mathcal{L}_{opt}^{\geq k}} \leftarrow \infty$ $\widehat{V_{opt}^{\geq k}} \leftarrow \emptyset$ (for $ADIM_{\geq k}$)

     or

     $\widehat{V_{opt}^{\geq k}} \leftarrow \emptyset$; $\widehat{k_{opt}} \leftarrow 0$ (for $ADIM$)

3   **repeat**

4   Select a node v_i uniformly at random from the n nodes

5   execute step 4 to step 15 of Algorithm 1   (for $ADIM_{\geq k}$)

        or

     execute step 4 to step 16 of Algorithm 2   (for $ADIM$)

6   **until** $\lceil \frac{2n \ln n}{k} \rceil$ *times*

7   **return** *the best of all solutions found in the previous steps*

Algorithm 4: $O(n^3)$ time $(1 + ln(n - 1))$-approximation algorithm for $ADIM_{=1}$ [16]

---

1 Compute $\mathbf{d}_i$ for all $i = 1, 2, \ldots, n$ in $O(n^3)$ time using Floyd-Warshall algorithm [17]

2 $\widehat{L_{opt}^{=1}} \leftarrow \infty$; $\widehat{V_{opt}^{=1}} \leftarrow \emptyset$
   (** *Guessing that set $\{v_i\}$ belongs to $\prod_{V \setminus V_{opt}^{=1}, -V_{opt}^{=1}}^{=}$* **)

3 **for** *each $v_i \in V$* **do**

4 $\quad$ create an instance of standard set cover problem containing $n - 1$ elements and $n - 1$ sets:

5 $\quad$ $\mathcal{U} = \{ a_{v_j} \mid v_j \in V \setminus \{v_j\}\}$

6 $\quad$ $S_{v_j} = \{ a_{v_j} \} \cup \{ a_{v_l} | dist_{v_i, v_j} \neq dist_{v_l, v_j} \}$ for $j \in \{1, 2, ..., n\} \setminus \{i\}$

7 $\quad$ **if** $\cup_{j \in \{1,2,...,n\} \setminus \{i\}} S_{v_j} = \mathcal{U}$ **then**

8 $\quad\quad$ run the greedy approximation algorithm in [19] for the instance of the set cover problem giving a solution $\mathcal{J} \subseteq \{1, 2, ..., n\} \setminus \{i\}$

9 $\quad\quad$ $V' = \{ v_j \mid j \in \mathcal{J} \}$

10 $\quad\quad$ **if** $|V'| < \widehat{\mathcal{L}_{opt}^{=1}}$ **then**

11 $\quad\quad\quad$ $\widehat{\mathcal{L}_{opt}^{=1}} \leftarrow |V'|$

12 $\quad\quad\quad$ $\widehat{V_{opt}^{=1}} \leftarrow V'$

13 $\quad\quad$ **end**

14 $\quad$ **end**

15 **end**

16 **return** $\widehat{\mathcal{L}_{opt}^{=1}}$ *and* $\widehat{V_{opt}^{=1}}$ *as our solution*

The results for both Problem 1 and Problem 2 for the networks in **Table 1** are depicted in **Table 2**. Results in **Table 2** provide the following interesting insights with respect to resistance against privacy attacks in real-world social networks.

- All networks, with the exception of "Enron Email Data" network, will have a significant percentage of their users compromised if an adversary gains control of *only* one node (varying between 2.6% of users compromised in "University Rovira i Virgili emails" network to 26.5% of users compromised in "Zachary Karate Club" network).

Name	Number of nodes	Number of edges	Description
Zachary Karate Club [20]	34	78	Network of friendship between 34 members of a karate club
San Juan Community [21]	75	144	Network for visiting relations between families living in farms in San Juan Sur, Costa Rica, 1948
Jazz Musician Network [22]	198	2842	A social network of jazz musicians
University Rovira i Virgili emails [23]	1133	10903	The network of email interchanges between members of university
Enron Email Data Set [24]	1088	1767	Enron email network
Email Eu Core [25]	986	24989	Emails from a large European research institute
UC Irvine College Message platform [26]	1896	59835	Messages on a Facebook-like platform at UC-Irvine
Hamsterster friendships [27]	1788	12476	Friendships between users of the website

Table 1.
List of 8 social networks studied in [18].

Name	n	k_{opt}	$p_{\text{opt}} = \frac{1}{k_{\text{opt}}}$	$\mathcal{L}^{\geq k_{\text{opt}}}_{\text{opt}} = \mathcal{L}^{=k_{\text{opt}}}_{\text{opt}}$	$\frac{k_{\text{opt}}}{n}$
Zachary Karate Club [20]	34	9	0.111	1	26.5%
San Juan Community [21]	75	7	0.143	1	9.3%
Jazz Musician Network [22]	198	12	0.084	1	6.0%
University Rovira i Virgili emails [23]	1133	29	0.035	1	2.6%
Enron Email Data Set [24]	1088	153	0.007	935	14.1%
Email Eu Core [25]	986	39	0.026	1	3.4%
UC Irvine College Message platform [26]	1896	55	0.019	1	2.9%
Hamsterster friendships [27]	1788	4	0.25	1	0.22%

n depict the number of nodes, k_{opt} is the largest value of k such that $V^{\geq k}_{opt} \neq \emptyset$, and $\mathcal{L}^{\geq k_{opt}}_{opt}$ is minimum number of attacker nodes for corresponding k.
[a]*n denotes the number of nodes in the social graph.*
[b]*k_{opt} is the largest value of k such that $V^{\geq k}_{opt} \neq \emptyset$.*

Table 2.
Results for ADIM *using Algorithm 1 [18].*

	k	4	5	10	20	40	60	100	120	153
Enron Email Data Set	$p_k = \frac{1}{k}$	0.25	0.2	0.1	0.05	0.025	0.017	0.01	0.009	0.007
	$\mathcal{L}^{\geq k}_{opt}$	1	334	463	567	683	842	935	935	935

Table 3.
$\mathcal{L}^{\geq k}_{opt}$ *values recorded for $k > 1$ for the "Enron Email Data" network [18]. The values shown are subject to $\mathcal{L}^{\geq k}_{opt} \neq \mathcal{L}^{\geq k-1}_{opt}$.*

- For all networks with the exception of "Enron Email Data" network, the minimum privacy violation probability is notably higher than 0 (varying between 0.019 for the "UC Irvine College Message platform" network to 0.25 for the "Hamsterster friendships" network). The value for minimum privacy violation probability in "Hamsterster friendships" network is notably higher compare to all other networks.

- In comparison to other networks, the "Zachary Karate Club" and the "San Juan Community" have higher percentage of their users compromised if subjected to a privacy attack.

The exception network is the "Enron Email Data" network which due to a high value of $\mathcal{L}^{\geq k}_{opt}$ is very resilient against an attack. An adversary needs to control at least 86% of the network to achieve a value of $p_{opt} = 0.007$, which is not feasible in practice. This interesting observation in the "Enron Email Data" network motivated further inspections in different values of k. As shown in **Table 3**, $\mathcal{L}^{\geq k}_{opt}$ in the "Enron Email Data" network does not decrease significantly until k is set to a much smaller value compare to k_{opt}, which further emphasizes that **violating the privacy of the "Enron Email Data" network is not guaranteed in practice.** The authors in [18] also investigated the (k, ℓ)-anonymity measure in *synthetic* networks constructed based on both Erdös-Rényi random graphs and Barabási-Albert scale-free networks. We refer the reader to the original paper for more information.

6. Conclusions

Since their emergence about a decade ago, social networks have rapidly grown and infiltrated every aspect of our daily lives. With rapidly expanding reliance on their platforms, social networks like Facebook and Twitter are becoming a goldmine of personal information and user behavior data which makes the study of these networks of prime importance. The valuable information stored within these platforms makes them the target of malicious entities which try to compromise the privacy of the users which may further lead to unwanted disclosure of the sensitive attributes of the network.

In this chapter, we have reviewed a novel privacy measure that quantifies the resistance of a large social network against a privacy violating attack. We reviewed some efficient algorithms to compute this measure in social graph and revisited the privacy violation properties in 8 real-world networks. The current theoretical and empirical results for (k, ℓ)-anonymity pave the way for further investigation of this measure, as well as addressing its shortcomings and limitations.

Acknowledgements

Author details

Tanima Chatterjee[†], Nasim Mobasheri[†] and Bhaskar DasGupta[*†]
Department of Computer Science, University of Illinois at Chicago,
Chicago, IL, USA

*Address all correspondence to: bdasgup@uic.edu

[†]These authors contributed equally.

IntechOpen

References

[1] Trujillo-Rasua R, Yero IG. k-metric antidimension: A privacy measure for social graphs. Information Sciences. 2016;**328**:403-417

[2] Bar-Yehuda R, Chor B, Kushilevitz E, Orlitsky A. Privacy, additional information and communication. IEEE Transactions on Information Theory. 1993;**39**(6):1930-1943

[3] Comi M, DasGupta B, Schapira M, Srinivasan V. On communication protocols that compute almost privately. Theoretical Computer Science. 2012; **457**:45-58

[4] Feigenbaum J, Jaggard AD, Schapira M. Approximate privacy: Foundations and quantification. In: Proceedings of the 11th ACM Conference on Electronic Commerce; ACM. 2010. pp. 167-178

[5] Kushelvitz E. Privacy and communication complexity. SIAM Journal on Discrete Mathematics. 1992; **5**(2):273-284

[6] Yao AC. Some complexity questions related to distributive computing (preliminary report). In: Proceedings of the 11th Annual ACM Symposium on Theory of Computing; ACM. 1979. pp. 209-213

[7] Newman ME. The structure and function of complex networks. SIAM Review. 2003;**45**(2):167-256

[8] Albert R, Barabási AL. Statistical mechanics of complex networks. Reviews of Modern Physics. 2002; **74**(1):47

[9] Zhou B, Pei J, Luk W. A brief survey on anonymization techniques for privacy preserving publishing of social network data. ACM SIGKDD Explorations Newsletter. 2008;**10**(2): 12-22

[10] Backstrom L, Dwork C, Kleinberg J. Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. In: Proceedings of the 16th International Conference on World Wide Web; ACM. 2007. pp. 181-190

[11] Netter M, Herbst S, Pernul G. Analyzing privacy in social networks— An interdisciplinary approach. In: 2011 IEEE 3rd International Conference on Privacy, Security, Risk and Trust and 2011 IEEE 3rd International Conference on Social Computing; IEEE. 2011. pp. 1327-1334

[12] Wu X, Ying X, Liu K, Chen L. A survey of privacy-preservation of graphs and social networks. In: Managing and Mining Graph Data; Springer, Boston, MA. 2010. pp. 421-453

[13] Slater PJ. Leaves of trees. Congressus Numerantium. 1975;**14** (549-559):37

[14] Harary F, Melter RA. On the metric dimension of a graph. Ars Combinatoria. 1976;**2**(191-195):1

[15] Yu H, Gibbons PB, Kaminsky M, Xiao F. Sybillimit: A near-optimal social network defense against sybil attacks. In: 2008 IEEE Symposium on Security and Privacy; IEEE. 2008. pp. 3-17

[16] Chatterjee T, DasGupta B, Mobasheri N, Srinivasan V, Yero IG. On the computational complexities of three problems related to a privacy measure for large networks under active attack. Theoretical Computer Science. 2019; **775**:53-67

[17] Leiserson CE, Rivest RL, Cormen TH, Stein C. Introduction to Algorithms. Cambridge, MA: MIT Press; 2001

[18] DasGupta B, Mobasheri N, Yero IG. On analyzing and evaluating privacy

measures for social networks under active attack. Information Sciences. 2019;**473**:87-100

[19] Johnson DS. Approximation algorithms for combinatorial problems. Journal of Computer and System Sciences. 1974;**9**(3):256-278

[20] Zachary WW. An information flow model for conflict and fission in small groups. Journal of Anthropological Research. 1977;**33**(4):452-473

[21] Loomis CP, Morales JO, Clifford RA, Leonard OE. Turrialba: Social Systems and the Introduction of Change. Glencoe, IL: Free Press; 1953

[22] Gleiser PM, Danon L. Community structure in jazz. Advances in Complex Systems. 2003;**6**(04):565-573

[23] Guimera R, Danon L, Diaz-Guilera A, Giralt F, Arenas A. Self-similar community structure in a network of human interactions. Physical Review E. 2003;**68**(6):065103

[24] Enron email network. Available from: UC Berkeley Enron Email Analysis website http://bailando.sims.berkeley.edu/enron_email.html

[25] Paranjape A, Benson AR, Leskovec J. Motifs in temporal networks. In: Proceedings of the 10th ACM International Conference on Web Search and Data Mining; ACM. 2017. pp. 601-610

[26] Panzarasa P, Opsahl T, Carley KM. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. Journal of the American Society for Information Science and Technology. 2009;**60**(5):911-932

[27] Hamsterster friendships network dataset–KONECT, April 2017. Available from: http://konect.uni-koblenz.de/networks/petster-friendships-hamster)