On the Computational Complexities of Three Problems Related to a Privacy Measure for Large Networks Under Active Attack

Tania Chatterjee\textsuperscript{a,1}, Bhaskar DasGupta\textsuperscript{a,1,*}, Nasim Mobasheri\textsuperscript{a,1}, Venkatkumar Srinivasan\textsuperscript{a,1}, Ismael G. Yero\textsuperscript{b,2}

\textsuperscript{a}Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA
\textsuperscript{b}Departamento de Matemáticas, Escuela Politécnica Superior, Universidad de Cádiz, 11202 Algeciras, Spain

Abstract

With the arrival of modern internet era, large public networks of various types have come to existence to benefit the society as a whole and several research areas such as sociology, economics and geography in particular. However, the societal and research benefits of these networks have also given rise to potentially significant privacy issues in the sense that malicious entities may violate the privacy of the users of such a network by analyzing the network and deliberately using such privacy violations for deleterious purposes. Such considerations have given rise to a new active research area that deals with the quantification of privacy of users in large networks and the corresponding investigation of computational complexity issues of computing such quantified privacy measures. In this paper, we formalize three natural problems related to such a privacy measure for large networks and provide non-trivial theoretical computational complexity results for solving these problems. Our results show the first two problems can be solved efficiently, whereas the third problem is provably hard to...
solve within a logarithmic approximation factor. Furthermore, we also provide computational complexity results for the case when the privacy requirement of the network is severely restricted, including an efficient logarithmic approximation.

Keywords: Privacy measure, social networks, active attack, computational complexity

2010 MSC: 68Q25, 68W25, 05C85

1. Introduction

Social networks have become an important center of attention in our modern information society by transforming human relationships into a huge interchange of, very often, sensitive data. There are many truly beneficial consequences when social network data are released for justified mining and analytical purposes. For example, researchers in sociology, economics and geography, as well as vendors in service-oriented systems and internet advertisers can benefit and improve their performances by a fair study of the social network data. But, such benefits are not free of cost as dishonest individuals or organizations may compromise the privacy of its users while scrutinizing a public social network and may deliberately use such privacy violations for harmful or other unfair commercial purposes. A common way to handle this kind of unwelcome intrusion on the user’s privacy is to somehow anonymize the data by removing most potentially identifying attributes. However, even after such anonymization, often it may still be possible to infer many sensitive attributes of a social network that may be linked to its users, such as node degrees, inter-node distances or network connectivity, and therefore further privacy-preserving methods need to be investigated and analyzed. These additional privacy-preserving methods of social networks are based on the concept of $k$-anonymity introduced for microdata in [10], aiming to ensure that no record in a database can be re-identified with a probability higher than $1/k$. 

Crucial to modelling a social network anonymization process are the ad-
versary’s background knowledge of any object and the structural information about the network that is available. For example, assuming the involved social network as a simple graph in which individuals are represented by nodes and relationships between pairs of individuals are represented by edges, the adversary’s background knowledge about a target (a node) could be the node degree [12], the node neighborhood [24], etc. In such scenarios, it frequently suffices to develop attacks to re-identify the individuals and their relationships. Such attacks are usually called passive (see [14] for more information). Some examples of passive attacks and the corresponding privacy-preserving methods for social networks can be found in references [12, 24, 25].

In contrast, Backstrom et al. introduced the concept of the active attacks in [1]. Such attacks are mainly based on creating and inserting in a network some nodes (the “attacker nodes”) under control by the adversary. These attacker nodes could be newly created accounts with pseudonymous or spoofed identities (commonly called Sybil nodes), or existing legitimate individuals in the network which are in the adversary’s proximity. The goal is to establish links with some other nodes in the network (or even links between other nodes) in order to create some sort of “fingerprints” in the network that will be further released. Clearly, once the releasing action has been achieved, the adversary could retrieve the fingerprints already introduced, and use them to re-identify other nodes in the network. Backstrom et al. in [1] showed that $O(\sqrt{\log n})$ attacker nodes in a network could in fact seriously compromise the privacy of any arbitrary node.

In recent years, several research works have appeared that deal with decreasing the impact of these active attacks (see, for instance, [20]). For other related publications on privacy-preserving methods in social networks, see [15, 21, 24].

There are already many well-known active attack strategies for social networks in order to find all possible vulnerabilities. However, somewhat surprisingly, not many prior research works have addressed the goal of measuring how resistant is a given social network against these kinds of active attacks to the privacy. Very recently a novel privacy measure for social networks was introduced in [18]. The privacy measure proposed there was called the $(k, \ell)$-anonymity,
where $k$ is a number indicating a privacy threshold and $\ell$ is the maximum number of attacker nodes that can be inserted into the network; $\ell$ may be estimated through some statistical methods. As claimed by Trujillo-Rasua and Yero in [18], graphs satisfying $(k, \ell)$-anonymity can prevent adversaries who control at most $\ell$ nodes in the network from re-identifying individuals with probability higher than $\frac{1}{k}$. This privacy measure relies on a graph parameter called the $k$-metric anti-dimension.

Consider a simple connected unweighted graph $G = (V, E)$ and let $\text{dist}_{u,v}$ be the length (number of edges) of a shortest path between two nodes $u, v \in V$. For an ordered set $S = u_1, \ldots, u_t$ of nodes of $G$ and a node $v \in V$, the vector $\mathbf{d}_v, - S = (\text{dist}_{v, u_1}, \ldots, \text{dist}_{v, u_t})$ is called the metric representation of $v$ with respect to $S$. Based on the above definition, an ordered set $S \subset V$ of nodes is called a $k$-anti-resolving set for $G$ if $k$ is the largest positive integer such that for every node $v \in V \setminus S$ there exist at least $k-1$ different nodes $v_1, \ldots, v_{k-1} \in V \setminus S$ such that $\mathbf{d}_v, - S = \mathbf{d}_{v_1, - S} = \cdots = \mathbf{d}_{v_{k-1}, - S}$, i.e., $v$ and $v_1, \ldots, v_{k-1}$ have the same metric representation with respect to $S$. The $k$-metric anti-dimension of $G$, denoted by $\text{adim}_k(G)$, is the minimum cardinality of any $k$-anti-resolving set in $G$. Note that $k$-anti-resolving sets may not exist in a graph for every $k$.

The connection between $(k, \ell)$-anonymity privacy measure and the $k$-metric anti-dimension can be understood in the following way. Suppose that an adversary takes control of a set of nodes $S$ of the graph (i.e., $S$ plays the role of attacker nodes), and the background knowledge of such an adversary regarding a target node $v$ is the metric representation of the node $v$ with respect to $S$. The $(k, \ell)$-anonymity privacy measure is a privacy metric that naturally evolves from the adversary’s background knowledge. Intuitively, if $S$ (the attacker nodes of an adversary) is a $k$-anti-resolving set then the adversary cannot uniquely re-identify other nodes in the network (based on the metric representation) from

\footnote{Note that other different privacy notions with the same name also exists, e.g., Feder and Nabar in [6] investigated $(k, \ell)$-anonymity where $\ell$ represented the number of common neighbors of two nodes.}
these attacker nodes with a probability higher than $1/k$ (based on uniform sampling of other nodes), and if the $k$-metric anti-dimension of the graph is $\ell$ then the adversary must use at least $\ell$ attacker nodes to get the probability of privacy violation down to $1/k$.

1.1. Organization of the Paper

It is desirable to know how secure a given social network is against active attacks. This necessitates the study of computational complexity issues for computing $(k,\ell)$-anonymity. Currently known results only include some heuristic algorithms with no provable guarantee on performances such as in [18], or algorithms for very special cases. In fact, it is not even known if any version of the related computational problems is $\text{NP}$-hard. We formalize three computational problems related to measuring the $(k,\ell)$-anonymity of graphs and present non-trivial computational complexity results for these problems. The rest of the paper is organized as follows:

- In Section 2 we review some basic terminologies and notations and present the three computational problems that we consider in this paper. For the benefit of the reader, we also briefly review some standard algorithmic complexity concepts and results that will be used later.
- In Section 3 we state the results in this paper mathematically precisely along with some informal remarks. We group our results based on the problem definitions and the expected size of the attacker nodes.
- Sections 4–6 are devoted to the proofs of the results stated in Section 3.
- We finally conclude in Section 8 with some possible future research directions.

Historical note on the results of this paper While our paper was still under submission/review, Zhang and Gao in [23] independently and without knowing our results provided an alternate proof to the $\text{NP}$-completeness result reported in Theorem 2(a) using a different reduction.
2. Basic Terminologies, Notations and Problem Definitions

In this section, we first describe the terminologies and notations required to describe our computational problems, and subsequently describe several versions of the problems we consider.

2.1. Basic Terminologies and Notations

Let $G = (V, E)$ be our undirected unweighted input graph over $n$ nodes $v_1, v_2, \ldots, v_n$. We use $\text{dist}_{v_i, v_j}$ to denote the distance (number of edges in a shortest path) between nodes $v_i$ and $v_j$. For illustrating various notations, we use the example in Fig. 1.

- $\mathbf{d}_{v_i} = (\text{dist}_{v_i, v_1}, \text{dist}_{v_i, v_2}, \ldots, \text{dist}_{v_i, v_n})$. For example, $\mathbf{d}_{v_2} = (1, 0, 2, 1, 2)$.

- $\text{diam}(G) = \max_{v_i, v_j \in V} \{ \text{dist}_{v_i, v_j} \}$ is the diameter (length of a longest shortest path) of the graph $G = (V, E)$. For example, $\text{diam}(G) = 3$.

- $\text{Nbr}(v_i) = \{ v_j \mid \{v_i, v_j\} \in E \}$ is the (open) neighborhood of node $v_i$ in $G = (V, E)$. For example, $\text{Nbr}(v_2) = \{v_1, v_4\}$.

- For a subset of nodes $V' \subseteq V$ and any $v_i \in V \setminus V'$, $\mathbf{d}_{v_i, -V'}$ denotes the metric representation of $v_i$ with respect to $V'$, i.e., the vector of $|V'|$ elements obtained from $\mathbf{d}_{v_i}$ by deleting $\text{dist}_{v_i, v_j}$ for every $v_j \in V \setminus V'$. For example, $\mathbf{d}_{v_2, -\{v_1, v_3\}} = (1, 2)$.

- $\mathcal{D}_{V'',-V'} = \{ \mathbf{d}_{v_i, -V'} \mid v_i \in V'' \}$ for any $V'' \subseteq V \setminus V'$. For example, if $V'' = \{v_2, v_4\}$ then $\mathcal{D}_{V'',-\{v_1, v_3\}} = \{(1, 2), (2, 1)\}$.
\[ \Pi = \{ V_1, V_2, \ldots, V_k \} \] is a partition of \( V' \subseteq V \) if and only if \( \bigcup_{i=1}^{k} V_i = V' \) and \( V_i \cap V_j = \emptyset \) for \( i \neq j \).

Partition \( \Pi' = \{ V'_1, V'_2, \ldots, V'_\ell \} \) is called a refinement of partition \( \Pi \), denoted by \( \Pi' \prec_r \Pi \), provided \( \bigcup_{t=1}^{\ell} V'_t \subset \bigcup_{t=1}^{k} V_t \) and the following two conditions are satisfied:

i. All the sets in \( \Pi' \) are pairwise disjoint.

ii. There exists a total and surjective function \( f : \{ 1, \ldots, \ell \} \mapsto \{ 1, \ldots, t \} \) such that \( \forall j \in \{ 1, \ldots, \ell \} : V'_j \subseteq V_{f(j)} \).

For example, if \( \Pi = \{ \{ v_1, v_2 \}, \{ v_3, v_4, v_5 \} \} \) and \( \Pi' = \{ \{ v_1, v_2 \}, \{ v_3 \}, \{ v_4 \} \} \) then \( \Pi' \prec_r \Pi \).

The equality relation over a set of vectors, all of same length, defines an equivalence relation. The following notations are used for such an equivalence relation over the set of vectors \( D \setminus V' \setminus -V' \) for some \( \emptyset \subset V' \subset V \).

The set of equivalence classes, which forms a partition of \( D \setminus V' \setminus -V' \), is denoted by \( \Pi_{D \setminus V' \setminus -V'} \). For example,
\[ \Pi_{\{ v_1, v_2, v_3 \}, -\{ v_4, v_5 \}} = \left\{ \{ (2, 3) \}, \{ (1, 2) \}, \{ (1, 1) \} \right\}. \]

Abusing terminologies slightly, two nodes \( v_i, v_j \in V \setminus V' \) will be said to belong to the same equivalence class if \( d_{v_i, -V'} \) and \( d_{v_j, -V'} \) belong to the same equivalence class in \( \Pi_{D \setminus V' \setminus -V'} \), and thus \( \Pi_{D \setminus V' \setminus -V'} \) also defines a partition into equivalence classes of \( V \setminus V' \). For example, \( \Pi_{\{ v_1, v_2, v_3 \}, -\{ v_4, v_5 \}} \) will also denote \( \left\{ \{ v_1 \}, \{ v_2 \}, \{ v_3 \} \right\} \).

The measure of the equivalence relation is defined as \( \mu(D \setminus V', -V') \) \( \overset{\text{def}}{=} \min_{Y \in \Pi_{D \setminus V' \setminus -V'}} \left\{ |Y| \right\} \). Thus, if a set \( S \) is a \( k \)-anti-resolving set then \( D \setminus S, -S \) defines a partition into equivalence classes whose measure is exactly \( k \). For example, \( \mu(D_{\{ v_1, v_2, v_3 \}, -\{ v_4, v_5 \}}) = 1 \) and \( \{ v_4, v_5 \} \) is a 1-anti-resolving set.

---

4Our definition is slightly different from the standard definition of refinement since we allow \( \bigcup_{i=1}^{\ell} V'_i \subset \bigcup_{i=1}^{k} V_i \).
2.2. Problem Definitions

It is desirable to know how secure a given social network is against active attacks. This necessitates the study of computational complexity issues for computing $(k, \ell)$-anonymity. We formalize three computational problems related to measuring the $(k, \ell)$-anonymity of graphs. For all the problem versions, let $G = (V, E)$ be the (connected undirected unweighted) input graph representing the social network under study.

**Problem 1 (metric anti-dimension or Adim)).** Given $G$, find a subset of nodes $V'$ that maximizes $\mu (D_{V \setminus V', V'})$.

**Notation related to Problem 1**

$k_{\text{opt}} = \max_{\emptyset \subset V' \subset V} \left\{ \mu (D_{V \setminus V', -V'}) \right\}$.

Problem 1 simply finds a $k$-anti-resolving set for the largest possible $k$. Intuitively, it sets an absolute bound on the privacy violation probability of an adversary assuming that the adversary can use any number of attacker nodes. In practice, however, the number of attacker nodes employed by the adversary may be limited, which leads us to the second problem formulation stated below.

**Problem 2 ($k_{\geq} -$metric anti-dimension or Adim$_{\geq k}$).** Given $G$ and a positive integer $k$, find a subset of nodes $V'$ of minimum cardinality such that $\mu (D_{V \setminus V', -V'}) \geq k$, if such a $V'$ exists.

**Notation and assumption related to Problem 2**

$L_{\text{opt}}^{\geq k} = \left| V_{\text{opt}}^{\geq k} \right| = \min \left\{ |V'| \left| \mu (D_{V \setminus V', -V'}) \geq k \right. \right\}$ for some $\emptyset \subset V_{\text{opt}}^{\geq k} \subset V$. If $\mu (D_{V \setminus V', -V'}) \geq k$ for no $V'$ then we set $L_{\text{opt}}^{\geq k} = \infty$ and $V_{\text{opt}}^{\geq k} = \emptyset$.

Problem 2 finds a $k$-anti-resolving set for a given $k$ while simultaneously minimizing the number of attacker nodes.

The remaining third version of our problem formulation relates to a trade-off between privacy violation probability and the corresponding minimum number of attacker nodes needed to achieve such a violation. To understand this motivation, suppose that $G$ has a $k$-metric anti-dimension of $\ell$, a $k'$-metric anti-dimension of $\ell'$, $k' > k$ and $\ell' < \ell$. This provides a trade-off between privacy...
and number of attacker nodes, namely we may allow a smaller privacy violation probability \( \frac{1}{k'} \) but the network can tolerate adversarial control of a fewer number \( \ell' \) of nodes or we may allow a larger privacy violation probability \( \frac{1}{k} \) but the network can tolerate adversarial control of a larger number \( \ell \) of nodes. Such a trade-off may be crucial for a network administrator in administering privacy of a network or for an individual in its decision to join a network. Clearly, this necessitates solving a problem of the following type.

**Problem 3 (k-\( \varepsilon \)-metric antidimension or \( \text{Adim}_{\varepsilon = k} \)).** Given \( G \) and a positive integer \( k \), find a subset of nodes \( V' \) of minimum cardinality such that \( \mu (D_{V\setminus V', V'}) = k \), if such a \( V' \) exists.

**Notation and assumption related to Problem 3**

\[
L_{\text{opt}} = k = |V_{\text{opt}}| = \min \left\{ |V'| \mid \mu (D_{V\setminus V', V'}) = k \right\} \text{ for some } \emptyset \subset V_{\text{opt}} \subset V. \text{ If } \mu (D_{V\setminus V', V'}) = k \text{ for no } V' \text{ then we set } L_{\text{opt}} = \infty \text{ and } V_{\text{opt}} = \emptyset.
\]

2.3. **Standard Algorithmic Complexity Concepts and Results**

For the benefit of the reader, we summarize the following concepts and results from the computational complexity theory domain. We assume that the reader is familiar with standard \( O \), \( \Omega \), \( o \) and \( \omega \) notations used in asymptotic analysis of algorithms (e.g., see [4]).

An algorithm \( A \) for a minimization (resp., maximization) problem is said to have an approximation ratio of \( \varepsilon \) (or is simply an \( \varepsilon \)-approximation) if \( A \) runs in polynomial time in the size of its input and produces a solution with an objective value no larger than \( \varepsilon \times \) (resp., no smaller than \( \frac{1}{\varepsilon} \times \)) the value of the optimum. \( \text{DTIME}(n^{\log \log n}) \) refers to the class of problem that can be solved by a deterministic algorithm running in \( (n^{\log \log n}) \) time when \( n \) is the size of the input instance; it is widely believed that \( \text{NP} \not\subseteq \text{DTIME}(n^{\log \log n}) \).

The minimum set-cover problem (Sc) is a well-known combinatorial problem that is defined as follows [1] [9]. Our input is an universe \( U = \{a_1, a_2, \ldots, a_n\} \) of \( n \) elements, and a collection of \( m \) sets \( S_1, S_2, \ldots, S_m \subseteq U \) over this universe with \( \bigcup_{j=1}^{m} S_j = U \). A valid solution of Sc is a subset of indices \( I \subseteq \{1, 2, \ldots, m\} \).
such that every element in $\mathcal{U}$ is “covered” by a set whose index is in $\mathcal{I}$, \textit{i.e.}, $\forall a_j \in \mathcal{U} \exists i \in \mathcal{I} : a_j \in S_i$. The objective of $Sc$ is to \textit{minimize} the number $|\mathcal{I}|$ of selected sets. We use the notation $\operatorname{opt}_{Sc}$ to denote the size (number of sets) in an optimal solution of an instance of $Sc$. On the inapproximability side, $Sc$ is NP-hard \cite{9} and, assuming $NP \not\subseteq \text{DTIME}(n^{\log \log n})$, $Sc$ does not admit a $(1 - \varepsilon) \ln n$-approximation for any constant $0 < \varepsilon < 1$ \cite{7}. On the algorithmic side, $Sc$ admits a $(1 + \ln n)$-approximation using a simple greedy algorithm \cite{10} that can be easily implemented to run in $O(\sum_{i=1}^{m} |S_i|)$ time \cite{3}.

Finally, in the context of proving NP-completeness, a “decision version” of an optimization problem by the standard method of using an additional parameter and formulating a decision \textit{(i.e., yes/no)} question on the value of the objective function with respect to this new parameter \textit{(e.g., see \cite{9})}. For reader’s convenience we explicitly write down these decision versions below.

**Problem 4 (decision version of metric anti-dimension or $Adim$).** \textit{Given $G$ and a positive integer $\zeta$, is there a subset of nodes $V'$ such that $\mu(D_{V \setminus V', -V'}) \geq \zeta$?}

**Problem 5 (decision version of $k \geq$-metric anti-dimension or $Adim_{\geq k}$).** \textit{Given $G$ and two positive integers $k$ and $\zeta$, is there a subset of nodes $V'$ such that $|V'| \leq \zeta$ and $\mu(D_{V \setminus V', -V'}) \geq k$?}

**Problem 6 (decision version of $k = $-metric antidimension or $Adim_{= k}$).** \textit{Given $G$ and two positive integers $k$ and $\zeta$, is there a subset of nodes $V'$ such that $|V'| \leq \zeta$ and $\mu(D_{V \setminus V', -V'}) = k$?}

It is also standard to state that an optimization problem is NP-complete to mean that the decision versions of the optimization problem is NP-complete, and therefore we will follow the same practice in this paper.

3. Our Results

In this section we provide precise statements of our results, leaving their proofs in Sections \cite{4,6}.
3.1. Polynomial Time Solvability of Adim and Adim≥k

**Theorem 1.**

(a) Both Adim and Adim≥k can be solved in \( O(n^4) \) time.

(b) Both Adim and Adim≥k can also be solved in \( O\left(\frac{n^4 \log n}{k}\right) \) time “with high probability” (i.e., with a probability of at least \( 1 - n^{-c} \) for some constant \( c > 0 \)).

**Remark 1.** The randomized algorithm in Theorem 1(b) runs faster than the deterministic algorithm in Theorem 1(a) provided \( k = \omega(\log n) \).

3.2. Computational Complexity of Adim=k

3.2.1. The Case of Arbitrary k

**Theorem 2.**

(a) Adim=k is NP-complete for any integer \( k \) 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.

(b) Assuming NP \( \not\subseteq \) DTIME \( (n^{\log \log n}) \), there exists a universal constant \( \delta > 0 \) such that Adim=k does not admit a \( (\frac{1}{\delta} \ln n) \)-approximation for any integer \( k \) 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.

(c) If \( k = n - c \) for some constant \( c \) then \( L^k_{opt} = c \) if a solution exists and Adim=k can be solved in polynomial time.

**Remark 2.**

(a) For \( k = 1 \), the inapproximability ratio in Theorem 2(b) is asymptotically optimal up to a constant factor because of the \( (1 + \ln(n-1)) \)-approximation of Adim=1 in Theorem 3(a).

(b) The result in Theorem 2(b) provides a much stronger inapproximability result compared to that in Theorem 2(a) at the expense of a slightly weaker complexity-theoretic assumption (i.e., NP \( \not\subseteq \) DTIME \( (n^{\log \log n}) \) vs. P \( \neq \) NP).
3.2.2. The Case of \( k = 1 \)

Note that even when \( k = 1 \) \( \text{ADIM}_k \) is \text{NP}-hard and even hard to approximate within a logarithmic factor due to Theorem 2. We show the following algorithmic results for \( \text{ADIM}_k \) when \( k = 1 \).

**Theorem 3.**

(a) \( \text{ADIM}_1 \) admits a \((1 + \ln(n - 1))\)-approximation in \( O(n^3) \) time.

(b) If \( G \) has at least one node of degree 1 then \( \mathcal{L}_{\text{opt}}^1 = 1 \) and thus \( \text{ADIM}_1 \) can be solved in \( O(n^3) \) time.

(c) If \( G \) does not contain a cycle of 4 edges then \( \mathcal{L}_{\text{opt}}^1 \leq 2 \) and thus \( \text{ADIM}_1 \) can be solved in \( O(n^3) \) time.

4. Proof of Theorem 1

(a) We first consider the claim for \( \text{ADIM}_k \geq 1 \). We begin by proving some structural properties of valid solutions for \( \text{ADIM}_k \geq 1 \).

**Proposition 1.** Consider two subsets of nodes \( \emptyset \subset V_1 \subset V_2 \subset V \). Let \( v_i, v_j \in V_2 \) be two nodes such that they do not belong to the same equivalence class in \( \Pi_{V \setminus V_1, V_2 \setminus V_1}^\equiv \). In this case \( v_i \) and \( v_j \) do not belong to the same equivalence class in \( \Pi_{V \setminus V_2, V_2 \setminus V_2}^\equiv \) also, and thus \( \Pi_{V \setminus V_1, V_2 \setminus V_1}^\equiv \otic \Pi_{V \setminus V_2, V_2 \setminus V_2}^\equiv \).

**Proof.** Since \( v_i \) and \( v_j \) are not in the same equivalence class in \( \Pi_{V \setminus V_1, V_2 \setminus V_1}^\equiv \), we have \( d_{v_i, V \setminus V_1} \neq d_{v_j, V \setminus V_1} \) which in turn implies (since \( V_1 \subset V_2 \)) \( d_{v_i, V_2 \setminus V_1} \neq d_{v_j, V_2 \setminus V_1} \) which implies \( v_i \) and \( v_j \) are not in the same equivalence class in \( \Pi_{V \setminus V_2, V_2 \setminus V_2}^\equiv \).  

Note that \( \Pi_{V \setminus V_2, V_2 \setminus V_1} \otic \Pi_{V \setminus V_1, V_2 \setminus V_1}^\equiv \) in Proposition 1 implies does not necessarily imply that \( \mu(D_{V \setminus V_2, V_2}) < \mu(D_{V \setminus V_1, V_1}) \). For example, for the example in Fig. 1 \( \Pi_{\{v_1, v_2\}, \{v_3, v_4, v_5\}} = \{\{v_1\}, \{v_2\}\} \otic \Pi_{\{v_1, v_2, v_3\}, \{v_4, v_5\}}^\equiv = \{\{v_1\}, \{v_2\}, \{v_3\}\} \) but \( \mu(D_{\{v_1, v_2, v_3\}, \{v_4, v_5\}}) = \mu(D_{\{v_1, \{v_2, v_3\}, \{v_4, v_5\}}) = 1 \). The following proposition gives some condition for this to happen.
Proposition 2. Consider two subsets of nodes \( \emptyset \subset V_1 \subset V_2 \subset V \), let \( S_1, \ldots, S_\ell \subseteq V \setminus V_1 \) be all equivalence classes (subsets of nodes) in \( \Pi_{V \setminus V_1} = V \setminus V_1 \setminus V_2 \), such that \( |S_1| = |S_2| = \cdots = |S_\ell| = \mu(D_{V \setminus V_1}) \), and assume that \( \emptyset \subset V_2 \cap S_j \subset S_j \) for some \( j \in \{1, \ldots, \ell\} \). In this case \( \mu(D_{V \setminus V_1}) < \mu(D_{V \setminus V_2}) \).

Proof. By Proposition 1, \( \Pi^\sim = V \setminus V_2 \prec_r \Pi^\sim = V \setminus V_1 \). If there exists a \( S_j \) such that \( \emptyset \subset V_2 \cap S_j \subset S_j \) then \( \Pi^\sim = V \setminus V_2 \) contains an equivalence class \( \emptyset \subset S_j' \subset S_j \). This implies \( \mu(D_{V \setminus V_2}) \leq |S_j'| < |S_j| = \mu(D_{V \setminus V_1}) \).

Based on the above structural properties, we design Algorithm I for \( \text{Adim} \geq k \) as shown below.

Algorithm I: \( O(n^4) \) time deterministic algorithm for \( \text{Adim} \geq k \).

1. Compute \( d_i \) for all \( i = 1, 2, \ldots, n \) in \( O(n^3) \) time using the Floyd-Warshall algorithm [4, p. 629]

2. \( \hat{L}^\geq_{\text{opt}} \leftarrow \infty \); \( \hat{V}^\geq_{\text{opt}} \leftarrow \emptyset \)

3. for each \( v_i \in V \) do (* we guess \( v_i \) to belong to \( V_{\text{opt}}^{\geq k} \) *)

   3.1 \( V' = \{v_i\} \); done \( \leftarrow \) FALSE

   3.2 while ( \( (V \setminus V') \neq \emptyset \) AND (NOT done) ) do

   3.2.1 compute \( \mu(D_{V \setminus V', - V'}) \)

   3.2.2 if \( \left( (\mu(D_{V \setminus V', - V'}) \geq k \right) \) and \( (|V'| < \hat{L}^\geq_{\text{opt}}) \) \)

   3.2.3 then \( \hat{L}^\geq_{\text{opt}} \leftarrow |V'| \); \( \hat{V}^\geq_{\text{opt}} \leftarrow V' \); done \( \leftarrow \) TRUE

   3.2.4 else let \( V_1, \ldots, V_\ell \) be all equivalence classes in \( \Pi^\sim = V \setminus V', - V' \),

   such that \( |V_1| = \cdots = |V_\ell| = \mu(D_{V \setminus V', - V'}) \)

   3.2.5 \( V' \leftarrow V' \cup (\bigcup_{i=1}^\ell V_i) \)

4. return \( \hat{L}^\geq_{\text{opt}} \) and \( \hat{V}^\geq_{\text{opt}} \) as our solution

Lemma 4 (Proof of correctness). Algorithm I returns an optimal solution
for $\text{ADIM}_{\geq k}$.

**Proof.** Assume that $V_{\text{opt}}^{\geq k} \neq \emptyset$ since otherwise our returned solution is correct. Fix any optimal solution (subset of nodes) $V_{\text{opt}}^{\geq k}$ of measure $\mu \left( D_{V \setminus V_{\text{opt}}^{\geq k}} \right) \geq k$ and select any arbitrary node $v_t \in V_{\text{opt}}^{\geq k}$. Consider the iteration of the for loop in Step 3 when $v_i$ is equal to $v_t$. We now analyze the run of this particular iteration.

Let $\{v_t\} = V_1 \subset V_2 \subset \cdots \subset V_k$ be the $k$ subsets of nodes that were assigned to $V'$ in successive iterations of the while loop in Step 3.2. We have the following cases to consider.

**Case 1:** $V_{\text{opt}}^{\geq k} = V_t$ for some $t \in \{1, 2, \ldots, k\}$. Our solution is a set $V_{\text{opt}}^{\geq k}$ such that $\mu \left( D_{V \setminus V_{\text{opt}}^{\geq k}} \right) \geq k$ and $\ell_{\text{opt}}^{\geq k} \subseteq \ell_{\text{opt}}^{\geq k}$.

**Case 2:** $V_{\text{opt}}^{\geq k} \neq V_t$ for any $t \in \{1, 2, \ldots, k\}$. Since $V_1 = \{v_t\} \subset V_{\text{opt}}^{\geq k}$ and $V_t \neq V_{\text{opt}}^{\geq k}$ for any $t \in \{1, 2, \ldots, k\}$, only one of the following cases is possible:

**Case 2.1:** there exists $r \in \{1, \ldots, k - 1\}$ such that $V_r \subset V_{\text{opt}}^{\geq k}$ but $V_{r+1} \not\subset V_{\text{opt}}^{\geq k}$. Let $V_{r,1}, V_{r,2}, \ldots, V_{r,p} \subseteq V \setminus V_r$ be all the $p > 0$ equivalence classes (subsets of nodes) in $\Pi_{V \setminus V_r}^r$ such that $|V_{r,1}| = |V_{r,2}| = \cdots = |V_{r,p}| = \mu \left( D_{V \setminus V_r} \right)$. Now we note the following:

- By Step 3.2.5, $V_{r+1} = V_r \cup V_{r,1} \cup V_{r,2} \cup \cdots \cup V_{r,p}$.
- Thus, $V_r \subset V_{\text{opt}}^{\geq k}$ and $V_{r+1} \not\subset V_{\text{opt}}^{\geq k}$ implies $V_{r,1} \cup V_{r,2} \cup \cdots \cup V_{r,p} \not\subset V_{\text{opt}}^{\geq k}$, and therefore there exists an index $1 \leq s \leq p$ such that $Z = V_{r,s} \setminus V_{\text{opt}}^{\geq k} \neq \emptyset$. Let $Z' = V_{r,s} \setminus Z$ ($Z'$ could be empty). For this case, for some $\emptyset \subset Z'' \subseteq Z$, $Z''$ is an equivalence class in $\Pi_{V \setminus (V_r \cup Z') \cup (V_r \cup Z'')}^r$ implying

$$\mu \left( D_{V \setminus (V_r \cup Z') \cup (V_r \cup Z'')} \right) \leq |Z''| \leq |Z| \quad (1)$$

Since $V_r \cup Z' \subseteq V_{\text{opt}}^{\geq k}$, we have

$$\Pi_{V \setminus V_{\text{opt}}^{\geq k}}^{\geq k} \preceq_r \Pi_{V \setminus (V_r \cup Z') \cup (V_r \cup Z'')}^{\geq k}$$

(in Proposition \[proposition:3.2.5\] set $V_2 = V_{\text{opt}}^{\geq k}$ and $V_1 = V_r \cup Z'$)
\[
\Rightarrow k \leq \mu(D_{V \setminus V_{opt} \geq k}) \leq \mu(D_{V \setminus (V \cup Z') \geq k}) \leq |Z| \leq |V_{r,s}| = \mu(D_{V \setminus V_r \geq k})
\]

Thus, \(\mu(D_{V \setminus V_r \geq k}) \geq k\) and \(|V_r| < |V_{opt}^k| = V_{opt}^k\), contradicting the optimality of \(V_{opt}^k\).

**Case 2.2:** \(V_k \subset V_{opt}^k\). If done was set to TRUE at the last iteration of the while loop, then \(\mu(D_{V \setminus V_{opt} \geq k}) \geq k\) and \(|V_k| < |V_{opt}^k| = V_{opt}^k\), contradicting the optimality of \(V_{opt}^k\). Thus, done must have remained FALSE after the last iteration of the while loop, implying \(\mu(D_{V \setminus V_{opt} \geq k}) \geq k\). Let \(V_{k,1}, V_{k,2}, \ldots, V_{k,p} \subseteq V \setminus V_k\) be all the \(p > 0\) equivalence classes (subsets of nodes) in \(\Pi_{V \setminus V_k \geq k}\) such that \(|V_{k,1}| = |V_{k,2}| = \cdots = |V_{k,p}| = \mu(D_{V \setminus V_{opt} \geq k})\). Since \(V_k \subset V_{opt}\), we have

\[
\Pi_{V_k \setminus V_{opt} \geq k} \prec \Pi_{V \setminus V_k \geq k}
\]

(in Proposition by setting \(V_2 = V_{opt}\) and \(V_1 = V_k\))

\[
\Rightarrow k \leq \mu(D_{V_{k,1} \setminus V_{opt} \geq k}) \leq \mu(D_{V \setminus V_{opt} \geq k}) \leq |Z| \leq |V_{k,p}| = \mu(D_{V \setminus V_{opt} \geq k})
\]

Thus, \(\mu(D_{V \setminus V_{opt} \geq k}) \geq k\) contradicting our assumption of \(\mu(D_{V \setminus V_{opt} \geq k}) < k\).

\[
\square
\]

**Lemma 5 (Proof of time complexity).** Algorithm I runs in \(O(n^4)\) time.

**Proof.** There are \(n\) choices for the for loop in Step 3. For each such choice, we analyze the execution of the while loop in Step 3.2. The running time in each iteration of the while loop is dominated by the time taken to compute \(\Pi_{V \setminus (V' \cup \{u_{i-1} \cup v_i\}) \setminus V' \cup \{u_{i-1} \cup v_i\}}\) from \(\Pi_{V \setminus V' \setminus V' \setminus V'}\). Suppose that \(\cup_{i=1}^n V_i = \{v_1, v_2, \ldots, v_p\}\). By Proposition

\[
\Pi_{V \setminus (V' \cup \{v_1, v_2, \ldots, v_{p-1}, v_p\}) \setminus V' \cup \{v_1, v_2, \ldots, v_{p-1}, v_p\}} \prec \Pi_{V \setminus (V' \cup \{v_1, v_2, \ldots, v_{p-1}, v_p\}) \setminus V' \cup \{v_1, v_2, \ldots, v_{p-1}, v_p\}} \prec \cdots
\]

15
Thus, it follows that the total time to execute all iterations of the while loop for a specific choice of \(v_i\) in Step 3 is of the order of \(n\) times the time taken to solve a problem of the following kind:

for a subset of nodes \(\emptyset \subset V_1 \subset V\), given \(\Pi^=_{V\setminus V_1, V'}\), and a node \(v_{i} \in V\setminus V_1\), compute \(\Pi^=_{V\setminus (V_1 \cup \{v_{i}\})}, -(V_1 \cup \{v_{i}\})}\).

Since \(\Pi^=_{V\setminus (V_1 \cup \{v_{i}\}), -V'}\) is a refinement of \(\Pi^=_{V\setminus V_1, -V'}\) by Proposition 1, we can use the following simple strategy. For every set \(S \in \Pi^=_{V\setminus V_1, -V'}\), we split \(S \setminus \{v_{j}\} = \{v_{i_1}, v_{i_2}, \ldots, v_{i_s}\}\) into two or more parts, if needed, by doing a bucket-sort (with \(n\) bins) in \(O(n |S|)\) time on the sequence of values \(\text{dist}_{v_{i_1}, v_{j}}, \ldots, \text{dist}_{v_{i_s}, v_{j}}\). The total time taken for all sets in \(\Pi^=_{V\setminus V_1, -V'}\) is thus \(\sum_{S \in \Pi^=_{V\setminus V_1, -V'}} O(n |S|) = O(n^2)\).

This completes the proof for \(\text{ADIM}_{\geq k}\). Now we consider the claim for \(\text{ADIM}\). Note that \(\text{ADIM}\) can be solved in \(O(n^5)\) time by solving \(\text{ADIM}_{\geq k}\) for \(k = n - 1, n - 2, \ldots, 1\) in this order and selecting the largest \(k\) as \(k_{\text{opt}}\) for which \(L_{\text{opt}}^2 < \infty\). However, we can modify the steps of Algorithm I directly to solve \(\text{ADIM}\) in \(O(n^4)\) time, as shown in Algorithm II.

---

**Algorithm II:** \(O(n^4)\) time deterministic algorithm for \(\text{ADIM}\)

(changes from Algorithm-I are shown enclosed in [])

1. Compute \(d_i\) for all \(i = 1, 2, \ldots, n\) in \(O(n^3)\) time using the Floyd-Warshall algorithm [4, p. 629]

2. \(V_{\text{opt}}^{\geq k} \leftarrow \emptyset\) \(\quad k_{\text{opt}} \leftarrow 0\)

3. for each \(v_i \in V\) do (* we guess \(v_i\) to belong to \(V_{\text{opt}}^{\geq k}\) *)

   3.1 \(V' = \{v_i\}\)

   3.2 while \((V \setminus V' \neq \emptyset)\) do

   3.2.1 compute \(\mu \left(\Pi^=_{V\setminus V', V'}\right)\)
The proof of correctness is very similar (and, in fact simpler due to elimination of some cases) to that of $\text{ADM} \geq k$.

(b) Our solution is the obvious randomization of Algorithm II (for $\text{ADM} \geq k$) or Algorithm-II (for $\text{ADM}$) as shown below.

Algorithm III (resp. Algorithm-IV): $O \left( n^4 \log \frac{n}{k} \right)$ time randomized algorithm for $\text{ADM} \geq k$ (resp. $\text{ADM}$)

1. Compute $d_i$ for all $i = 1, 2, \ldots, n$ in $O \left( n^3 \right)$ time using the Floyd-Warshall algorithm [4, p. 629]

2. $\hat{k}_{\text{opt}} \leftarrow \infty$; $V_{\text{opt}} \leftarrow \emptyset$ (for $\text{ADM} \geq k$) or $V_{\text{opt}} \leftarrow \emptyset$; $\hat{k}_{\text{opt}} \leftarrow 0$ (for $\text{ADM}$)

3. repeat $\lceil \frac{2n \ln n}{k} \rceil$ times

   3.1 select a node $v_i$ uniformly at random from the $n$ nodes

   3.2 execute Step 3.1 and Step 3.2 (and its sub-steps) of Algorithm I (for $\text{ADM} \geq k$) or execute Step 3.1 and Step 3.2 (and its sub-steps) of Algorithm II (for $\text{ADM}$)

4. return the best of all solutions found in Step 3
The success probability $p$ is given by

$$p = \Pr \left[ v_i \in V_{\text{opt}}^\geq k \text{ in at least one of the } \left\lceil \frac{2n \ln n}{k} \right\rceil \text{ iterations} \right] = 1 - \Pr \left[ v_i \notin V_{\text{opt}}^\geq k \text{ in each of the } \left\lceil \frac{2n \ln n}{k} \right\rceil \text{ iterations} \right] \geq 1 - \left( 1 - \frac{k}{n} \right) \left\lceil \frac{2n \ln n}{k} \right\rceil > 1 - \frac{1}{e^2 \ln n} = 1 - \frac{1}{n^2}$$

5. Proof of Theorem 2

The standard NP-complete minimum dominating set (MDS) problem for a graph is defined as follows [9]. Our input is a connected undirected unweighted graph $G = (V, E)$. A subset of nodes $V' \subseteq V$ is called a dominating set if and only if every node in $V \setminus V'$ is adjacent to some node in $V'$. The objective of MDS is to find a dominating set of nodes of minimum cardinality. Let $\nu(G)$ denote the cardinality of a minimum dominating set for a graph $G$. It is well-known that the MDS and Sc problems have precisely the same approximability via approximation-preserving reductions in both directions and, in particular, there exists a standard reduction from Sc to MDS as follows. Given an instance $U = \{a_1, a_2, \ldots, a_n\}$ and $S_1, S_2, \ldots, S_m \subseteq U$ of Sc, we create the following instance $G_1 = (V_1, E_1)$ of MDS. $V_1$ has an element node $v_{a_i}$ for every element $a_i \in U$ and a set node $v_{S_j}$ for every set $S_j$ with $j \in \{1, 2, \ldots, m\}$. There are two types of edges in $E_1$. Every set node $v_{S_j}$ has an edge to every other set node $v_{S_i}$ and the collection of these edges is called the set of clique edges. Moreover, a set node $v_{S_j}$ is connected to an element node $v_{a_i}$ if and only if $a_i \in S_j$ and the collection of these edges is called the set of membership edges. A standard straightforward argument shows that $I \subseteq \{1, 2, \ldots, m\}$ is a solution of Sc if and only if the collection of set nodes $\{ v_{S_i} \mid i \in I \}$ is a solution of MDS on $G_1$ and thus $\text{opt}_{\text{Sc}} = \nu(G_1)$.

(a) $\text{Adim}_k$ belongs to NP for any $k$ since given any solution $V'$ it is straightforward to verify if $|V'| \leq \zeta$ and $\mu(D_{V \setminus V', -V'}) = k$. Thus we need to show that it is also NP-hard.
For the purpose of our NP-hardness reduction, it would be more convenient to work with a restricted version of Sc known as the exact cover by 3-sets (X3C) problem. Here we have exactly \( n \) elements and exactly \( n \) sets where \( n \) is a multiple of three, every set contains exactly 3 elements and every element occurs in exactly 3 sets. Note that we need at least \( \frac{n}{3} \) sets to cover all the \( n \) elements. Letting \( \text{opt}_{X3C} \) to denote the number of sets in an optimal solution of X3C, it is well-known that problem of deciding whether \( \text{opt}_{X3C} = \frac{n}{3} \) is in fact NP-complete (e.g., see [9, p. 221]).

Let \( n_1 = \frac{-6k + \sqrt{36k^2 + 24(n-k)}}{4} \) be the real-valued solution of the quadratic equation \( n_1 \left( 2k + \frac{2n_1}{3} \right) + k = n \). Note that since \( k \leq n^\varepsilon \) for some constant \( \varepsilon < \frac{1}{2} \), we have \( n_1 = \Theta(\sqrt{n}) \), i.e., \( n \) and \( n_1 \) are "polynomially related".

Figure 2: Illustration of the NP-hardness reduction in Theorem 2(a). Only a part of the graph \( G \) is shown for visual clarity (for example, non-member edges are not shown).

We assume without loss of generality that \( n_1 \) is an even integer, and start with an instance of X3C of \( \frac{n_1}{2} \) elements and transform it to an instance graph
$G_1 = (V_1, E_1)$ having $n_1$ nodes of $MDS$ via the reduction outlined before. Since $\frac{n_1}{2}$ is polynomially related to $n$, such an instance of $X3C$ is $NP$-complete with respect to $n$ being the input size. We reduce $G_1$ to an instance $G = (V, E)$ of $ADIM = k$ in polynomial time as follows (see Fig. 2 for an illustration):

- We “clone” each element node $v_{a_j} \in V_1$ to get $2k + \frac{2n_1}{3}$ copies, i.e., every node $v_{a_j}$ is replaced by $2k + \frac{2n_1}{3}$ new nodes $v_{a_j, 1}, v_{a_j, 2}, \ldots, v_{a_j, 2k + \frac{2n_1}{3}}$. We refer to these nodes as clones of the element node $v_{a_j}$ (or, sometimes simply as element-clone nodes). There are precisely $n_1 \left( k + \frac{n_1}{3} \right)$ such nodes.

- We “clone” each set node $v_{S_j} \in V_1$ to get $2k + \frac{2n_1}{3}$ copies, i.e., every node $v_{S_j}$ is replaced by $2k + \frac{2n_1}{3}$ new nodes $v_{S_j, 1}, v_{S_j, 2}, \ldots, v_{S_j, 2k + \frac{2n_1}{3}}$. We refer to these nodes as clones of the set node $v_{S_j}$ (or, sometimes simply as set-clone nodes). There are precisely $n_1 \left( k + \frac{n_1}{3} \right)$ such nodes.

- We add $k$ new nodes $u_1, u_2, \ldots, u_k$. We refer to these nodes as clique nodes.

- We add an edge between every pair of clique nodes $u_i$ and $u_j$. We refer to these edges as clique edges. There are precisely $\binom{k}{2}$ such edges.

- We add an edge between every clique node and every non-clique node, i.e., we add every edge in the set

$$\left\{ \{u_i, v_{a_j, \ell}\} \mid 1 \leq i \leq k, 1 \leq j \leq \frac{n_1}{2}, 1 \leq \ell \leq 2k + \frac{2n_1}{3} \right\}$$

$$\bigcup \left\{ \{u_i, v_{S_j, \ell}\} \mid 1 \leq i \leq k, 1 \leq j \leq \frac{n_1}{2}, 1 \leq \ell \leq 2k + \frac{2n_1}{3} \right\}$$

We refer to these edges as the partition-fixing edges. There are precisely $k n_1 \left( k + \frac{n_1}{3} \right)$ such edges.

- We add an edge between every pair of distinct element-clone nodes $v_{a_j, \ell}$ and $v_{a_j, \ell'}$. We refer to these as the element-clone edges. There are precisely $\binom{2k + (2n_1)/3}{2}$ such edges.
• For every element $a_i$ and every set $S_j$ such that $a_i \notin S_j$, we add the following $(2k + \frac{2n_1}{3})^2$ edges:

$$\{v_{S_j,\ell}, v_{a_i,p}\} \text{ for } 1 \leq \ell, p \leq 2k + \frac{2n_1}{3}$$

We refer to these edges as the non-member edges corresponding to the element node $a_i$ and the set node $S_j$. There are precisely $\frac{3n_1}{2} (2k + \frac{2n_1}{3})^2$ such edges.

Note that $G$ has precisely $n_1 (2k + \frac{2n_1}{3}) + k = n$ nodes and thus our reduction is polynomial time in $n$. Since any clique node is adjacent to every other node in $G$, it follows that $\text{diam}(G) = 2$. We now show the validity of our reduction by showing that

\[ (*) \nu(G_1) = \frac{n_1}{3} \text{ if and only if } L_{\text{opt}}^3 \leq \frac{n_1}{3} \]

**Proof of** $\nu(G_1) = \frac{n_1}{3} \Rightarrow L_{\text{opt}}^3 < \frac{n_1}{3}$

Consider an optimal solution $V'_1 \subset \{v_{S_1}, v_{S_2}, \ldots, v_{S_{n_1}}\}$ of $\text{Mds}$ on $G_1$ with $\nu(G_1) = |V'_1| = \frac{n_1}{3}$. We now construct a solution $V' \subseteq V$ of $\text{Adim} = k$ on $G$ by setting $V' = \{v_{S_{j,1}} \mid v_{S_j} \in V'_1\}$. Note that $|V'| = |V'_1| = \frac{n_1}{3}$. We claim that $V'$ is a valid solution of $\text{Adim} = k$ by showing that

(a) $\{u_1, u_2, \ldots, u_k\} \in \Pi_{V \setminus V', \neg V'}$ and

(b) any other equivalence class in $\overline{\Pi_{V \setminus V', \neg V'}}$ has at least $k$ nodes.

To prove (a), consider a clique node $u_i$ and any other non-clique node. Then, the following cases apply:

• Suppose that the non-clique node is a element-clone node $v_{a_j, \ell} \in V \setminus V'$ for some $j$ and $\ell$. Since $V'_1$ is a solution of $\text{Mds}$ on $G_1$, there exists a set node $v_{S_p} \in V'_1$ such that $\{v_{S_p}, v_{a_j}\} \in E_1$ and consequently $\{v_{S_p,1}, v_{a_j,\ell}\} \notin E$. This implies that there exists a node $v_{S_{p,1}} \in V'$ such that $1 = \text{dist}_{u_i, v_{a_j, \ell}} = \text{dist}_{v_{S_{p,1}}, v_{a_j, \ell}}$, and therefore $v_{a_j, \ell}$ cannot be in the same equivalence class with $u_i$. 

21
Suppose that the non-clique node is a set-clone node \( v_{S,j,p}^* \in V \setminus V' \). Pick any set-clone node \( v_{S,j,p} \in V \setminus V' \). Then, \( 1 = \text{dist}_{u_i,v_{S,j,p}} \neq \text{dist}_{v_{S,j,p},v_{S,j,1}} \), and therefore \( v_{S,j,p} \) cannot be in the same equivalence class with \( u_i \).

To prove (b), note the following:

- Since \( \text{diam}(G) = 2 \), \( \text{dist}_{v_{S,i,p},v_{S,j,q}} = 2 \) for any two distinct set-clone nodes \( v_{S,i,p} \) and \( v_{S,j,q} \), and thus all the set nodes in \( V \setminus V' \) belong together in the same equivalence class in \( \Pi^{=\setminus V',-V'} \). There are at least \( n_1 (k + \frac{n_1}{3}) - \frac{n_1}{3} > k \) such nodes in \( V \setminus V' \). Thus, any equivalence class that contains these set-clone nodes cannot have less than \( k \) nodes.

- Consider now an equivalence class in \( \Pi^{=\setminus V',-V'} \), that contains a copy \( v_{a,i,j} \) of the element node \( v_{a,i} \) for some \( i \) and \( j \). Consider another copy \( v_{a,i,\ell} \) of the element node \( v_{a,i} \) for some \( \ell \neq j \). For any set node \( v_{S,p,1} \in V' \), if \( a_i \notin S_p \), then \( \text{dist}_{v_{S,p,1},v_{a,i,j}} = \text{dist}_{v_{S,p,1},v_{a,i,\ell}} = 1 \), whereas if \( a_i \in S_p \) then, since \( \text{diam}(G) = 2 \), it follows that \( \text{dist}_{v_{S,p,1},v_{a,i,j}} = \text{dist}_{v_{S,p,1},v_{a,i,\ell}} = 2 \). Thus, any equivalence class that contains at least one clone of an element node must contain all the \( 2k + \frac{2n_1}{3} > k \) clones of that element node and thus such an equivalence class cannot have a number of nodes that is less than \( k \).

**Proof of** \( L=^k \leq \frac{n_1}{3} \Rightarrow \nu(G_1) = \frac{n_1}{4} \)

Since we know that \( \nu(G_1) \) is always at least \( \frac{n_1}{3} \), it suffices to show that \( L=^k \leq \frac{n_1}{3} \Rightarrow \nu(G_1) \leq \frac{n_1}{4} \). Consider an optimal solution \( V_{opt}^{=k} \subset V \) with \( L_{opt}^{=k} = |V_{opt}^{=k}| = \frac{n_1}{3} \). Since \( V_{opt}^{=k} \) is a solution of \( \text{ADIM}=k \) on \( G \), there exists a subset of nodes, say \( \hat{V} \subset V \setminus V_{opt}^{=k} \), such that \( |\hat{V}| = k \) and \( \hat{V} \in \Pi^{=\setminus V_{opt}^{=k},-V_{opt}^{=k}} \).

**Proposition 3.** \( \hat{V} \) does not contain any set-clone or element-clone nodes and thus \( \hat{V} = \{u_1, u_2, \ldots, u_k\} \).

**Proof.** Suppose that \( \hat{V} \) contains at least one element-clone node \( v_{a,i,j} \) for some \( i \) and \( j \). But, \( V \setminus V_{opt}^{=k} \) contains at least \( 2k + \frac{2n_1}{3} - \frac{n_1}{3} - 1 > k \) other clones of
the element node $a_i$ and all these clones must belong together with $v_{a_i,j}$ in the same equivalence class. This implies $|\hat{V}| \geq 2k + \frac{2n_1}{3} - \frac{n_1}{4} + 2n_1^3 > k$, a contradiction.

Similarly, suppose that $\hat{V}$ contains at least one set-clone node $v_{S_i,j}$ for some $i$ and $j$. But, $V \setminus V_{opt}^{=k}$ contains at least $2k + \frac{2n_1}{3} - \frac{n_1}{4} - 1 > k$ other clones of the set node $S_i$ and all these clones must belong together with $v_{S_i,j}$ in the same equivalence class. This implies $|\hat{V}| \geq 2k + \frac{2n_1}{3} - \frac{n_1}{4} - 1 > k$, a contradiction. $\square$

**Proposition 4.** $V_{opt}^{=k}$ does not contain two or more clones of the same set node.

**Proof.** Suppose that $V_{opt}^{=k}$ contains two set-clone nodes $v_{S_j,p}$ and $v_{S_j,q}$ of the same set node $v_{S_j}$. But, $V \setminus V_{opt}^{\leq k}$ contains at least $2k + \frac{2n_1}{3} - \frac{n_1}{4} - 1 > k$ other clones of the element node $a_i$ and all these clones must belong together in the same equivalence class $S$. If we remove $v_{S_j,p}$ from $V_{opt}^{=k}$ then $v_{S_j,p}$ gets added to this equivalence class. Thus, such a removal produced another valid solution but with one node less than $L_{opt}$, contradicting the optimality of $L_{opt}^{=k}$. $\square$

**Proposition 5.** $V_{opt}^{=k}$ does not contain any element-clone node.

**Proof.** Suppose that $V_{opt}^{=k}$ contains at least one element-clone node and thus at most $\frac{n_1}{4} - 1$ set-clone nodes. Note that $V \setminus V_{opt}^{=k}$ contains at least $2k + \frac{2n_1}{3} - \frac{n_1}{4}$ clones of every element node $a_i$. Consider an element-clone node $v_{a_i,p} \in V \setminus V_{opt}^{=k}$ and a clique node $u_j$. Since $\hat{V} = \{u_1, u_2, \ldots, u_k\} \in \Pi^{=k}_{V_{opt}^{=k}}$, there must be a node in $V_{opt}^{=k}$ such that the distance of this node to $u_j$ is different from the distance to $v_{a_i,p}$. Such a node in $V_{opt}^{=k}$ cannot be an element-clone node, say $v_{u_{r,s}}$ since $dist_{v_{u_{r,s}}, v_{a_i,q}} = dist_{v_{a_i,q}, v_{a_i,p}} = 1$. Since there is an edge between every set-clone node and every clique node, such a node must be a set-clone node, say $v_{S_{r,s}}$ for some $r$ and $s$, such that $dist_{v_{a_i,p}, v_{S_{r,s}}} = 2$, i.e., $a_i \in S_r$. Since every set in X3c contains exactly 3 elements and $3 \times \left(\frac{n_1}{3} - 1\right) < n_1$, there must exist an element-clone node $v_{a_i,p}$ such that the distance of $v_{a_i,p}$ to any node in $V_{opt}^{=k}$ is exactly the same as the distance of $u_j$ to that node in $V_{opt}^{=k}$. This implies $v_{a_i,p} \in \hat{V}$, contradicting Proposition 3. $\square$
By Proposition 4 and Proposition 5, \( V_{\text{opt}} \) contains exactly one clone of a subset of set nodes. Without loss of generality, assume that \( V_{\text{opt}} = \{ v_{S,j} \mid j \in J, J \subset \{ 1, 2, \ldots, n \} \} \) and let \( V'_{1} = \{ v_{S,j} \mid v_{S,j} \in V_{\text{opt}} \} \). Note that \( |V'_{1}| = |V_{\text{opt}}| \). We are now ready to finish our proof by showing \( V'_{1} \) is indeed a valid solution of \( M_{\text{DS}} \) on \( G_{1} \). Suppose not, and let \( v_{a,i} \) be an element-node that is not adjacent to any node in \( V'_{1} \). For this case,

\[
\forall v_{S,j} \in V'_{1}: \{ v_{a,i}, v_{S,j} \} \notin E_{1} \Rightarrow \forall v_{S,j,1} \in V_{\text{opt}}^{\equiv k}: \{ v_{a,i,1}, v_{S,j,1} \} \in E
\]

\[
\Rightarrow \forall v_{S,j,1} \in V_{\text{opt}}^{\equiv k}: \text{dist}_{v_{a,i,1},v_{S,j,1}} = 1 \Rightarrow v_{a,i,1} \in \hat{V}
\]

which contradicts Proposition 3.

\textbf{(b)} The proof is similar to that of (a) but this time we start with a general version of \( \text{Sc} \) as opposed to the restricted \( \text{X3C} \) version, and show that the reduction is approximation-preserving in an appropriate sense. In the sequel, we use the standard notation \( \text{poly}(n) \) to denote a polynomial \( n^{c} \) of \( n \) (for some constant \( c > 0 \)). We recall the following details of the inapproximability reduction of Feige in [7]. Given an instance formula \( \phi \) of the standard Boolean satisfiability problem (\( \text{SAT} \)), Feige reduces \( \phi \) to an instance \( U, S_{1}, S_{2}, \ldots, S_{m} \) of \( \text{SC} \) (with \( m = \text{poly}(n) \)) in \( O(n^{\log \log n}) \) time such that the following properties are satisfied for any constant \( 0 < \varepsilon < 1 \):

- For some \( Q > 0 \), either \( \text{opt}_{\text{Sc}} = \frac{n}{Q} \) or \( \text{opt}_{\text{Sc}} > \left( \frac{n}{Q} \right) (1 - \varepsilon) \ln n \).
- The reduction satisfies the following completeness and soundness properties:

  \textbf{(completeness)} If \( \phi \) is satisfiable then \( \text{opt}_{\text{Sc}} = \frac{n}{Q} \).

  \textbf{(soundness)} If \( \phi \) is not satisfiable then \( \text{opt}_{\text{Sc}} > \left( \frac{n}{Q} \right) (1 - \varepsilon) \ln n \).

Since \( m = \text{poly}(n) \), by adding duplicate copies of a set, if necessary, we can ensure that \( m = n^{c} - n \) for some constant \( c \geq 1 \). Our reduction from \( \text{Sc} \) to \( \text{MDS} \) is same as in (a) except that some details are different, which we show here.
• We start with an instance of Sc as given by Feige in [7] with \(n_1\) elements
and \(m = (n_1)^c - n_1\) sets, where
\[
n_1 = \left(\frac{-k + \sqrt{k^2 + 2(n-k)}}{2}\right)^{1/c}
\]
is a real-valued solution of the equation
\[
(n_1)^{2c} + k(n_1)^c - \frac{n-k}{2} = 0.
\]
Note that since \(k \leq n^c\) for some constant \(\varepsilon < \frac{1}{2}\), we have
\(n_1 = \Theta\left(n^{1/(2c)}\right)\), i.e., \(n\) and \(n_1\) are polynomially related.

• We make \(2(n_1)^c + 2k\) copies of each element node and each set node as opposed
to \(2k + \frac{2n_1}{3}\) copies that we made in the proof of (a). Note that \(G\) has again
precisely \((n_1)^c(2k + 2(n_1)^c) + k = n\) nodes.

• Let \(\delta > 0\) be the constant given by \(\delta = \ln n (1 - \varepsilon) \ln n_1\). Our claim (⋆) in the proof
of (a) is now modified to

\[
\text{(completeness)} \quad \text{if } \nu(G_1) = \frac{n_1}{Q} \text{ then } L^=_{\text{opt}} \leq \frac{n_1}{Q}
\]

\[
\text{(soundness)} \quad \text{if } \nu(G_1) > \left(\frac{n_1}{Q}\right)(1 - \varepsilon) \ln n_1
\]
\[
\text{then } L^=_{\text{opt}} > \left(\frac{n_1}{Q}\right)(1 - \varepsilon) \ln n_1 = \left(\frac{n_1}{Q}\right)\frac{1}{\delta} \ln n
\]

• Our proof of the completeness claim follows the “Proof of \(\nu(G_1) = \frac{n_1}{3} \Rightarrow L^=_{\text{opt}} \leq \frac{n_1}{3}\)” in the proof of (a) with the obvious replacement of \(\frac{n_1}{3}\) by \(\frac{n_1}{Q}\).

• Note that our soundness claim is equivalent to its contra-positive

\[
\text{if } L^=_{\text{opt}} \leq \left(\frac{n_1}{Q}\right)(1 - \varepsilon) \ln n_1 \text{ then } \nu(G_1) \leq \left(\frac{n_1}{Q}\right)(1 - \varepsilon) \ln n_1
\]

and the proof of this contra-positive follows the “Proof of \(L^=_{\text{opt}} \leq \frac{n_1}{3} \Rightarrow \nu(G_1) = \frac{n_1}{3}\)” in the proof of (a). In the proof, the quantity \(2k + \frac{2n_1}{3}\) corresponding to the number of copies for each set and element node needs to be replaced by \(2(n_1)^c + 2k\); note that \((2(n_1)^c + 2k) - n_1 \gg k\).

(c) Since \(k = n - c\) for some constant \(c\), \(\Pi^=_{V \setminus V^=_{\text{opt}}} \cdot V^=_{\text{opt}}\) contains a single equivalence class \(V' \subset V\) such that \(|V'| = k\). Thus, we can employ the straightforward
exhaustive method of selecting every possible subset \(V'\) of \(k\) nodes to be in
\(\Pi^=_{V \setminus V', \cdot V^=_{\text{opt}}}\) and checking if the chosen subset of nodes provide a valid solution.

There are \(\binom{n}{k} < n^c\) such possible subsets and therefore the asymptotic running
time is \(O\left(n^c + n^3\right)\) which is polynomial in \(n\). Note that for this case \(L^=_{\text{opt}} = c\
if a solution exists.
6. Proof of Theorem 3

(a) Note that $L_{opt} = 1 ≤ n - 1$ and thus $V_{opt} ≠ ∅$. Our algorithm, shown as Algorithm V, uses the greedy logarithmic approximation of Johnson [10] for $Sc$ that selects, at each successive step, a set that contains the maximum number of elements that are still not covered.

Algorithm V: $O(n^3)$-time $(1 + \ln(n - 1))$-approximation algorithm for $Adim=1$.

1. Compute $d_i$ for all $i = 1, 2, \ldots, n$ in $O(n^3)$ time using Floyd-Warshall algorithm.

2. $\hat{L}_{opt} ← \infty ; \hat{V}_{opt} ← ∅$

3. for each node $v_i \in V$ do
   (* we guess the set $\{v_i\}$ to belong to $\Pi_{V \setminus \hat{V}_{opt}, \hat{V}_{opt}}$ *)
   3.1 create the following instance of $Sc$ containing $n - 1$ elements and $n - 1$ sets:
      $U = \{a_{v_j} \mid v_j \in V \setminus \{v_i\} \}$,
      $S_{v_j} = \{a_{v_j} \} \cup \{a_{v_j} \mid \text{dist}_{v_i,v_j} \neq \text{dist}_{v_i,v_j} \}$ for $j \in \{1, 2, \ldots, n\} \setminus \{i\}$
   3.2 if $\cup_{j \in \{1, 2, \ldots, n\} \setminus \{i\}} S_{v_j} = U$ then
      3.2.1 run the greedy approximation algorithm [10] for this instance of $Sc$
      giving a solution $I \subseteq \{1, 2, \ldots, n\} \setminus \{i\}$
      3.2.2 $V' = \{v_j \mid j \in I\}$
      3.2.3 if $(|V'| < \hat{L}_{opt})$ then $\hat{L}_{opt} ← |V'| ; \hat{V}_{opt} ← V'$

4. return $\hat{L}_{opt}$ and $\hat{V}_{opt}$ as our solution

Lemma 6 (Proof of correctness). Algorithm V returns a valid solution for $Adim=1$.

Proof. Suppose that our algorithm returns an invalid solution in the iteration of the for loop in Step 3 when $v_i$ is equal to $v_\ell$ for some $v_\ell \in V$. We claim that this cannot be the case since $\{v_\ell\} \in \Pi_{V \setminus V', \hat{V}_{opt}}$. Indeed, since $I$ is a valid
solution of the SC instance, for every \( j \notin \{ \ell \} \cup \mathcal{I} \), the following holds:

\[
\exists t \in \mathcal{I} : a_{vJ} \in S_{vI} \Rightarrow \exists v_t \in V' : \text{dist}_{vJ,v_t} \neq \text{dist}_{vJ,v_t}
\]

and thus \( v_t \) cannot be together with any other node in any equivalence class in \( \Pi_{V \setminus V'} \). □

Lemma 7 (Proof of approximation bound). Algorithm V solves \( \text{ADIM}=1 \)
with an approximation ratio of \( 1 + \ln(n - 1) \).

Proof. Fix any optimal solution \( V^=_{\text{opt}} \). Since \( \mu \left( D_{V \setminus V^=_{\text{opt}}}, V^=_{\text{opt}} \right) = 1 \), \( \{ v_t \} \in \Pi_{V \setminus V^=_{\text{opt}}} \) for some \( v_t \in V \). Consider the iteration of the for loop in Step 3 when \( v_i \) is equal to \( v_t \). We now analyze the run of this particular iteration, and claim that the set-cover instance created during this iteration satisfies \( \text{opt}_{\text{SC}} \leq |V^=_{\text{opt}}| = L^=_{\text{opt}} \). To see this, construct the following solution of the set-cover instance from \( V_{\text{opt}} \) containing exactly \( L^=_{\text{opt}} \) sets:
\[
v_i \in V^=_{\text{opt}} \equiv i \in \mathcal{I}
\]

To see that this is indeed a valid solution of the set-cover instance, consider any \( a_{vJ} \in \mathcal{U} = \{ a_{v1}, a_{v2}, \ldots, a_{vn} \} \setminus \{ a_{vJ} \} \). Then, the following cases apply showing that \( a_{vJ} \) belongs to some set selected in our solution of SC:

- if \( j \in \mathcal{I} \) then \( a_{vJ} \in S_{vJ} \) and \( S_{vJ} \) is a selected set in the solution.
- if \( j \notin \mathcal{I} \) then \( vJ \in V \setminus V_{\text{opt}} \Rightarrow \exists v_t \in V_{\text{opt}} : \text{dist}_{vJ,v_t} \neq \text{dist}_{vJ,v_t} \Rightarrow \exists t \in \mathcal{I} : a_{vJ} \in S_{vJ} \).

Using the approximation bound of the algorithm of [10] it now follows that the quality of our solution \( \hat{L}^=_{\text{opt}} \) satisfies
\[
\hat{L}^=_{\text{opt}} = |V^=_{\text{opt}}| = |\mathcal{I}| < (1 + \ln(n - 1)) \text{opt}_{\text{SC}} \leq (1 + \ln(n - 1))L^=_{\text{opt}}
\]

□

Lemma 8 (Proof of time complexity). Algorithm V runs in \( O(n^3) \) time.
Proof. There are a total of \( n \) instances of set cover that we need to build in Step 3.1 and solve by the greedy heuristic in Step 3.2.1. Building the set-cover instance can be done in \( O(n^2) \) time by comparing \( \text{dist}_{v_i,v_j} \) for all appropriate pairs of nodes \( v_i \) and \( v_j \). Since the set-cover instance in Step 3.1 has \( n - 1 \) sets each having no more than \( n - 1 \) elements, each implementation of the greedy heuristic in Step 3.2.1 takes \( O(n^2) \) time.

(b) Let \( v_i \) be the node of degree 1. Let \( v_{\ell} \) be the unique node adjacent to \( v_i \) (i.e., \( \{v_i,v_{\ell}\} \in E \)). Consider the following solution of \( \text{Adim}=1 \): \( V' = \{v_i\} \). We claim that is a valid solution of \( \text{Adim}=1 \) by showing that \( \{v_{\ell}\} \in \Pi_{V' \setminus V, -V} \). Consider any node \( v_j \in V \setminus \{v_i,v_{\ell}\} \). Then, \( 1 = \text{dist}_{v_{\ell},v_i} \neq \text{dist}_{v_j,v_i} \).

(c) Since \( G \) does not contain a 4-cycle, \( \text{diam}(G) \geq 2 \). Thus, there exists two nodes \( v_i, v_j \in V \) such that \( \text{dist}_{v_i,v_j} = 2 \). Let \( v_{\ell} \) be a node at a distance of 1 from both \( v_i \) and \( v_j \) on a shortest path between \( v_i \) and \( v_j \) (see Fig. 3). Consider the following solution of \( \text{Adim}=1 \): \( V' = \{v_i,v_j\} \). Note that \( v_{\ell} \in V \setminus V' \). We claim that this is a valid solution of \( \text{Adim}=1 \) by showing that \( \{v_{\ell}\} \in \Pi_{V \setminus V', -V'} \) (i.e., no node \( v_p \in V \setminus \{v_i,v_j,v_{\ell}\} \) can belong together with \( v_{\ell} \) in the same equivalence class of \( \Pi_{V \setminus V', -V'} \)) in the following manner:
\begin{itemize}
  \item If \( v_p \in \text{Nbr}(v_i) \setminus \{v_l\} \) then \( \text{dist}_{v_l,v_i} = 1 \) but \( \text{dist}_{v_p,v_j} \neq 1 \) since \( G \) has no 4-cycle (see the edges marked × in Fig. 3).
  
  \item If \( v_p \in \text{Nbr}(v_j) \setminus \{v_t\} \) then \( \text{dist}_{v_t,v_i} = 1 \) but \( \text{dist}_{v_p,v_i} \neq 1 \) since \( G \) has no 4-cycle (see the edges marked × in Fig. 3).
  
  \item If \( v_p \in \text{Nbr}(v_l) \setminus \{v_i,v_j\} \) then \( v_p \) cannot be adjacent to both \( v_i \) and \( v_j \) since \( G \) does not contain a 4-cycle. This implies that \( \text{dist}_{v_l,v_i} = \text{dist}_{v_l,v_j} = 1 \) but at least one of \( \text{dist}_{v_p,v_i} \) and \( \text{dist}_{v_p,v_j} \) is not equal to 1.
  
  \item If \( v_p \) is any node not covered by the above cases, we have \( \text{dist}_{v_p,v_i} > 1 \) but \( \text{dist}_{v_l,v_i} = 1 \).
\end{itemize}

7. Related Works: Other Privacy Concepts and Measures

There is a rich literature on theoretical investigations of privacy measures and privacy preserving computational models in several other application areas such as multi-party communications, distributed computing and game-theoretic settings (e.g., see [2, 11, 22, 8, 3]). However, none of these settings apply directly to our application scenario of active attack model for social networks. The differential privacy model, introduced by Dwork [5] in the context of privacy preservation in statistical databases against malicious database queries, works by computing the correct answer to a query and adding a noise drawn from a specific distribution, and is quite different from the anonymization approach studied in this paper.

8. Concluding Remarks

Prior to our work, known results for the three problems considered in this paper only included some heuristic algorithms with no provable guarantee on performances such as in [18], or algorithms for very special cases. In fact, it was not even known if any version of these computational problems is \( \text{NP} \)-hard. Our work provides the first non-trivial computational complexity results for effective solution of these problems. Theorem 1 shows that both \( \text{ADIM} \) and \( \text{ADIM}_{\geq k} \)
are provably computationally easier problems than $\text{ADIM}_k$. In contrast, Theorem 2(a)–(b) and Theorem 3 show that $\text{ADIM}_k$ is in general computationally hard but admits approximations or exact solution for specific choices of $k$ or graph topology. We believe that our results will stimulate further research on quantifying and computing privacy measures for networks. In particular, our results raise the following interesting research questions:

- We have only provided a logarithmic approximation algorithm for $\text{ADIM}_{=1}$. Is it possible to design a non-trivial approximation algorithm for $\text{ADIM}_{=k}$ for $k > 1$?

- We have provided a logarithmic inapproximability result for $\text{ADIM}_{=k}$ for every $k$ roughly up to $\sqrt{n}$. Can this approximability result be further improved when $k$ is not a constant?

References


