Analysis of Privacy Measures for Multi-Agent and Networked Systems

BY

VENKATAKUMAR SRINIVASAN
B.E., Bharathiyar University, 2002

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2017

Chicago, Illinois

Defense Committee:

Bhaskar DasGupta, Chair and Advisor
Robert H. Sloan
Jon A. Solworth
V.N. Venkatakrishnan
Ismael Gonzalez Yero, Universidad de Cádiz, Spain

I dedicate this thesis to my parents and my family.

ACKNOWLEDGMENT

I would like to express my profund gratitude to my advisor Prof. Bhaskar DasGupta for his guidance and support. I would also like to thank my committee members for their valuble suggestions and support.

VS

PREFACE

This thesis is based on the following publications

- Marco Comi, Bhaskar DasGupta, Michael Schapira and Venkatakumar Srinivasan, On Communication Protocols that Compute Almost Privately, Theoretical Computer Science, 457, 45-58, 2012.
- Marco Comi, Bhaskar DasGupta, Michael Schapira, Venkatakumar Srinivasan, On Communication Protocols that Compute Almost Privately, 4th Symposium on Algorithmic Game Theory, G. Persiano (Ed.), LNCS 6982, Springer-Verlag, 44-56, 2011
- Bhaskar DasGupta and Venkatkumar Srinivasan, A review of some approximate privacy measures of multi-agent communication protocols, in Frontiers of Intelligent Control and Information Processing, Derong Liu, Cesare Alippi, Dongbin Zhao, and Huaguang Zhang (editors), Chapter 10, 267-283, World Scientific Publishing, 2014.
- Tanima Chatterjee, Bhaskar DasGupta, Nasim Mobasheri, Venkatakumar Srinivasan and Ismael G. Yero, On the Computational Complexities of Three Privacy Measures for Large Networks Under Active Attack, arXiv:1510.08779 [cs.CC]

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF ABBREVIATIONS

PAR	Privacy Approximation Ratio
BSP	Binary Space Partition
X3C	Exact Cover by 3-Sets
MDS	Minimum Dominating Set
SC	Set Cover
ADMIN	Metric anti-dimension
SAT	Boolean Satisfiability Problem

SUMMARY

Privacy preserving computation is an important research area and it has become a natural question to ask about the class of functions that are privately computable. Another active research area 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 thesis we present techniques for analysing and quantifying privacy measures in multi-agent and networked systems.

In the first part of the thesis, we provide the results of our investigation into the approximate privacy model introduced by Feigenbaum, Jaggard and Schapira (1). Our results indicate that for a large class of functions which we call as the *tiling functions*, a protocol exists that provides a constant *average* privacy approximation ratio and such a protocol involves a number of communication rounds linear in the number of monochromatic regions of the function; however, we show that such a good privacy approximation ratio for tiling functions do not exist in the *worst case*. We also discuss extension of the basic setup to more than two players as well as to non-tiling functions, and provide calculations of average and worst case privacy approximation ratios of the bisection protocol for several new non-tiling functions.

In the second part of the thesis, we formalize three optimization problems concerning a privacy measure used for quantifying privacy of users in large networks and provide non-trivial theoretical computational complexity results for solving these optimization problems. Our results show the first two optimization problems can be computed efficiently, whereas the third

problem is provably hard to compute 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.

CHAPTER 1

INTRODUCTION

With the arrival of modern internet era, large public data stores 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. There is widespread usage of sensitive data in networked environments, as evidenced by distributed computing applications, game-theoretic settings (e.g., auctions) and more. On the other hand, malicious entities may violate the privacy of the users of such a network by analyzing the network or intercepting communication between multiple agents and deliberately using such privacy violations for deleterious purposes. Hence protecting the privacy of data in such systems is of much practical importance. There is lot of research focussed on privacy preserving computations and an important task in that area is to define privacy measures to quantify privacy (and loss of privacy). In this thesis we present the results of our investigations into two such privacy measures

---

[1]The contents of this chapter are taken from (2; 3), arXiv:1510.08779 [cs.CC]

[2]Reprinted from Theoretical Computer Science, Vol 457, M. Comi *et al.* ,On communication protocols that compute almost privately, 45-58, 2012, with permission from Elsevier

[3]Algorithmic Game Theory: 4th International Symposium, SAGT 2011, Amalfi, Italy, October 17-19, 2011. Proceedings, On Communication Protocols That Compute Almost Privately, 2011, M. Comi *et al.* (© Springer-Verlag Berlin Heidelberg 2011) With permission of Springer

- privacy measures for quantifying privacy of communication protocol between multiple agents performing a computation
- privacy measures for quantifying privacy of user infomation in large social networks

1.1 Privacy Preserving Communication Protocols

Consider the following interaction between two parties, Alice and Bob. Each of the two parties, Alice and Bob, holds a *private* input, x_{bob} and y_{alice} respectively, not known to the other party. The two parties aim to compute a function f of the two private inputs(Figure 1). Alice and Bob alternately query each other to make available a *small* amount of information about their private inputs, e.g., an answer to a range query on their private inputs or a few bits of their private inputs. This process ends when each of them has seen enough information to be able to compute the value of $f(x_{bob}, y_{alice})$.

This raises the following question:

> *Can we design a communication protocol whose execution reveals, to both Alice and Bob, as well as to any eavesdropper, as little information as possible about the others private input beyond what is necessary to compute the function value?*

Note that there are two conflicting constraints: Alice and Bob need to communicate sufficient information for computing the function value, but would prefer not to communicate too much information about their private inputs. This setting can be generalized in an obvious manner to $d > 1$ parties $party_1, party2, \ldots, party_d$ computing a d-ary f by querying the parties

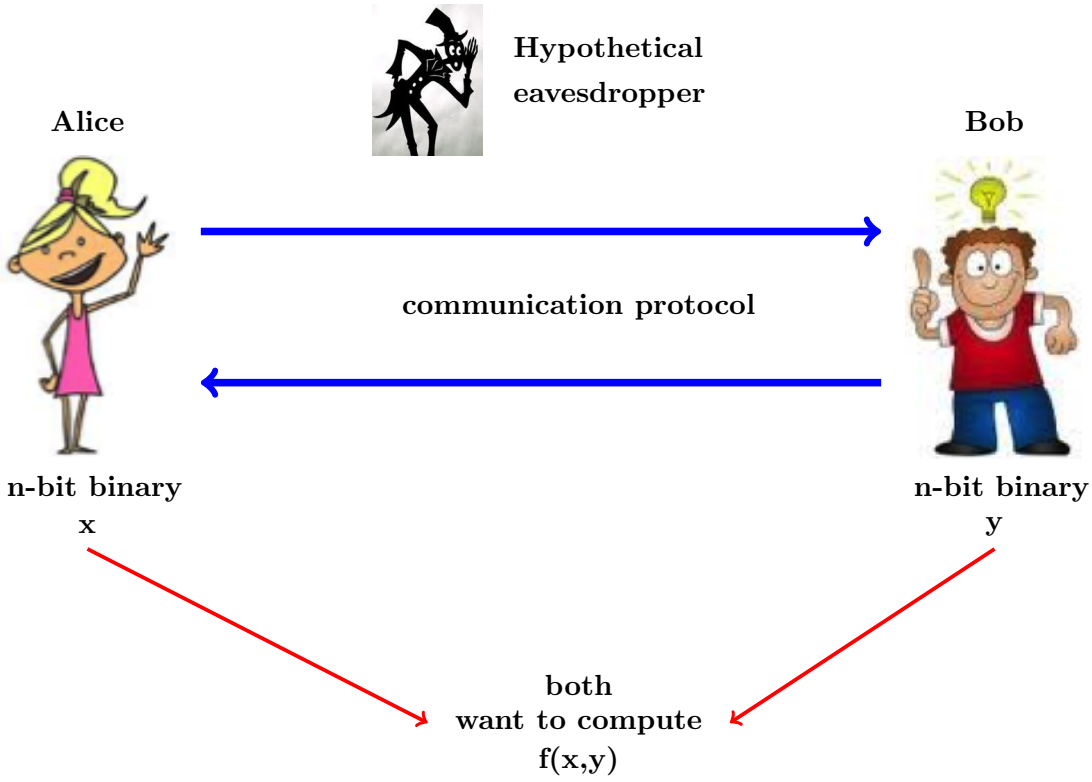


Figure 1. Two Agents

in round-robin order, allowing each party to broadcast information about its private input (via a public communication channel).

Over the years computer scientists have explored many *quantifications* of privacy in computation. Much of this research focused on designing *perfectly* privacy-preserving protocols, *i.e.* , protocols whose execution reveals *no* information about the parties private inputs beyond that implied by the outcome of the computation. Unfortunately, perfect privacy is often either *impossible*, or *infeasibly costly* to achieve. To overcome this, researchers have also investigated

various notions of *approximate privacy* (1; 4). The first part of the thesis explains the notion of *approximate privacy* and lists the results of our research in privacy preserving protocols for evaluating *tiling* and *non-tiling* functions.

1.2 Privacy of Social Networks

Social networks have certainly 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 certainly benefit and improve their performances by a fair study of the social network data. But, such benefits are definitely *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 (5). The objective is to make sure that *no* database record can be identified again with a probability greater than $1/k$.

Crucial to modelling a social network anonymization process are of course the adversary'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 (6), the node neighborhood (7), *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 (8) for more information). Some examples of passive attacks and the corresponding privacy-preserving methods for social networks can be found in references (6; 7; 9).

In contrast, Backstrom *et al.* introduced the concept of the so-called *active* attacks in (10). Such attacks are mainly based on creating and inserting some nodes (the "attacker nodes") controlled by the adversary into the network. These attacker nodes could be newly created accounts with pseudonymous or forged identities (commonly called Sybil nodes), or existing legitimate individuals in the network that are in the adversary's proximity. The goal is then 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 (10) 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, (11)). For other related publications on privacy-preserving methods in social networks, see (7; 12; 13).

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. To this effect, very recently a novel privacy measure for social networks was introduced in (14). The privacy measure proposed there was called the (k, ℓ)-*anonymity*, where k is a number indicating a privacy threshold and ℓ is the *maximum* number of attacker nodes that can be inserted into the network; ℓ may be estimated through some statistical methods[1]. Trujillo-Rasua and Yero in (14) showed that graphs satisfying (k, ℓ)-anonymity can prevent adversaries who control at most ℓ nodes in the network from re-identifying individuals with probability higher than $1/k$. 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 $\mathrm{dist}_{u,v}$ be the length (number of edges) of a shortest path between two nodes $u, v \in V$. For an ordered sequence $S = u_1, \ldots, u_t$ of nodes of G and a node $v \in V$, the vector $\mathbf{d}_{v,-S} = (\mathrm{dist}_{v,u_1}, \ldots, \mathrm{dist}_{v,u_t})$ is called the *metric representation* of v with respect to S. Based on the above definition, a 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 $v_1, \ldots, v_{k-1} \in V \setminus S$

---

[1]Note that other different privacy notions with the *same* name also exists, *e.g.* , Feder and Nabar in (15) investigated (k, ℓ)-anonymity where ℓ represented the number of common neighbors of two nodes.

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 $\mathrm{adim}_k(G)$, is then 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, ℓ)-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, ℓ)-anonymity privacy measure is then 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 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 ℓ then the adversary must use at least ℓ attacker nodes to get the probability of privacy violation down to $1/k$.

The second part of the thesis formalizes three computational problems related to measuring (k, ℓ)-anonymity of graphs, presents algorithms and non-trivial computational complexity results for these problems

CHAPTER 2

PRIVACY OF MULTI AGENT COMMUNCATION PROTOCOLS

Privately computable functions has been studied in the literature in the past based on combinatorial characterization (*e.g.* , see (16)), communication-complexity analysis (*e.g.* , see (17)) or information-theoretic analysis (*e.g.* , see (18)). Unfortunately, the results of such investigations have showed that many interesting classes of functions either do not have a "perfect" privacy-preserving protocols or such protocols require impractically large communication rounds. Thus, it behooves to formalize an appropriate notion of approximate privacy and study its properties.

Recently, in (1) Feigenbaum, Jaggard and Schapira described a notion of approximate privacy for protocols computing a function of two variables based on a geometric and combinatorial interpretation of the protocol. We study its application to *tiling functions* and *bisection protocol* designed based on Bsp.

---

[1]The contents of this chapter are taken from (2; 3)

[2]Reprinted from Theoretical Computer Science, Vol 457, M. Comi *et al.* ,On communication protocols that compute almost privately, 45-58, 2012, with permission from Elsevier

[3]Algorithmic Game Theory: 4th International Symposium, SAGT 2011, Amalfi, Italy, October 17-19, 2011. Proceedings, On Communication Protocols That Compute Almost Privately, 2011, M. Comi *et al.* (© Springer-Verlag Berlin Heidelberg 2011) With permission of Springer

2.1 Basic Definitions for the Two-party Model

We have two parties, each one of them holding a private information represented by a k-bit string that represents a value in $\{0, 1, \ldots, 2^k - 1\}$. In each communication round, one of the parties alternately sends out a bit that is computed as a function of that party's input and communication history. The last message sent in a protocol P is assumed to contain the value of the function, and therefore may require a larger number of bits. The final outcome of the protocol is denoted by the function s.

Denoting the domain of inputs and outputs by Σ_{in} and Σ_{out}, respectively, any function $f : \Sigma_{\text{in}} \times \Sigma_{\text{in}} \mapsto \Sigma_{\text{out}}$ can be visualized as $|\Sigma_{\text{in}}| \times |\Sigma_{\text{in}}|$ matrix with entries from Σ_{out} in which the first dimension represents the possible values of player 1, ordered by some permutation Π_1, while the second dimension represents the possible values of player 2, ordered by some permutation Π_2; each entry contains the value of f associated with a particular set of inputs from the 2 players. This matrix will be denoted by $A_{\Pi_1,\Pi_2}(f)$.(Figure 2)

Definition 1 *(1) Let $A = A_{\Pi_1,\Pi_2}(f)$ be the matrix as described above.*

Region: *a region of A is any subset of entries in A (not necessarily a submatrix).*

Partition: *a partition of A is a collection of disjoint regions in A whose union equals to A.*

Monochromaticity: *a region R of A in called monochromatic if all entries in R are of the same value. A monochromatic partition of A is a partition all of whose regions are monochromatic.*

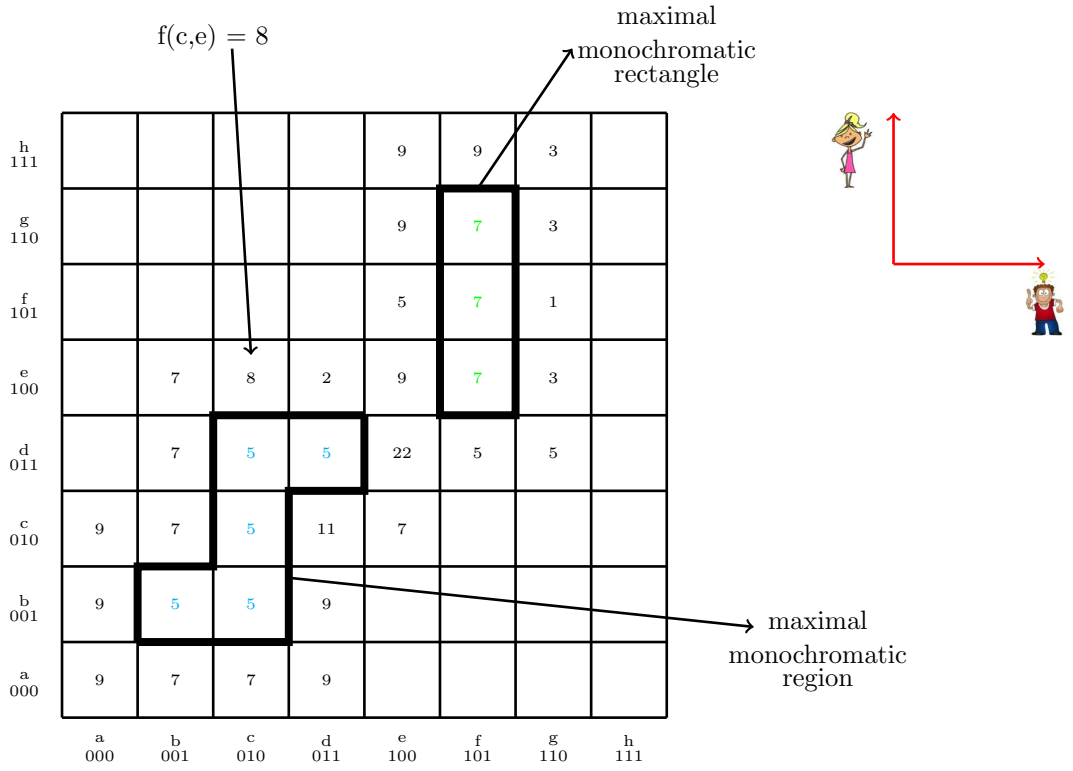


Figure 2. Function Matrix

Rectangle: *a rectangle in A is a submatrix of A.*

Tiling: *a tiling of A is a partition of A into rectangles.*

Refinements: *a tiling T_1 of A is said to be a refinement of another tiling T_2 of A if every rectangle in T_1 is contained in some rectangle in T_2.*

Perfect privacy: *P achieves perfect privacy if, for every two sets of inputs (x_1, x_2) and (x'_1, x'_2) such that $f(x_1, x_2) = f(x'_1, x'_2)$, it holds that $s(x_1,\ x_2) = s(x'_1,\ x'_2)$.*

Ideal monochromatic partitions: *a monochromatic region of A is said to be a maximal monochromatic region if no monochromatic region in A properly contains it. The ideal monochromatic partition of A is made up of the maximal monochromatic regions.*

Perfectly privacy-preserving protocol: *a communication protocol P for f is perfectly privacy-preserving if the monochromatic tiling induced by P is the ideal monochromatic partition of $A(f)$.*

Worst case* Par *of a protocol P: *let $R^P(x_1, x_2)$ be the monochromatic rectangle induced by P for $(x_1, x_2) \in \{0,1\}^k \times \{0,1\}^k$ and $R^I(x_1, x_2)$ be the monochromatic region containing $A(x_1, y_1)$ in the ideal monochromatic partition of A. Then P has a worst-case privacy-approximation-ratio (*Par*) of α_{worst} if $\alpha_{\text{worst}} = \max_{(x_1,x_2)} \left[\frac{|R^I(x_1,x_2)|}{|R^P(x_1,x_2)|} \right]$.*

Average case* Par *of P: *let $\mathcal{D}$ be a probability distribution over the space of inputs. The average case privacy-approximation-ratio (*Par*) of a communication protocol P under distribution $\mathcal{D}$ for function f is $\alpha_{\mathcal{D}} = E_{\mathcal{D}} \left[\frac{|R^I(x_1,x_2)|}{|R^P(x_1,x_2)|} \right]$.*

Worst case* Par *for a function: *the worst case* Par *for a function f is the minimum, over all protocols P for f, of the worst case* Par *of P.*

2.2 Dissection Protocols and Tiling Functions

Often in communication complexity settings the input of each party has a natural ordering, *e.g.* , the set of party i's inputs $\{0,1\}^k$ can represent the numbers $0, \ldots, 2^k - 1$ (as is the case when computing the maximum/minimum of two inputs, in the millionaires problem, in second-price auctions, and more). When designing protocols for such environments, a natural

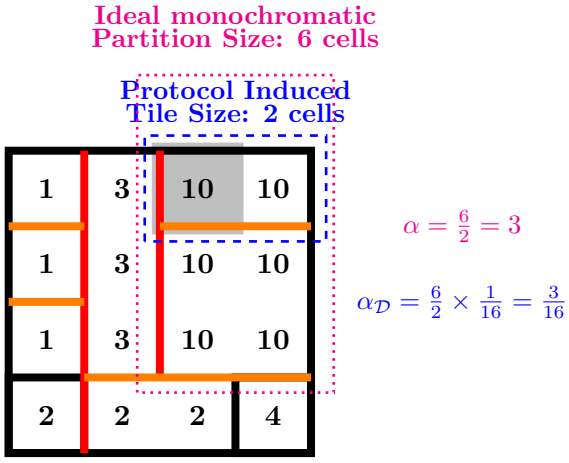


Figure 3. Privacy Approximation Ratio

restriction is to only the allow protocol to ask each party questions of the form "Is your input between a and b (in the natural order over possible inputs)?", where $a, b \in \{0,1\}^k$. The bisection protocol for the millionaires problem (1) and the bisection auction (19; 20) both fall within this category of protocols. We call this type of protocols as "dissection protocols".(Figure 4)

We now formally present dissection protocols. Given a permutation Π of $\{0,1\}^k$, we let $\prec_\Pi$ denote the order over $\{0,1\}^k$ that Π induces, *i.e.* , $\forall\, a, b \in \{0,1\}^k$, $a \prec_\Pi b$ provided b comes after a in Π. We call a subset $I \subseteq \{0,1\}^k$ *contiguous* with respect to Π if for every $a, b \in I$ and for every $c \in \{0,1\}^k$ it holds that $a \prec_\Pi c \prec_\Pi b \implies c \in I$.

Definition 2 (dissection protocol) *Given a function* $f : \{0,1\}^k \times \{0,1\}^k \mapsto \{0,1\}^t$ *and permutations* Π_1 *and* Π_2 *over* $\{0,1\}^k$*, we call a communication protocol for* f *a* dissection

protocol *with respect to* Π_1, Π_2 *if, at each communication step, the maintained subset of inputs* S_i *of each party* i *is contiguous with respect to* Π_i.

Observe that *every* protocol can be regarded as a dissection protocol with respect to *some* permutations over inputs by simply reverse-engineering to construct the permutation so that they be consistent with the way the protocol updates the maintained sets of inputs. However, not every protocol is a dissection protocol with respect to *specific* permutations. Consider, for example, the case that both Π_1 and Π_2 are the permutation over $\{0,1\}^k$ that orders the elements from lowest to highest binary values. Observe that a protocol that is a dissection protocol with respect to these permutations cannot ask questions of the form "Is your input odd or even?", for these questions partition the space of inputs into non-contiguous subsets.

We next introduce the concept of *tiling* functions. Recall that a tiling is defined to be a partition of a matrix into monochromatic rectangles.

Definition 3 (tiling function) *A function* $f : \{0,1\}^k \times \{0,1\}^k \mapsto \{0,1\}^t$ *is called a tiling function with respect to two permutations* Π_1, Π_2 *of* $\{0,1\}^k$ *(or, simply a tiling function if* Π_1 *and* Π_2 *are clear from the context) if the monochromatic regions in* $A_{\Pi_1,\Pi_2}(f)$ *form a tiling.*

For example, for a prime p, the following function is a tiling function

$$f(x_0, x_1, \ldots, x_{k-1}, y_0, y_1, \ldots, y_{k-1}) \equiv \sum_{i=0}^{k-1} (x_i + y_i) \pmod{p}$$

Here, Π_1 orders the inputs $(x_0, x_1, \ldots, x_{k-1})$ in increasing order on the value $\sum_{i=0}^{k-1} x_i \pmod{p}$; Π_2 is defined analogously.(Figure 5)

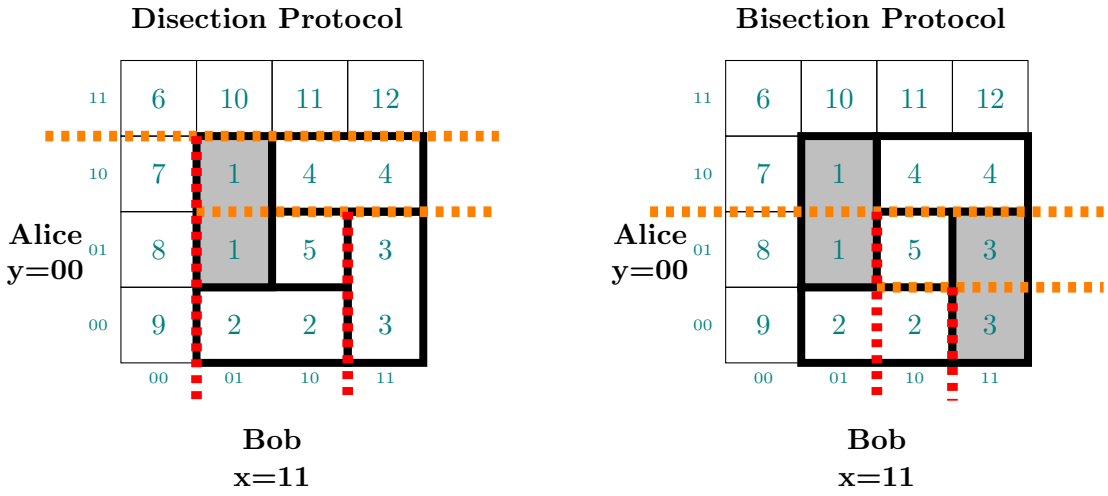


Figure 4. Bisection and Disection Protocols

Finally, a special case of interest of the dissection protocol is the bisection protocol that has been investigated in the literature in various contexts.

Definition 4 (bisection protocol) *A dissection protocol with respect to the permutations* Π_1, Π_2 *is a* bisection protocol *for a tiling function with respect to the same permutations* Π_1 *and* Π_2.

2.2.1 Some Remarks on Tiling Functions and Bisection/Dissection Protocols

Obviously it is possible to have functions f that are tiling with respect to two permutations Π_1, Π_2 and with respect to another two permutations Π_1', Π_2' where $\Pi_i \neq \Pi_i'$ and the number of monochromatic regions in the two cases differ. As a result, approximate privacy measures may differ for the two sets of permutations. For example, consider the tiling function

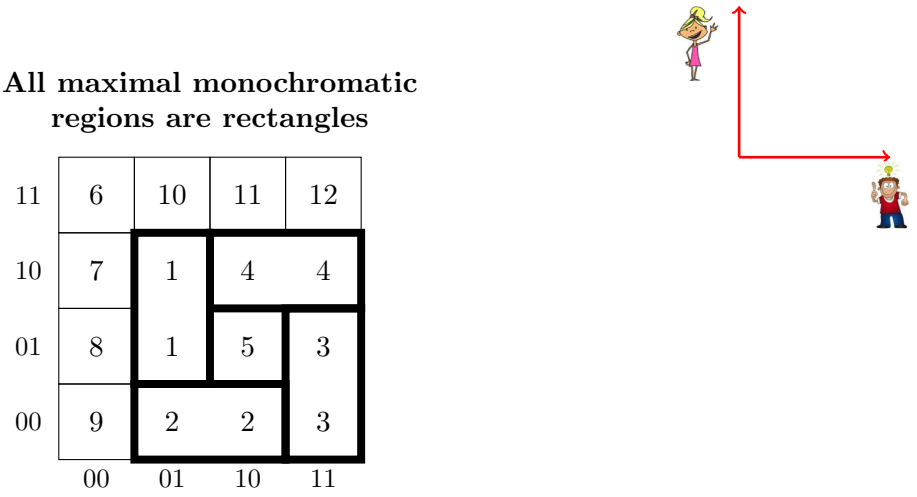


Figure 5. Tiling Function

as shown in Figure 6. With respect to the permutations Π_1, Π_2 in Figure 6(a) the function is not privately computable. But, with respect to the permutations Π_1', Π_2' in Figure 6(b) the function *is* privately computable. Thus, the tiling function in Definition 3 specifies the two permutations Π_1, Π_2 with respect to which the tiles are given so that the quantification of privacy is unambiguous; `note that these permutations are not necessarily the same that the two players use in a dissection protocol`. However, our results for upper bounds of worst-case and average privacy values are stronger: `the upper bounds hold for any two permutations that produce a tiling of the function`.

Our definition of bisection protocol is more general than the one in (1)[1] since, for example, we do not require the protocol to divide the maintained input space into two *equal* halves and the permutations of the two players for a dissection could be different.

2.2.2 Boolean Tiling Functions

We show that Boolean tiling functions are *nicely-behaving* in terms of privacy preservation.

Lemma 5 *Every Boolean tiling function can be computed in a perfectly privacy-preserving manner.*

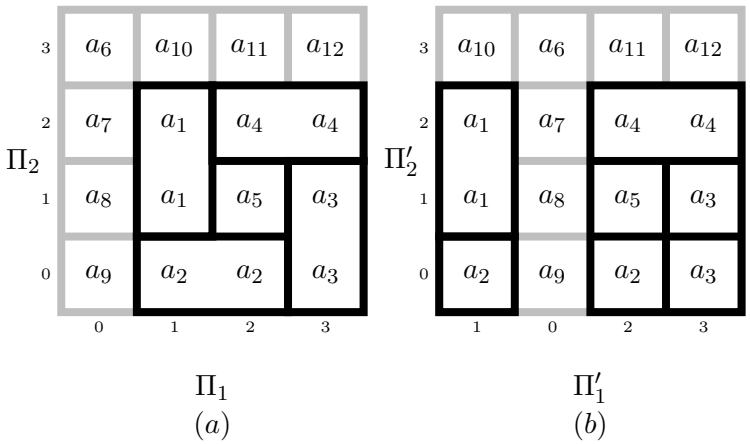


Figure 6. A function that is a tiling function with respect to two permutation pairs Π_1, Π_2 and Π'_1, Π'_2.

Proof. Consider the specific tiling in Figure 7, that contains rectangles A, B and C (the rest of the tiling is not specified in the figure). We observe that this tiling *cannot* be a tiling of the input space of a Boolean function. This is because A, B and C are maximal monochromatic regions and thus

---

[1]When both players have the same permutation, such protocols are sometimes referred to as ε-approximate bisection ($\varepsilon > 1$) provided each protocol step divides the maintained input space into two halves with each half being between $\frac{1}{\varepsilon}$ and $1 - \frac{1}{\varepsilon}$ fraction of the original.

cannot border rectangles that are of the same "color" (*i.e.* , for which f outputs the same outcome). However, because f only has two possible outcomes (0 and 1), and every two rectangles in the set $\{A, B, C\}$ border one another, we have that the existence of a tiling as in Figure 7 is impossible for a Boolean function.

This line of argument applies more generally, as we now show. Two rectangles in a tiling can be neighbors in one of two ways: either on the x axis (one is "to the left" of the other) or on the y axis (one is "above" the other). The same argument as above shows that every two rectangles that neighbor on the x axis must have the same upper and lower line-boundaries, whereas every two rectangles that neighbor on the y axis must have the same left and right line-boundaries. This implies that the tiling of every Boolean tiling must be in the form of a checkers board, and thus perfectly privately computable. ❑

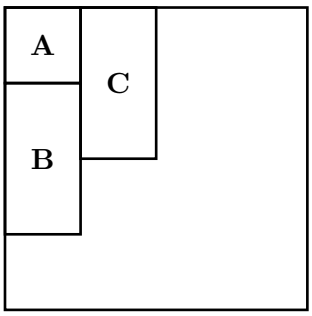


Figure 7. Illustration of the argument in the proof of Lemma 5.

Remark 1 *The claim of the above lemma does not hold if the range of the function has at least three values.*

2.3 Average and Worst Case Par for Tiling Functions

In this section, we show that *any* tiling function admits a protocol that has a *small constant* average case privacy approximation ratio. Moreover, we show that this result *cannot* be extended to the case of worst-case privacy approximation ratios.

2.3.1 Constant Average-case Par for Tiling Functions

Let $f : \{0,1\}^k \times \{0,1\}^k \mapsto \{0,1\}^t$ be a given tiling function with permutations Π_1, Π_2, and let r denote the number of monochromatic rectangles in $A_{\Pi_1,\Pi_2}(f)$ $(0 < r \leq 2^k)$. We will denote the uniform distribution over all input pairs by D_u. A c-approximate uniform distribution $\mathsf{D}_u^{\sim c}$ is a distribution in which the probabilities of two input pairs are close as a function of $c \geq 0$, namely $\max_{(\mathbf{x},\mathbf{y}),\,(\mathbf{x}',\mathbf{y}')\in\{0,1\}^k\times\{0,1\}^k} |\, \mathsf{D}_u^{\sim c}(\mathbf{x},\mathbf{y}) - \mathsf{D}_u^{\sim c}(\mathbf{x}',\mathbf{y}') \,| \leq c\, 2^{-2k}$.

Theorem 6 *The following results hold:*

(a) *For any tiling function f. there is a* `bisection` *protocol P using at most $8r$ communication steps such that*

- $\alpha_{\mathsf{D}_u^{\sim c}} \leq 4 + 4c$, *and*
- *P can be computed in $O\left(k4^k\right)$ time.*

(b) *For $0 \leq c < \frac{9}{8}$, there exists a tiling function f such that for every dissection protocol $\alpha_{\mathsf{D}_u^{\sim c}} \geq \frac{11}{9} + \frac{2}{81}c$.*

Proof. For $i = 1, 2, \ldots, r$, let the i^{th} monochromatic rectangle $R^{\text{I}}(\mathbf{x},\mathbf{y})$ in $A_{\Pi_1,\Pi_2}(f)$ contain $y_i \times 2^{2k}$ elements and suppose that a communication protocol partitions this rectangle into $t_i \geq 1$ rectangles containing $z_1, \ldots, z_{t_i}$ elements, respectively. Then the contribution of all cells in R^{I} to α_{D_u} is $\sum_{j=1}^{t_i}\left(\frac{y_i 2^{2k}}{z_j} \times \frac{z_j}{2^{2k}}\right) = t_i y_i$. Thus $\alpha_{\mathsf{D}_u} = \sum_{i=1}^{r} t_i y_i$. Similarly, one can see that $\alpha_{\mathsf{D}_u^{\sim c}} \leq \sum_{i=1}^{r}\sum_{j=1}^{t_i}\left(\frac{y_i 2^{2k}}{z_j} \times \frac{(1+c)\, z_j}{2^{2k}}\right) = \sum_{i=1}^{r}(1+c)\, t_i y_i$.

[1]A binary space partition (Bsp) for a collection of *disjoint* rectangles in the two-dimensional plane is defined as follows. The plane is divided into two parts by cutting rectangles with a line if necessary. Each fragment of the rectangle belongs solely to one of the parts it falls in. The two resulting parts of the plane are divided recursively in a similar manner; the process continues until at most one fragment of the original rectangles remains in any part of the plane. This division process can be naturally represented as a binary tree (Bsp-tree) where a node represents a part of the plane and stores the cut that splits the plane into two parts that its two children represent; each leaf of the Bsp-tree represents the final partitioning of the plane and stores at most one fragment of an input rectangle. The *size* of a Bsp is the number of leaves in the Bsp-tree. The following result was proved in (22).

Theorem 7 *(22)*[1] *Assume that we have a set $\mathcal{S}$ of r disjoint axis-parallel rectangles in the plane. Then, a* Bsp *of $\mathcal{S}$ can be computed in $O(r \log r)$ time such that every rectangle in $\mathcal{S}$ is partitioned into at most* 4 *rectangles.*

(a) Notice that any Bsp defines a bisection protocol. Thus, using $\max_i\{t_i\} \leq 4$ we get $\alpha_{\mathrm{D}^{\sim^c}_u} \leq \sum_{i=1}^{r} 4\,(1+c)\,y_i = 4\,(1+c)$. Also, each communication step either partitions a sub-rectangle of the monochromatic rectangles or partitions the sub-rectangles, thus we need at most $8r$ steps.

---

[1]As defined in (21)

[1]*Note that we cannot use the slightly stronger bounds in (21) since that applies to the* average Bsp *size.*

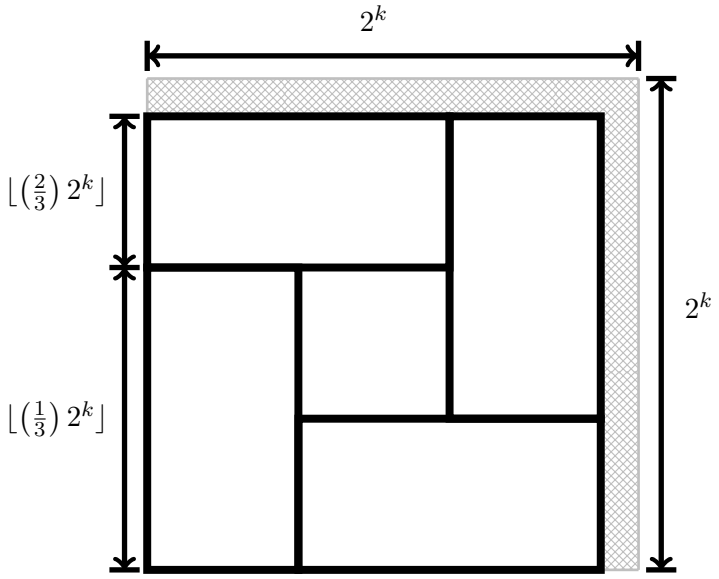


Figure 8. Example for $\alpha_{\mathsf{D}_u^{\sim c}} \geq \frac{11}{9} + \frac{2}{9}c$. The crosshatched area is covered by unit area squares.

(b) Consider the function f whose ideal monochromatic rectangles are shown in Figure 8. Each of the four non-square rectangles contain about $\left(\frac{2}{9}\right)2^{2k}$ elements and the remaining squares contain about $\left(\frac{1}{9}\right)2^{2k}$ elements. We assign a probability of about $\left(1+\frac{c}{9}\right)/2^{2k}$ to every point in the four non-square rectangles and assign a probability of $\left(1-\frac{8c}{9}\right)/2^{2k}$ to the remaining rectangles. The very first step of any dissection protocol *must* partition at least one border rectangle, giving

$$\alpha_{\mathsf{D}_u^{\sim c}} \geq \left(2 \times \frac{2+\frac{2c}{9}}{9}\right) + \frac{7-\frac{2c}{9}}{9} = \frac{11}{9} + \frac{2}{81}c \qquad \square$$

2.3.2 Large Worst-case Par for Tiling Functions

Can we extend the results of the previous section to show that for every tiling function there exists a dissection protocol that achieves a good PAR even in the worst case? We now show that the answer to this question is negative. We present a tiling function for which *every* dissection protocol has *exponential* worst-case PAR.

Theorem 8 *There exists a tiling function* $f : \{0,1\}^k \times \{0,1\}^k \mapsto \{0,1\}^t$ *such that for any two permutations* Π_1, Π_2 *of* $\{0,1\}^k$*, every protocol* P *for* f *that is a dissection protocol with respect to* Π_1, Π_2 *has* $\alpha_{\text{worst}} = \Omega\left(2^{k/2}\right)$.

Proof. Recall the example in Figure 8 that showed that there exist functions that cannot be computed in a privacy preserving manner. Our construction of the function f in the statement of the theorem is based on the function in Figure 8. We consider the specific permutations Π_1, Π_2 over $\{0,1\}^k$ that order the elements in $\{0,1\}^k$ by binary value (from 0 to $2^k - 1$). We now use the construction in Figure 8 "recursively" to create a tiling of the input space. We first embed $2^{k-1} - 1$ instances of the construction in Figure 8 recursively within one another, as described in Figure 9(a), leaving a 2×2 square at the center. We then partition each of the outermost rectangles in Figure 9(a) into two "nearly" equal-sized rectangles as described in Figure 9(b).

Consider the function f such that the monochromatic rectangles of $A_f(\Pi_1, \Pi_2)$ are the tilings in Figure 9(b) (f outputs a different outcome for each (minimal) rectangle in the figure). Clearly, f is a tiling function. We now first show that any protocol P for f that is a dissection protocol with respect to Π_1, Π_2 has an $\Omega\left(2^{k-1}\right)$ worst-case PAR.

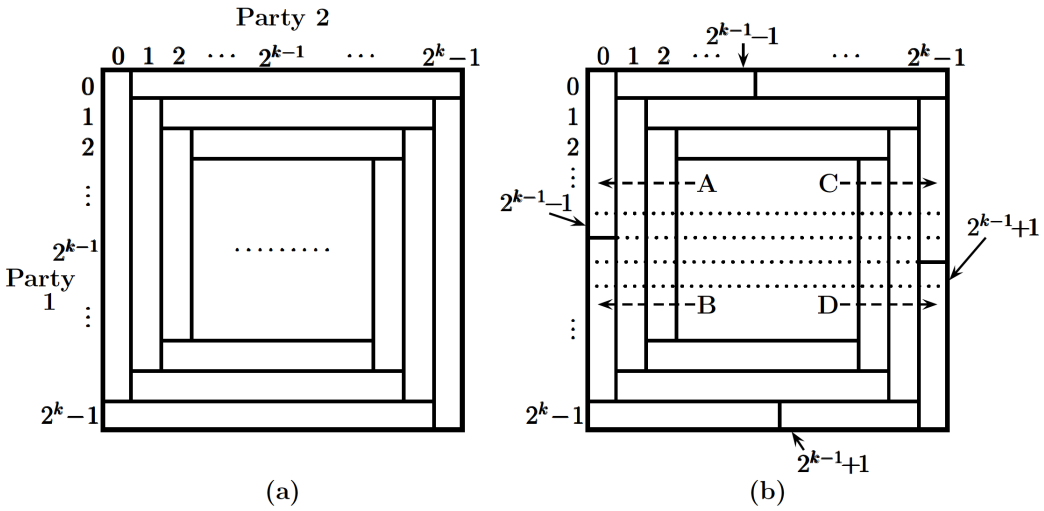


Figure 9. Illustrations of the arguments in the proof of Theorem 8. The dotted lines in **(b)** are shown for visual clarities only.

Consider a protocol P for f that is a dissection protocol with respect to Π_1, Π_2. Consider the first (meaningful) bit transmitted in the execution of P. Suppose that this bit is transmitted by party 1 (the case that the bit is transmitted by party 2 is analogous). This bit effectively partitions the input space into two nonempty rectangles, where one rectangle (the "upper rectangle") are all inputs of the form $\{0, \ldots, \Gamma\} \times \{0, \ldots, 2^k\}$ for some $0 \leq \Gamma < 2^k$. Consider the rectangles A and B in Figure 9(b). We have the following cases.

Case I: the first bit does not partition rectangle A or rectangle B. This case can be divided into subcases: either the first bit dissects 1's input space just between A and B, or it dissects 1's input space just below B.

Case I(a): the first bit dissects 1's input space just between A and B. Observe that if 2's input is $2^k - 1$ then this results in the partitioning of a rectangle of size 2^{k-1} (rectangle C) into two rectangles, one of which is of size exactly 2. Hence, $\alpha_{\text{worst}} \geq 2^{k-2}$.

Case I(b): the first bit dissects 1's input space just below B. Consider the case that 2's input is 2^{k-1}. Observe that in this case rectangle D is also partitioned similarly to subcase **I(a)**, and thus $\alpha_{\text{worst}} \geq 2^{k-1}$.

Case II: either rectangle A or rectangle B is partitioned. We focus on the case that rectangle A is partitioned (the case that rectangle B is partitioned is analogous). A is partitioned into two contiguous rectangles and so there must exist some $0 < \Gamma < 2^{k-1} - 1$ such that all of 1's inputs below (and equal to) Γ are separated from all of 1's inputs that lie above Γ.

Case II(a): $0 < \Gamma \leq 2^{k-1} - 2^{k/2}$. Observe that for every such value Γ there exists a vertical rectangle of width 1 (that is, an input of party 2) and of size (length) $2 \times (|A| - \Gamma)$ that is partitioned into two rectangles, one of which is of size exactly 1 (the greater the value Γ the more to the right the partitioned rectangle is). Since $|A| - \Gamma \geq 2^{\frac{k}{2}} - 1$ we have $\alpha_{\text{worst}} = \Omega\left(2^{k/2}\right)$.

Case II(b): $2^{k-1} - 2^{\frac{k}{2}} < \Gamma < 2^{k-1} - 1$. Observe that, for every such value of Γ, A (which is of size 2^{k-1}) is partitioned into two rectangles, of which one is of size at most $2^{k/2}$. Thus, in this case too $\alpha_{\text{worst}} = \Omega\left(2^{k/2}\right)$.

The above argument shows that for Π_1, Π_2 the statement of the theorem holds. To show that the statement of the theorem holds for every tiling-inducing permutations we observe that all tiling-inducing permutations for f induce (roughly) the same tiling. In this case, the first meaningful bit transmitted by party 1 partitions the input space into two nonempty regions, where one region consists of all inputs from $\Delta \times \{0, \ldots, 2^k\}$ for some $\emptyset \subset \Delta \subset \{0, 1, 2 \ldots, 2^k - 1\}$. A similar analysis can now be carried out by considering the intersection of this region with rectangles A and B. ❑

2.4 Extensions of the Basic Two-player Setup

In this section, we briefly discuss two extensions to the basic setup of two-party communication model described before.

2.4.1 Non-tiling Functions

A natural extension of the classes of functions to be computed involves relaxing the constraint of the tiling functions, namely that each monochromatic region *must* be a rectangle.

Definition 9 (Δ-approximate tiling function) *A function $f : \{0,1\}^k \times \{0,1\}^k \mapsto \{0,1\}^t$ is a called a Δ-approximate tiling function provided there exists two permutations Π_1, Π_2 of $\{0,1\}^k$ such that each monochromatic region in $A_{\Pi_1,\Pi_2}(f)$ is an* ***union of at most Δ disjoint rectangle****.*

Proposition 1 *For any Δ-approximate tiling function f with r monochromatic regions, there is a bisection protocol P using at most $8r\Delta$ communication steps such that*

- $\alpha_{\mathrm{D}_u^{\sim c}} \leq (4 + 4c)\Delta$, *and*

- *P can be computed in $O\left(k4^k\right)$ time.*

Proof. We use the algorithm of Theorem 6 on the set of at most $r\Delta$ rectangles obtained by partitioning each monochromatic region into rectangles. Since each rectangle is partitioned at most 4 times, each monochromatic region of f will be partitioned at most 4Δ times. ❑

2.4.2 Multi-party Computation

Another natural extension of the basic two-player setup is to consider the case of $d > 2$ players computing a d-argument function $f : \underbrace{\{0,1\}^k \times \{0,1\}^k \times \cdots \times \{0,1\}^k}_{d \text{ times}} \mapsto \{0,1\}^t$. We need to adjust some definitions in the following natural manner:

- Players are sequentially ordered 1 to d.
- The tiling function has permutation Π_i for the i^{th} argument of f (or, equivalently for player i) for $1 \leq i \leq d$. The input space is now the d-dimensional space $\{0,1\}^k \times \{0,1\}^k \times \cdots \times \{0,1\}^k$ and each tile is a d-dimensional hyper-rectangles (Cartesian product of d intervals).
- A dissection protocol is generalized to a "round robin" dissection protocol in the following manner. In one "mega" round of communications, players communicate in the order player 1, player 2, ..., player d, and the mega round is repeated if necessary. Any communication made by any player is available to *all* the remaining players. Thus, each communication of the dissection protocol partitions a d-dimensional space into two d-dimensional space by an appropriate $(d-1)$-dimensional hyperplane.

- All the definitions of Sections 2.1 and 2.2 can now be easily generalized in a similar manner.

How good is the average PAR for d-dimensional tiling functions? For general d, it is non-trivial to compute precise bounds because each player i has her/his own permutation Π_i of the input, the tiles are boxes of "full" dimension and separating hyperplanes must be of dimension exactly $d-1$. Nonetheless, we show that the average PAR is very high for dissection protocols even for 3 players and uniform distribution, thereby suggesting that this quantification of privacy may not provide good bounds for three or more players.

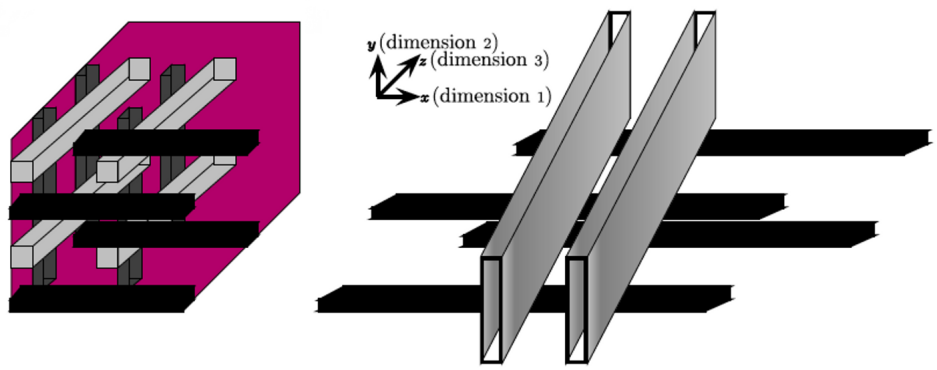


Figure 10. **(a)** Illustration of the 3-dimensional tiling function in Lemma 10 for $k = 3$. The non-trivial hyper-rectangles for each dimension are shown colored by light gray, dark gray and black; the trivial hyper-rectangles cover the region colored magenta. **(b)** Hyper-rectangles corresponding to one protocol step for player 1.

Lemma 10 (large average Par for dissection protocols with 3 players) *There are tiling functions* $f : \{0,1\}^k \times \{0,1\}^k \times \{0,1\}^k \mapsto \{0,1\}^{O(k)}$ *such that for every dissection protocol* $\alpha_{\mathrm{D}_u} = \Omega\left(2^k\right)$.

Proof. The tiling function for our claim is adopted from an example of the paper by Paterson and Yao (23; 24) with appropriate modifications[1]. For convenience, we refer to the arguments of function f as decimal equivalent of the corresponding binary numbers. The three arguments of f are referred to as dimension 1, 2 and 3, respectively. The area of a hyper-rectangle $R = [a, a'] \times [b, b'] \times [c, c'] \in \{0, 1, \ldots, 2^k - 1\}^3$ is $\mathsf{Area}(R) = (a' - a + 1) \times (b' - b + 1) \times (c' - c + 1)$. We will show the tiling for the function f; see Figure 10 for an illustration of our construction for $k = 3$.

For each dimension, we have a set of $\Theta\left(2^k\right)$ hyper-rectangles; we refer to these hyper-rectangles as *non-trivial* hyper-rectangles for this dimension. For dimension 1, these hyper-rectangles are of the form $[0, 2^k - 1] \times [4x, 4x] \times [4y, 4y]$ for every $0 \leq x, y < \frac{2^k-1}{4}$. Similarly, for dimensions 2 and 3, the non-trivial hyper-rectangles are of the form $[4x + 1, 4x + 1] \times [0, 2^k - 1] \times [4y + 1, 4y + 1]$ and $[4x + 2, 4x + 2] \times [4y + 2, 4y + 2] \times [0, 2^k - 1]$, respectively. It is easy to see that the non-trivial hyper-rectangles are mutually disjoint, each of them is of area 2^k and the sum of their areas is $\Omega(2^{3k})$. The remaining "trivial" hyper-rectangles are each of unit area such that they together cover the remaining input space. It now also follows that the number of monochromatic regions is $\Theta\left(2^{3k}\right)$. Let $\mathcal{S}_i$ be the set of all non-trivial hyper-rectangles

---

[1] According to (23; 24), the example came from a private communication with W. Thurston.

corresponding to the i^{th} dimension and $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$. Suppose that a dissection protocol partitions, for $i = 1, 2, \ldots, |\mathcal{S}|$, the j^{th} non-trivial hyper-rectangle R_j into t_j hyper-rectangles; then the same argument as in Theorem 6 leads us to $\alpha_{\mathrm{D}_u} = \sum_{j=1}^{|\mathcal{S}|} \frac{\mathsf{Area}(R_j)}{2^{3k}} y_j = \frac{\sum_{j=1}^{|\mathcal{S}|} y_j}{2^{2k}}$. Thus, it suffices to show that $\sum_{j=1}^{|\mathcal{S}|} y_j = \Omega\left(2^{3k}\right)$.

Consider any player, say player 1 corresponding to the first dimension. For each communication step of this player, the maintained set of inputs is a subset of the maintained set of inputs in the previous step. Thus, each meaningful communication step of player 1 geometrically corresponds to a *set* of 3-dimensional hyper-rectangles as illustrated in Figure 10(b). Notice that each of these hyper-rectangles produce $\Omega\left(2^{2k}\right)$ *new* fragments of the primary hyper-rectangles corresponding to the first dimension. To finish the proof, consider the line $y = z = 4$ and place points on this line starting from $x = 0$ consecutively at a distance of 4 apart. It can be easily seen that the final set of hyper-rectangles produced by player 1 cannot contain two such points. Thus, the number of hyper-rectangles must be $\Omega\left(2^{k}\right)$ and consequently $\sum_{j=1}^{|\mathcal{S}|} y_j = \Omega\left(2^{3k}\right)$. ❑

Remark 2 *A generalized version of the example in d dimension (e.g. , see (25)) can be used to provide a slightly improved lower bound on* α_{D_u} *for bisection protocols more than three players; the bound asymptotically approaches* $\Omega\left(2^{2k}\right)$ *for large d.*

2.5 Analysis of the Bisection Protocol for Two Boolean Functions

In Section 2.2.2 we showed that any Boolean function can be computed with perfect privacy by a dissection protocol. In this section, we analyze the *bisection* protocol (19; 20), a special case of the general dissection protocol, for two Boolean functions that appear in the literature.

In bisection protocol, each party does a binary search on the ordering of its inputs until the result is revealed. As before, D_u denotes the uniform distribution. In (1) the authors provided calculated bounds on α_{worst} and α_{D_u} for the bisection protocol on a few functions. In this section, we show similar calculations for two Boolean functions. Letting $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \{0,1\}^k$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n) \in \{0,1\}^k$, the functions that we consider are the following:

AND-OR function: $f_{\wedge,\vee}(\mathbf{x},\mathbf{y}) = \bigwedge_{i=1}^{n}(x_i \vee y_i)$. For example, each bit may indicate the availability of a specific resource and a 1 output of the function ensures that every resource is available to at least one of the parties.

Equality function: $f_{=}(\mathbf{x},\mathbf{y}) = \begin{cases} 1 & \text{if } \forall\, i:\ x_i = y_i \\ 0 & \text{otherwise} \end{cases}$.

A summary of our bounds are as follows.

Function	α_{D_u}	α_{worst}
$f_{\wedge,\vee}$	$\geq \left(\frac{3}{2}\right)^{2k}$	
$f_{=}$	$= 2^k - 2 + 2^{1-k}$	$= 2^{2k-1} - 2^{k-1}$

We will use the formula for α_{D_u} that we derived in the proof of Theorem 6: *letting r denote the number of monochromatic regions in an ideal partition of the function if, for $i = 1, 2, \ldots, r$, the i^{th} monochromatic region contain $y_i \times 2^{2k}$ elements and the bisection protocol partitions this region into $t_i \geq 1$ rectangles containing $z_1, \ldots, z_{t_i}$ elements, respectively, then $\alpha_{\mathsf{D}_u} = \sum_{i=1}^{r} t_i y_i$.* In the sequel, by "contribution of a rectangle (of the bisection protocol) to the (average PAR)" we mean the size of the ideal monochromatic region that the rectangle is a part.

2.5.1 AND-OR Function

Theorem 11 $\alpha_{\mathrm{D}_u} \geq (3/2)^{2k}$.

Proof. We begin by showing the geometry of the tilings for small values of k which easily generalizes to larger k. The ideal tiling for $f_{\wedge,\vee}$ is shown in Figure 11(a) for $k = 3$ with the value of the function for each input pair. The sizes of the ideal monochromatic partition are shown in Figure 11(b) for $k = 1, 2, 3, 4$. The contributions to the average PAR of various inputs after applying the bisection protocol are illustrated in Figure 12 for $k = 1, 2, 3, 4$. We observe the following:

- The tiles colored *light gray* for the case when $k = 4$ are referred to as the "background tiles". For $k = 1, 2, 3, 4$ each such tile contributes $3, 9, 27$ and 81, respectively, to the average PAR. In general, this contribution is given by 3^k and all these tiles have size 1.
- The contributions of the tiles in the upper-left region of the matrix are given by the sum of the first $2^k - 1$ natural numbers; thus each of these tiles contribute $2^{2k-1} - 2^{k-1}$.
- For any k, observe that the matrix can be decomposed into 4 quadrants; the following observations can be repeated recursively on each resulting quadrant, except for the first quadrant:
  - The first quadrant is a monochromatic region that contributes $2^{2k-1} - 2^{k-1}$ to the average PAR.
  - The fourth quadrant has the same structure as the original matrix, but the contributions for the *non-background* tiles will be related to the case of a matrix with j

bits instead of k, where the size of the quadrant is 2^j. For example, notice that the fourth quadrant of a matrix with $k = 4$ is the same as a whole matrix with $k = 3$, except for the "background tiles", that always contribute for 3^k, with the original value of k.

– The second and third quadrants are similar to the fourth quadrant case, but in this case the values in the upper-left portion of the quadrants will remain the same as the original matrix, instead of going down as with the fourth quadrant case.

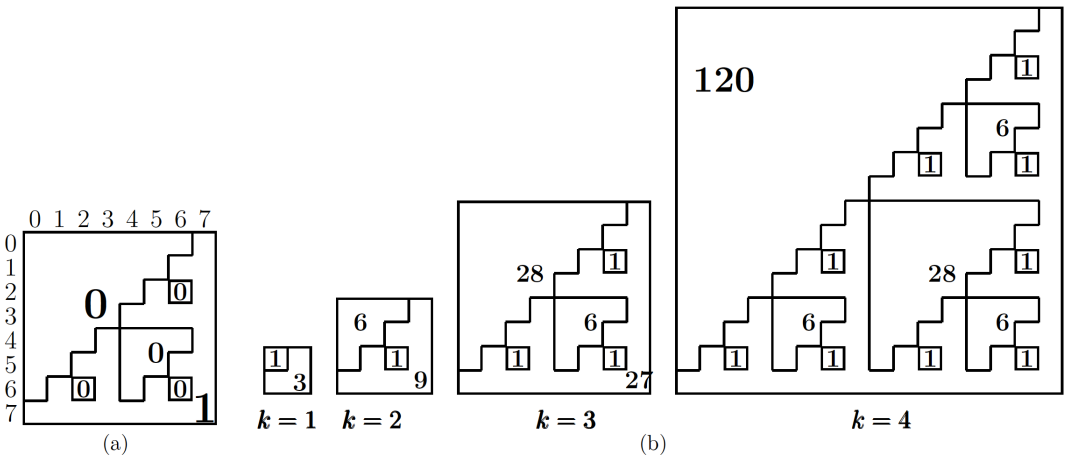


Figure 11. **(a)** Ideal monochromatic partition for $f_{\wedge,\vee}$ when $k = 3$. **(b)** Sizes of ideal monochromatic partition for $f_{\wedge,\vee}$.

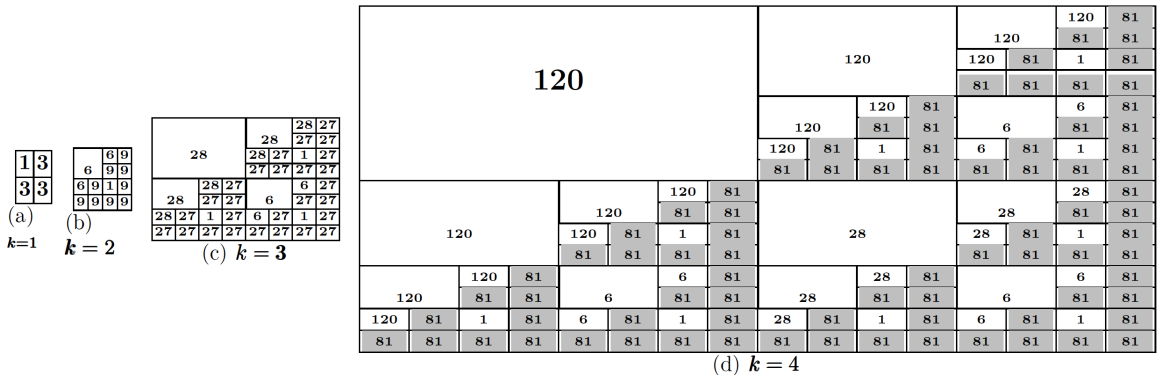


Figure 12. Contribution to PAR for $k = 0, 1, 2, 3, 4$.

Based on these observations, we can obtain a recurrence for the total contribution to the average PAR of all the tiles in a generic matrix. We need the following parameters:

- The number of bits in the original matrix, that we denote by k;
- The number of bits corresponding to the size of the matrix, or submatrix being considered, that we denote by i;
- The number of bits to be used in the calculation of the contribution of the upper-left portion of the matrix, or submatrix, being considered; we denote this by j.

The recurrence that computes the total contribution to the PAR of all the tiles in the matrix is:

$$g(i,j,k) = \begin{cases} 3^k, & \text{if } i = 0 \\ 2^{2j-1} - 2^{j-1} + 2g(i-1,j,k) + g(i-1,i-1,k), & \text{otherwise} \end{cases}$$

The values of i and j are initially set to the value of k. The interpretation of each term in the above recurrence is as follows:

- 3^k is the contribution of each "background tile";
- $2^{2j-1} - 2^{j-1}$ is the contribution of the first quadrant;
- $g\,(i-1, j, k)$ is the contribution of each one of the second and third quadrants and
- $g\,(i-1, i-1, k)$ is the contribution of the fourth quadrant.

Remember that, for a given k, the recurrence equation is initialized with $i = j = k$. Thus, we have:

Case: k = 0: $g\,(k, k, k) = 3^k = 3^{2k}$.

Case: k > 0: $g(k,k,k) = g(k-1,k-1,k) + 2g(k-1,k,k) + t(k)$. The second parameter to the function indicates how to generate the $t(k)$ terms; the value of such terms is proportional to that parameter. Thus, for $a \geq b$, $g(k,a,k) \geq g(k,b,k)$. For our lower bound, we can neglect the terms $t(k)$. Thus, we obtain:

$$g(k,k,k) \geq 3g(k-1,k-1,k) \geq 3g(k-2,k-2,k) \geq \cdots\cdots \geq 3g(1,1,k) \geq 3g(0,0,k)$$

For each step, the value of the first parameter decreased exactly by one unit, so after k iterations the value of the first parameter will be zero. Hence we have $g(k,k,k) \geq 3^k g(0,0,k)$. Since $g(0,0,k) = 3^k$ we finally obtain $g(k,k,k) \geq 3^k \times 3^k = 3^{2k}$.

Thus, $\alpha_{\mathsf{D}_u} = g(k,k,k)/2^{2k} \geq (3/2)^{2k}$. ❑

2.5.2 Equality function

Theorem 12 $\alpha_{\mathsf{D}_u} = 2^k - 2 + 2^{1-k}$ *and* $\alpha_{\text{worst}} = 2^{2k-1} - 2^{k-1}$.

Proof. An illustration of the ideal partition into monochromatic *regions* for equality function is shown in Figure 13 **(a)**. After running the bisection protocol, the induced tiling is (for $k = 3$) is shown in Figure 13 **(b)**. Excluding the diagonal, we have 2 tiles of size 16, 4 tiles of size 4, and 8 tiles of size 1. In general, it is easy to observe that, for each $0 \leq i < k$, we have exactly 2^{k-i} tiles of size 2^{2i}.

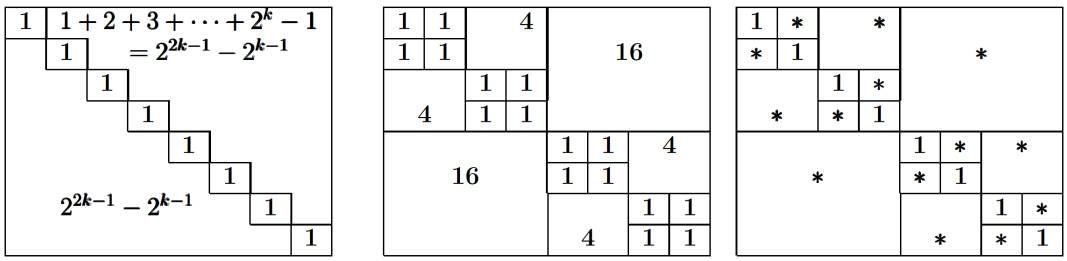


Figure 13. **(a)** Ideal tiling for equality function. **(b)** The induced tiling by the bisection protocol (shown for $k = 3$). **(c)** Contribution of each rectangle in protocol-induced tiling where $* \equiv \mathbf{2^{2k-1} - 2^{k-1}}$. The numbers in the figure denote the size of each tile.

The following accounting scheme can be used to simplify calculation. For uniform distribution D_u, α_{D_u} is the sum of the ratio $\frac{|R^I(i,j)|}{|R^P(i,j)|}$ over each element (i, j) in the matrix divided by the number of total elements 2^{2k} in the matrix, where $R^I(i, j)$ and $R^P(i, j)$ is the size of the ideal

and protocol-induced tiling that contains the cell (i, j). Consider a rectangle A of size m in the protocol-induced tiling and suppose that A is contained in a monochromatic region of the ideal partition of size m'. Then, the sum of contributions of the elements of A is $\sum_{i=1}^{m} m'/m = m'$. Thus, the total contribution of the rectangle A is simply the size of region of the ideal partition containing it.

Figure 13 **(c)** illustrates the contribution of each rectangle in the protocol-induced tiling to average PAR. We can calculate the total contribution to the average PAR of all the tiles in the matrix, except the diagonal, by multiplying $2^{2k-1} - 2^{k-1}$ by the number of tiles. The number of tiles is given by: $\sum_{i=0}^{k-1} 2^{k-i} = 2^{k+1} - 2$. The total contribution of those tiles is $(2^{k+1} - 2) \times \left(2^{2k-1} - 2^{k-1}\right) = 2^{3k} - 2^{2k+1} + 2^k$. The contribution of the diagonal is $\underbrace{1 + 1 + \cdots\cdots + 1}_{2^k \text{ times}} = 2^k$. Since the average objective PAR α_{D_u} is the sum of the total contributions divided by the number of cells in the matrix, we have

$$\alpha_{\mathrm{D}_u} = \frac{2^{3k} - 2^{2k+1} + 2^k + 2^k}{2^{2k}} = \frac{2^{3k} - 2^{2k+1} + 2^{k+1}}{2^{2k}} = 2^k - 2 + 2^{1-k}$$

It can be seen from the ideal and protocol tilings that the worst case for PAR is the one in which the ideal tile size is $2^{2k-1} - 2^{k-1}$, and the protocol tile size is 1. Thus $\alpha_{\text{worst}} = 2^{2k-1} - 2^{k-1}$. ❑

CHAPTER 3

SOCIAL NETWORKS PRIVACY

Online Social Networks have become very popular in the recent years. Such networks provide a platform for users to publicize their private informations. It is obviously desirable to know how secure a given social network is against active attacks.

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 (1; 2; 17; 18; 26)). The differential privacy model, introduced by Dwork (4) 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 chapter.

However, none of these settings apply directly to our application scenario of active attack model for social networks. This necessitates the study of computational complexity issues for computing (k, ℓ)-anonymity. Currently known results only include some heuristic algorithms with no provable guarantee on performances such as in (14), or algorithms for very special cases. In fact, it is not even known if any version of the related computational problems is NP-hard. To this effect, we formalize three computational problems related to measuring the

---

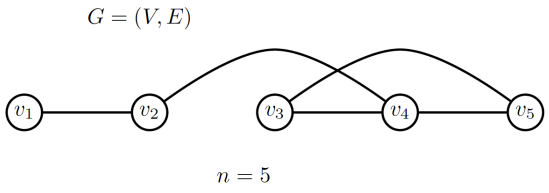
[0]The contents of this chapter are taken from arXiv:1510.08779 [cs.CC]

(k, ℓ)-anonymity of graphs and present non-trivial computational complexity results for these problems.

3.1 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.

3.1.1 Basic Terminologies and Notations



dist_{v_i,v_j} values

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

Figure 14. An example to illustrate the notations in Section 3.1.1.

Let $G = (V, E)$ be our *undirected unweighted* input graph over n nodes $v_1, v_2, \ldots, v_n$. We use 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 Figure 14.

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

- $\mathsf{diam}(G) = \max\limits_{v_i, v_j \in V} \left\{ \mathrm{dist}_{v_i, v_j} \right\}$ is the *diameter* (length of a longest shortest path) of the graph $G = (V, E)$. For example, $\mathsf{diam}(G) = 3$.

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

- For a subset of nodes $V' \subset V$ 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 dist_{v_i, v_j} for every $v_j \in V \setminus V'$. For example, $\mathbf{d}_{v_2, -\{v_1, v_3\}} = (1, 2)$.

- $\mathcal{D}_{V'', -V'} = \left\{ \mathbf{d}_{v_i, -V'} \,|\, v_i \in V'' \right\}$ for any $V'' \subseteq V \setminus V'$. For example, if $V'' = \{v_2, v_4\}$ then $\mathcal{D}_{V'', -\{v_1, v_3\}} = \big\{(1, 2), (2, 1)\big\}$.

- $\Pi = \{V_1, V_2, \ldots, V_k\}$ is a partition of $V' \subseteq V$ if and only if $\cup_{t=1}^{k} V_t = V'$ and $V_i \cap V_j = \emptyset$ for $i \neq j$.

  - Partition $\Pi' = \{V'_1, V'_2, \ldots, V'_\ell\}$ is called a *refinement*[1] of partition Π, denoted by $\Pi' \prec_r \Pi$, provided $\cup_{t=1}^{\ell} V'_t \subset \cup_{t=1}^{k} V_t$ and Π' can be obtained from Π in the following manner:

    - For every node $v_i \in \big(\cup_{t=1}^{k} V_t\big) \setminus \big(\cup_{t=1}^{\ell} V'_t\big)$, remove v_i from the set containing it in Π.
    - *Optionally*, for every set V_ℓ in Π, replace V_ℓ by a partition of V_ℓ.
    - Remove empty sets, if any.

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

---

[1] Our definition is slightly different from the standard definition of refinement since we have $\cup_{t=1}^{\ell} V'_t \subset \cup_{t=1}^{k} V_t$.

- The equality relation over a set of vectors, all of same length, obviously defines an *equivalence relation*. The following notations are used for such an equivalence relation over the set of vectors $\mathcal{D}_{V\setminus V',-V'}$ for some $\emptyset \subset V' \subset V$.

  - The set of equivalence classes, which forms a partition of $\mathcal{D}_{V\setminus V',-V'}$, is denoted by $\Pi^{=}_{V\setminus V',-V'}$. For example,

    $\Pi^{=}_{\{v_1,v_2,v_3\},-\{v_4,v_5\}} = \Big\{ \big\{(2,3)\big\}, \big\{(1,2)\big\}, \big\{(1,1)\big\} \Big\}$.

  - Abusing terminologies slightly, two nodes $v_i, v_j \in V \setminus V'$ will be said to belong to the *same* equivalence class if $\mathbf{d}_{v_i,-V'}$ and $\mathbf{d}_{v_j,-V'}$ belong to the same equivalence class in $\Pi^{=}_{V\setminus V',-V'}$, and thus $\Pi^{=}_{V\setminus V',-V'}$ 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 $\Big\{ \big\{v_1\big\}, \big\{v_2\big\}, \big\{v_3\big\} \Big\}$.

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

3.1.2 Problem Definitions

It is obviously desirable to know how secure a given social network is against active attacks. This necessitates the study of computational complexity issues for computing (k, ℓ)-anonymity. To this effect, we formalize three computational problems related to measuring the (k, ℓ)-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 $\mu\left(\mathcal{D}_{V\setminus V',-V'}\right)$.

Notation related to Problem 1 $k_{\text{opt}} = \max\limits_{\emptyset\subset V'\subset V}\left\{\mu\left(\mathcal{D}_{V\setminus V',-V'}\right)\right\}$.

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 $\text{Adim}_{\geq k}$) *Given G and a positive integer k, find a subset of nodes V' of* minimum *cardinality such that $\mu\left(\mathcal{D}_{V\setminus V',-V'}\right)\geq k$, if such a V' exists.*

Notation and assumption related to Problem 2 $\mathcal{L}_{\text{opt}}^{\geq k} = \left|V_{\text{opt}}^{\geq k}\right| = \min\left\{|V'| \;\middle|\; \mu\left(\mathcal{D}_{V\setminus V',-V'}\right)\geq k\right\}$ for some $\emptyset\subset V_{\text{opt}}^{\geq k}\subset V$. If $\mu\left(\mathcal{D}_{V\setminus V',-V'}\right)\geq k$ for *no* V' then we set $\mathcal{L}_{\text{opt}}^{\geq k}=\infty$ and $V_{\text{opt}}^{\geq k}=\emptyset$.

Problem 2 finds a k-anti-resolving set for largest 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 ℓ, a k'-metric anti-dimension of ℓ', $k'>k$ and $\ell'<\ell$. Then, this provides a trade-off between privacy and number of attacker nodes, namely we may allow a smaller

privacy violation probability $1/k'$ but the network can tolerate *adversarial control* of a *fewer* number ℓ' of nodes or we may allow a larger privacy violation probability $1/k$ but the network can tolerate adversarial control of a larger number ℓ 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 ($k_{=}$-metric antidimension or $\mathbf{Adim}_{=k}$) *Given G and a positive integer k, find a subset of nodes V' of* minimum *cardinality such that $\mu\left(\mathcal{D}_{V\setminus V',-V'}\right) = k$, if such a V' exists.*

Notation and assumption related to Problem 3 $\mathcal{L}_{\text{opt}}^{=k} = \left|V_{\text{opt}}^{=k}\right| = \min\left\{ |V'| \;\middle|\; \mu\left(\mathcal{D}_{V\setminus V',-V'}\right) = k\right\}$ for some $\emptyset \subset V_{\text{opt}}^{=k} \subset V$. If $\mu\left(\mathcal{D}_{V\setminus V',-V'}\right) = k$ for *no* V' then we set $\mathcal{L}_{\text{opt}}^{=k} = \infty$ and $V_{\text{opt}}^{=k} = \emptyset$

3.1.3 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, Ω, o and ω notations used in asymptotic analysis of algorithms (e.g. , see (27)).*

An algorithm $\mathcal{A}$ for a minimization (resp., maximization) problem is said to have an *approximation ratio* of ε (or is simply an *ε-approximation*) (28) provided $\mathcal{A}$ runs in polynomial time in the size of its input and produces a solution with an objective value *no larger than* ε times (resp., *no smaller than* $1/\varepsilon$ times) the value of the optimum. $\text{DTIME}\left(n^{\log\log n}\right)$ 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 $\mathsf{NP} \not\subset \text{DTIME}(n^{\log\log n})$.

The minimum set-cover problem (Sc) is a well-known combinatorial problem that is defined as follows (27; 29). Our input is an universe $\mathcal{U} = \{a_1, a_2, \ldots, a_n\}$ of n elements, and a collection

of m sets $S_1, S_2, \ldots, S_m \subseteq \mathcal{U}$ over this universe with $\cup_{j=1}^{m} S_j = U$. A valid solution of SC is a subset of indices $\mathcal{I} \subseteq \{1, 2, \ldots, m\}$ such that every element in $\mathcal{U}$ is "covered" by a set whose index is in $\mathcal{I}$, *i.e.* , $\forall\, a_j \in \mathcal{U}\ \exists\, i \in \mathcal{I}\ :\ a_j \in S_i$. The objective of SC is to *minimize* the number $|\mathcal{I}|$ of selected sets. We use the notation 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 (29) and, assuming $NP \not\subseteq \text{DTIME}\big(n^{\log\log n}\big)$, SC does not admit a $(1-\varepsilon)\ln n$-approximation for any constant $0 < \varepsilon < 1$ (30). On the algorithmic side, SC admits a $(1+\ln n)$-approximation using a simple greedy algorithm (31) that can be easily implemented to run in $O\big(\sum_{i=1}^{m} |S_i|\ \big)$ time (27).

3.2 Our Results

In this section we provide precise statements of our results, leaving their proofs in Sections 3.3–3.5.

3.2.1 Polynomial Time Solvability of Adim and $\text{Adim}_{\geq k}$

Theorem 13

(a) *Both* ADIM *and* $\text{ADIM}_{\geq k}$ *can be solved in* $O\left(n^4\right)$ *time.*

(b) *Both* ADIM *and* $\text{ADIM}_{\geq k}$ *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 3 *The randomized algorithm in Theorem* 13(b) *runs faster that the deterministic algorithm in Theorem* 13(a) *provided* $k = \omega(\log n)$*.*

3.2.2 Computational Complexity of $\text{Adim}_{=k}$

3.2.2.1 The Case of Arbitrary k

Theorem 14

(a) *$\text{ADIM}_{=k}$ 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 $\mathsf{NP} \not\subseteq \text{DTIME}\left(n^{\log\log n}\right)$, there exists a universal constant $\delta > 0$ such that $\text{ADIM}_{=k}$ does not admit a $\left(\frac{1}{\delta}\ln n\right)$-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 $\mathcal{L}_{\text{opt}}^{=k} = c$ if a solution exists and $\text{ADIM}_{=k}$ can be solved in polynomial time.*

Remark 4

(a) *For $k = 1$, the inapproximability ratio in Theorem 14(a) is asymptotically optimal up to a constant factor because of the $(1 + \ln(n-1))$-approximation of $\text{ADIM}_{=1}$ in Theorem 15(a).*

(b) *The result in Theorem 14(b) provides a much stronger inapproximability result compared to that in Theorem 14(a) at the expense of a slightly weaker complexity-theoretic assumption (* i.e. *, $\mathsf{NP} \not\subseteq \text{DTIME}\left(n^{\log\log n}\right)$ vs. $P \neq \mathsf{NP}$).*

3.2.2.2 The Case of $k = 1$

Note that even when $k = 1$ $\text{ADIM}_{=k}$ is NP-hard and even hard to approximate within a logarithmic factor due to Theorem 14. We show the following algorithmic results for $\text{ADIM}_{=k}$ when $k = 1$.

Theorem 15

(a) $\textsc{Adim}_{=1}$ *admits a* $(1+\ln(n-1)\,)$*-approximation in* $O\left(n^3\right)$ *time.*

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

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

3.3 Proof of Theorem 13

(a) We first consider the claim for $\textsc{Adim}_{\geq k}$. We begin by proving some structural properties of valid solutions for $\textsc{Adim}_{\geq k}$.

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* $\Pi^{=}_{V\setminus V_1,-V_1}$*. Then* v_i *and* v_j *do not belong to the same equivalence class in* $\Pi^{=}_{V\setminus V_2,-V_2}$ *also.*

Proof. Since v_i and v_j are not in the same equivalence class in $\Pi^{=}_{V\setminus V_1,-V_1}$, we have $\mathbf{d}_{v_i,-V_1}\neq\mathbf{d}_{v_j,-V_1}$ which in turn implies (since $V_1\subset V_2$) $\mathbf{d}_{v_i,-V_2}\neq\mathbf{d}_{v_j,-V_2}$ which implies v_i and v_j are not in the same equivalence class in $\Pi^{=}_{V\setminus V_2,-V_2}$. ❑

Corollary 16 *Proposition 1 implies* $\Pi^{=}_{V\setminus V_2,-V_2}\prec_r\Pi^{=}_{V\setminus V_1,-V_1}$.

Note that $\Pi^{=}_{V\setminus V_2,-V_2}\prec_r\Pi^{=}_{V\setminus V_1,-V_1}$ in Corollary 16 does not necessarily imply that $\mu\left(\mathcal{D}_{V\setminus V_2,-V_2}\right)\leq\mu\left(\mathcal{D}_{V\setminus V_1,-V_1}\right)$. The following proposition gives some condition for this to happen.

Proposition 2 *Consider two subsets of nodes $\emptyset \subset V_1 \subset V_2 \subset V$, and let $S_1, S_2, \ldots, S_\ell \subseteq V \setminus V_1$ be the only $\ell > 0$ equivalence classes (subsets of nodes) in $\Pi^{=}_{V\setminus V_1,-V_1}$ such that $|S_1| = |S_2| = \cdots = |S_\ell| = \mu\left(\mathcal{D}_{V\setminus V_1,-V_1}\right)$. Then,*

- $\triangleright$ *$\cup_{t=1}^{\ell} S_t \not\subseteq V_2 \setminus V_1$ implies $\mu\left(\mathcal{D}_{V\setminus V_2,-V_2}\right) \leq \mu\left(\mathcal{D}_{V\setminus V_1,-V_1}\right)$, and*
- $\triangleright$ *if $\emptyset \subset V_2 \cap S_j \subset S_j$ for some $j \in \{1, \ldots, \ell\}$ then $\mu\left(\mathcal{D}_{V\setminus V_2,-V_2}\right) < \mu\left(\mathcal{D}_{V\setminus V_1,-V_1}\right)$.*

Proof. Since $V_2 \cap S_j \subset S_j$, there exists a node v_p such that $v_p \in S_j$ and $v_p \notin V_2$. Similarly, since $\emptyset \subset V_2 \cap S_j$, there exists a node v_q such that $v_q \in S_j$ and $v_q \in V_2$. By Corollary 16, $\Pi^{=}_{V\setminus V_2,-V_2} \prec_r \Pi^{=}_{V\setminus V_1,-V_1}$ and thus the following implications hold:

- If $\cup_{t=1}^{\ell} V_t \not\subseteq V_2 \setminus V_1$ then $\Pi^{=}_{V\setminus V_2,-V_2}$ contains an equivalence class (subset of nodes) $S_{j'} \subseteq S_j$ such that $v_i \in S_{j'}$. This implies $\mu\left(\mathcal{D}_{V\setminus V_2,-V_2}\right) \leq \left|S_{j'}\right| \leq |S_j| = \mu\left(\mathcal{D}_{V\setminus V_1,-V_1}\right)$.
- If there exists a S_j such that $\emptyset \subset V_2 \cap S_j \subset S_j$ then $\Pi^{=}_{V\setminus V_2,-V_2}$ contains an equivalence class $\emptyset \subset S_{j'} \subset S_j$ with $v_p \in S_{t'}$. This implies $\mu\left(\mathcal{D}_{V\setminus V_2,-V_2}\right) \leq \left|S_{j'}\right| < |S_j| = \mu\left(\mathcal{D}_{V\setminus V_1,-V_1}\right)$.

❑

Based on the above structural properties, we design Algorithm I for ADIM$_{\geq k}$ as shown below.

Algorithm I: $O\left(n^4\right)$ time deterministic algorithm for ADIM$_{\geq k}$.

---

1. Compute $\mathbf{d}_i$ for all $i = 1, 2, \ldots, n$ in $O\left(n^3\right)$ time using Floyd-Warshall algorithm (27, p. 629)

2. $\widehat{\mathcal{L}^{\geq k}_{\text{opt}}} \leftarrow \infty$; $\widehat{V^{\geq k}_{\text{opt}}} \leftarrow \emptyset$

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

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

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

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

3.2.2 **if** $\Big(\left(\mu\left(\mathcal{D}_{V\setminus V',-V'}\right) \geq k\right)$ and $\left(|V'| < \widehat{\mathcal{L}_{\text{opt}}^{\geq k}}\right) \Big)$

3.2.3 **then** $\widehat{\mathcal{L}_{\text{opt}}^{\geq k}} \leftarrow |V'|$; $\widehat{V_{\text{opt}}^{\geq k}} \leftarrow V'$; done $\leftarrow$ TRUE

3.2.4 **else** let $V_1, V_2, \ldots, V_\ell$ be the *only* $\ell > 0$ equivalence classes (subsets of nodes) in $\Pi^{=}_{V\setminus V',-V'}$ such that $|V_1| = |V_2| = \cdots = |V_\ell| = \mu\left(\mathcal{D}_{V\setminus V',-V'}\right)$

3.2.5 $V' \leftarrow V' \cup \left(\cup_{t=1}^{\ell} V_t\right)$

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

---

Lemma 17 (Proof of correctness) *Algorithm I returns an optimal solution for* $\textsc{Adim}_{\geq k}$.

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

Let $\{v_\ell\} = V_1 \subset V_2 \subset \cdots \subset V_\kappa$ be the κ 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** $t \in \{1, 2, \ldots, \kappa\}$**.** Then, our solution is a set $\widehat{V_{\text{opt}}^{\geq k}}$ such that $\mu\left(\mathcal{D}_{V\setminus \widehat{V_{\text{opt}}^{\geq k}},-\widehat{V_{\text{opt}}^{\geq k}}}\right) \geq k$ and $\widehat{\mathcal{L}_{\text{opt}}^{\geq k}} \leq \mathcal{L}_{\text{opt}}^{\geq k}$.

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

Case 2.1: there exists $r \in \{1, 2, \ldots, \kappa - 1\}$ such that $V_r \subset V_{\text{opt}}^{\geq k}$ but $V_{r+1} \not\subseteq V_{\text{opt}}^{\geq k}$.

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 $\Pi^{=}_{V \setminus V_r, -V_r}$ such that $|V_{r,1}| = |V_{r,2}| = \cdots = |V_{r,p}| = \mu\left(\mathcal{D}_{V \setminus V_r, -V_r}\right)$. 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 $V_{r+1} \not\subseteq V_{\text{opt}}^{\geq k}$ implies $V_{r,1} \cup V_{r,2} \cup \cdots \cup V_{r,p} \not\subseteq V_{\text{opt}}^{\geq k}$, 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). Then, for some $\emptyset \subset Z'' \subseteq Z$, Z'' is an equivalence class in $\Pi^{=}_{V \setminus (V_r \cup Z'), -(V_r \cup Z')}$ implying

$$\mu\left(\mathcal{D}_{V \setminus (V_r \cup Z'), -(V_r \cup Z')}\right) \leq \left|Z''\right| \leq |Z| \tag{3.1}$$

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

$$\begin{aligned}
&\Pi^{=}_{V \setminus V_{\text{opt}}^{\geq k}, -V_{\text{opt}}^{\geq k}} \prec_r \Pi^{=}_{V \setminus (V_r \cup Z'), -(V_r \cup Z')} \quad \text{(in Corollary 16, set } V_2 = V_{\text{opt}}^{\geq k} \text{ and } V_1 = V_r \cup Z') \\
&\quad \Rightarrow\ k \leq \mu\left(\mathcal{D}_{V \setminus V_{\text{opt}}^{\geq k}, -V_{\text{opt}}^{\geq k}}\right) \\
&\qquad \leq \mu\left(\mathcal{D}_{V \setminus (V_r \cup Z'), -(V_r \cup Z')}\right) \\
&\qquad \leq |Z| \\
&\qquad \leq |V_{r,s}| \\
&\qquad = \mu\left(\mathcal{D}_{V \setminus V_r, -V_r}\right) \quad \text{by (Equation 3.1)}
\end{aligned}$$

Thus, $\mu\left(\mathcal{D}_{V\setminus V_r,-V_r}\right) \geq k$ and $|V_r| < \left|V_{\text{opt}}^{\geq k}\right| = \mathcal{L}_{\text{opt}}^{\geq k}$, contradicting the optimality of $\mathcal{L}_{\text{opt}}^{\geq k}$.

Case 2.2: $V_\kappa \subset V_{\text{opt}}^{\geq k}$**.** If done was set to TRUE at the last iteration of the **while** loop, then $\mu\left(\mathcal{D}_{V\setminus V_\kappa,-V_\kappa}\right) \geq k$ and $|V_\kappa| < \left|V_{\text{opt}}^{\geq k}\right| = \mathcal{L}_{\text{opt}}^{\geq k}$, contradicting the optimality of $\mathcal{L}_{\text{opt}}^{\geq k}$. Thus, done must have remained FALSE after the last iteration of the **while** loop, which implies $\mu\left(\mathcal{D}_{V\setminus V_\kappa,-V_\kappa}\right) < k$. Let $V_{\kappa,1}, V_{\kappa,2}, \ldots, V_{\kappa,p} \subseteq V \setminus V_\kappa$ be all the $p > 0$ equivalence classes (subsets of nodes) in $\Pi^{=}_{V\setminus V_\kappa,-V_\kappa}$ such that $|V_{\kappa,1}| = |V_{\kappa,2}| = \cdots = |V_{\kappa,p}| = \mu\left(\mathcal{D}_{V\setminus V_\kappa,-V_\kappa}\right)$. Since $V_\kappa \subset V_{\text{opt}}$, we have

$$\begin{array}{l} \Pi^{=}_{V\setminus V_{\text{opt}},-V_{\text{opt}}} \prec_r \Pi^{=}_{V\setminus V_\kappa,-V_\kappa} \\ \text{(in Corollary 16, set } V_2 = V_{\text{opt}} \text{ and } V_1 = V_\kappa) \\ \Rightarrow \quad k \leq \mu\left(\mathcal{D}_{V\setminus V_{\text{opt}},-V_{\text{opt}}}\right) \leq \mu\left(\mathcal{D}_{V\setminus V_\kappa,-V_\kappa}\right) \qquad \leq \qquad |Z| \leq |V_{\kappa,p}| = \mu\left(\mathcal{D}_{V\setminus V_\kappa,-V_\kappa}\right) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{by (Equation 3.1)} \end{array}$$

Thus, $\mu\left(\mathcal{D}_{V\setminus V_\kappa,-V_\kappa}\right) \geq k$ contradicting our assumption of $\mu\left(\mathcal{D}_{V\setminus V_\kappa,-V_\kappa}\right) < k$.

❑

Lemma 18 (Proof of time complexity) *Algorithm I runs in* $O\left(n^4\right)$ *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 $\Pi^{=}_{V\setminus\left(V'\cup\left(\cup_{t=1}^{\ell}V_t\right)\right),-V'\cup\left(\cup_{t=1}^{\ell}V_t\right)}$ from $\Pi^{=}_{V\setminus V',-V'}$. Suppose that $\cup_{t=1}^{\ell}V_t = \left\{v_{i_1}, v_{i_2}, \ldots, v_{i_p}\right\}$. By Corollary 16,

$$\Pi^{=}_{V\setminus(V'\cup\{v_{i_1},v_{i_2},\ldots,v_{i_{p-1}},v_{i_p}\}),-V'\cup\{v_{i_1},v_{i_2},\ldots,v_{i_{p-1}},v_{i_p}\}} \prec_r \Pi^{=}_{V\setminus(V'\cup\{v_{i_1},v_{i_2},\ldots,v_{i_{p-1}}\}),-V'\cup\{v_{i_1},v_{i_2},\ldots,v_{i_{p-1}}\}}$$
$$\prec_r \ldots \prec_r \Pi^{=}_{V\setminus(V'\cup\{v_{i_1},v_{i_2}\}),-V'\cup\{v_{i_1},v_{i_2}\}} \prec_r \Pi^{=}_{V\setminus(V'\cup\{v_{i_1}\}),-V'\cup\{v_{i_1}\}} \prec_r \Pi^{=}_{V\setminus V',-V'}$$

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* $\Pi^{=}_{V\setminus V_1,-V_1}$ *and a node* $v_j \in V \setminus V_1$*, compute* $\Pi^{=}_{V\setminus(V_1\cup\{v_j\}),-(V_1\cup\{v_j\})}$.

Since $\Pi^{=}_{V\setminus(V_1\cup\{v_j\}),-(V_1\cup\{v_j\})}$ is a refinement of $\Pi^{=}_{V\setminus V_1,-V_1}$ by Corollary 16, we can use the following simple strategy. For every set $S \in \Pi^{=}_{V\setminus V',-V'}$, 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 $\mathrm{dist}_{v_{i_1},v_j}, \mathrm{dist}_{v_{i_2},v_j}, \ldots, \mathrm{dist}_{v_{i_s},v_j}$,. The total time taken for all sets in $\Pi^{=}_{V\setminus V',-V'}$ is thus $\sum_{S\in\Pi^{=}_{V\setminus V',-V'}} O\left(n\,|S|\right) = O\left(n^2\right)$. ❑

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

Algorithm II: $O\left(n^4\right)$ time deterministic algorithm for ADIM

(changes from Algorithm-I are shown enclosed in ▭)

---

1. Compute $\mathbf{d}_i$ for all $i = 1, 2, \ldots, n$ in $O\left(n^3\right)$ time using Floyd-Warshall algorithm (27, p. 629)

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

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** $\boxed{\left(V \setminus V' \neq \emptyset\right)}$ **do**

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

3.2.2 **if** $\boxed{\left(\mu\left(\mathcal{D}_{V \setminus V', -V'}\right) > \widehat{k_{\text{opt}}}\right)}$

3.2.3 **then** $\boxed{\widehat{k_{\text{opt}}} \leftarrow \mu\left(\mathcal{D}_{V \setminus V', -V'}\right)}$; $\widehat{V_{\text{opt}}^{\geq k}} \leftarrow V'$

3.2.4 **else** let $V_1, V_2, \ldots, V_\ell$ be the *only* $\ell > 0$ equivalence classes (subsets of nodes) in $\Pi_{V \setminus V', -V'}^{=}$ such that $|V_1| = |V_2| = \cdots = |V_\ell| = \mu\left(\mathcal{D}_{V \setminus V', -V'}\right)$

3.2.5 $V' \leftarrow V' \cup \left(\cup_{t=1}^{\ell} V_t\right)$

4. **return** $\boxed{\widehat{k_{\text{opt}}}}$ and $\widehat{V_{\text{opt}}^{\geq k}}$ as our solution

---

The proof of correctness is very similar (and, in fact simpler due to elimination of some cases) to that of ADIM$_{\geq k}$.

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

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

---

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

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

or

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

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 ADIM$_{\geq k}$)

or

execute Step 3.1 and Step 3.2 (and its sub-steps) of Algorithm II (for ADIM)

4. **return** the best of all solutions found in Step 3

The success probability p is given by

$$p = \Pr\left[v_i \in V_{\text{opt}}^{\geq k} \text{ in } \textit{at least one} \text{ of the } \lceil \tfrac{2n \ln n}{k} \rceil \text{ iterations}\right]$$

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

3.4 Proof of Theorem 14

(a) ADIM$_{=k}$ trivially belongs to NP for any k, thus we need to show that it is also NP-hard.

The standard NP-complete *minimum dominating set* (MDS) problem for a graph is defined as follows (29). Our input is a connected undirected unweighted graph $G = (V, E)$. A subset of nodes $V' \subset V$ 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 $\mathcal{U} = \{a_1, a_2, \ldots, a_n\}$ and $S_1, S_2, \ldots, S_m \subseteq \mathcal{U}$ 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 $a_i \in \mathcal{U}$ 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_ℓ} 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 $\mathcal{I} \subset \{1, 2, \ldots, m\}$ is a solution of SC if and only if the collection of set nodes $\{ v_{S_i} \mid i \in \mathcal{I} \}$ is a solution of MDS on G_1 and thus $\mathsf{opt}_{\mathrm{SC}} = \nu(G_1)$.

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. Obviously we need at least $\frac{n}{3}$ sets to cover all the n elements. Letting $\mathsf{opt}_{\mathrm{X3C}}$ to denote the number of sets in an optimal solution of X3C, it is well-known that problem of deciding whether $\mathsf{opt}_{\mathrm{X3C}} = \frac{n}{3}$ is in fact NP-complete.

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".

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

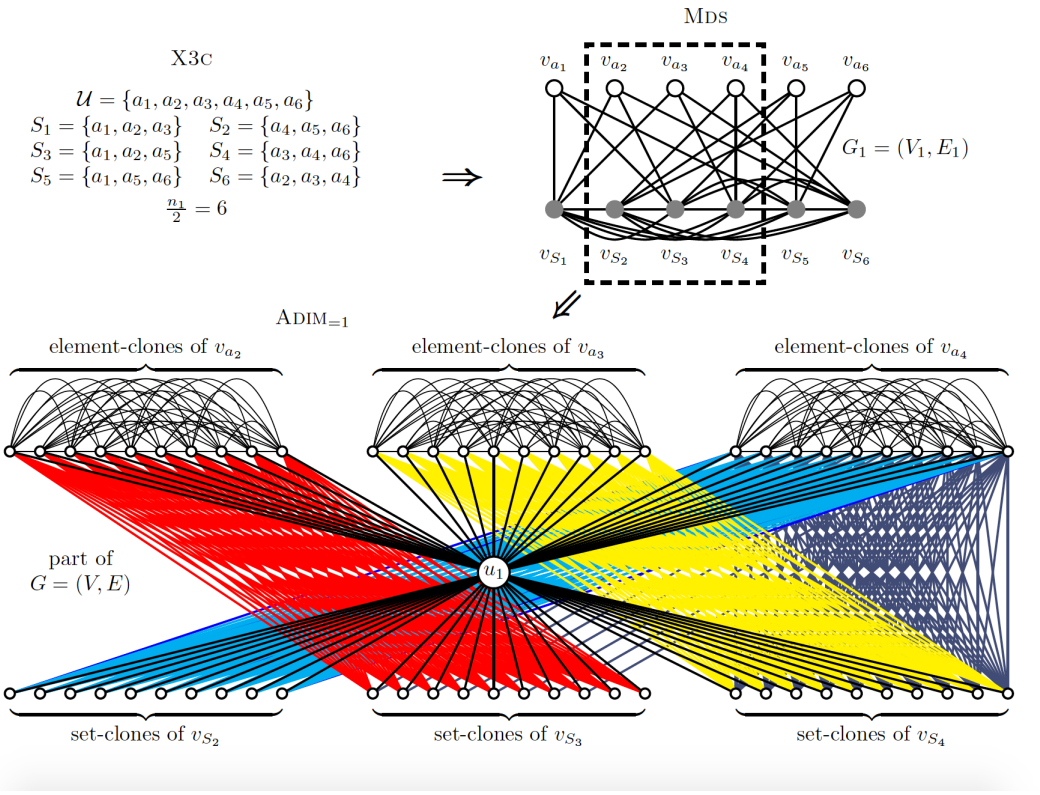


Figure 15. Illustration of the NP-hardness reduction in Theorem 14(a). Only a part of the graph G is shown for visual clarity.

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. Figure 15 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

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

  We refer to these edges as the *partition-fixing* edges. There are precisely $kn_1\left(k+\frac{n_1}{3}\right)$ 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 $\left(2k + \frac{2n_1}{3}\right)^2$ edges:

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

  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 $\frac{3n_1}{2}\left(2k + \frac{2n_1}{3}\right)^2$ such edges.

Note that G has precisely $n_1\left(2k + \frac{2n_1}{3}\right) + k = n$ 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 $\mathsf{diam}(G) = 2$. We now show the validity of our reduction by showing that

$$(\star)\ \nu\left(G_1\right) = \frac{n_1}{3} \ \text{ if and only if } \ \mathcal{L}_{\text{opt}}^{=k} \leq \frac{n_1}{3}$$

Proof of $\nu\left(G_1\right) = \frac{n_1}{3} \ \Rightarrow \ \mathcal{L}_{\text{opt}}^{=k} \leq \frac{n_1}{3}$

Consider an optimal solution $V_1' \subset \left\{v_{S_1}, v_{S_2}, \ldots, v_{S_{n_1}}\right\}$ of MDS on G_1 with $\nu\left(G_1\right) = |V_1'| = \frac{n_1}{3}$. We now construct a solution $V' \subset V$ of ADIM$_{=k}$ on G by setting $V' = \left\{v_{S_j,1} \,|\, v_{S_j} \in V_1'\right\}$. Note that $|V'| = |V_1'| = \frac{n_1}{3}$. We claim that V' is a valid solution of ADIM$_{=k}$ by showing that

(*a*) $\{u_1, u_2, \ldots, u_k\} \in \Pi^{=}_{V \setminus V', -V'}$ and

(*b*) any other equivalence class in $\Pi^{=}_{V \setminus V', -V'}$ 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 ℓ. Since V_1' is a solution of 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 $1 = \text{dist}_{u_i,v_{a_j,\ell}} \neq \text{dist}_{v_{S_p,1},v_{a_j,\ell}}$, and therefore $v_{a_j,\ell}$ *cannot* be in the same equivalence class with u_i.

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

To prove (b), note the following:

- Since $\mathsf{diam}(G) = 2$, $\text{dist}_{v_{S_i,p},v_{S_j,q}} = 2$ 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 $\Pi^{=}_{V\setminus V',-V'}$. There are at least $n_1\left(k + \frac{n_1}{3}\right) - \frac{n_1}{3} > k$ 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 $\Pi^{=}_{V\setminus V',-V'}$ 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 $\mathsf{diam}(G) = 2$, 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 $\mathcal{L}_{\text{opt}}^{=k} \leq \frac{n_1}{3} \Rightarrow \nu(G_1) = \frac{n_1}{3}$

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

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

Proof. Suppose that $\widehat{V}$ contains at least one element-clone node $v_{a_i,j}$ for some i and j. But, $V \setminus V_{\text{opt}}^{=k}$ 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 $|\widehat{V}| \geq 2k + \frac{2n_1}{3} - \frac{n_1}{3} > k$, a contradiction.

Similarly, suppose that $\widehat{V}$ contains at least one set-clone node $v_{S_i,j}$ for some i and j. But, $V \setminus V_{\text{opt}}^{=k}$ contains at least $2k + \frac{2n_1}{3} - \frac{n_1}{3} - 1 > k$ 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 $|\widehat{V}| \geq 2k + \frac{2n_1}{3} - \frac{n_1}{3} > k$, a contradiction. ❑

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

Proof. Suppose that $V_{\text{opt}}^{=k}$ contains two set-clone nodes $v_{S_j,p}$ and $v_{S_j,q}$ of the same set node v_{S_j}. But, $V \setminus V_{\text{opt}}^{=k}$ 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 in the *same* equivalence class S. If we remove

$v_{S_j,p}$ from $V_{\text{opt}}^{=k}$ 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 $\mathcal{L}_{\text{opt}}$, contradicting the optimality of $\mathcal{L}_{\text{opt}}^{=k}$. ❑

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

Proof. Suppose that $V_{\text{opt}}^{=k}$ contains at least one element-clone node and thus at most $\frac{n_1}{3} - 1$ set-clone nodes. Note that $V \setminus V_{\text{opt}}^{=k}$ contains at least $2k + \frac{2n_1}{3} - \frac{n_1}{3}$ clones of every element node a_i. Consider an element-clone node $v_{a_i,p} \in V \setminus V_{\text{opt}}^{=k}$ and a clique node u_j. Since $\widehat{V} = \{u_1, u_2, \ldots, u_k\} \in \Pi^{=}_{V \setminus V_{\text{opt}}^{=k}, -V_{\text{opt}}^{=k}}$, there must be a node in $V_{\text{opt}}^{=k}$ such that the distance of this node to u_j is different from the distance to $v_{a_i,p}$. Such a node in $V_{\text{opt}}^{=k}$ cannot be an element-clone node, say $v_{a_\ell,q}$ since $\text{dist}_{v_{a_i,p},v_{a_\ell,q}} = \text{dist}_{u_j,v_{a_\ell,q}} = 1$. 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 $\text{dist}_{v_{a_i,p},v_{S_r,s}} = 2$, *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 then exist an element-clone node $v_{a_i,p}$ such that the distance of $v_{a_i,p}$ to any node in $V_{\text{opt}}^{=k}$ is exactly the same as the distance of u_j to that node in $V_{\text{opt}}^{=k}$. This implies $v_{a_i,p} \in \widehat{V}$, contradicting Proposition 3. ❑

By Proposition 4 and Proposition 5, $V_{\text{opt}}^{=k}$ contains exactly one clone of a subset of set nodes. Without loss of generality, assume that $V_{\text{opt}}^{=k} = \left\{ v_{S_j,1} \,|\, j \in J,\, J \subset \left\{1, 2, \ldots, \frac{n_1}{2}\right\} \right\}$ and let $V_1' = \left\{ v_{S_j} \,|\, v_{S_j,1} \in V_{\text{opt}}^{=k} \right\}$. Note that $|V_1'| = \left|V_{\text{opt}}^{=k}\right|$. We are now ready to finish our proof

by showing V_1' is indeed a valid solution of MDS on G_1. Suppose not, and let v_{a_i} be an element-node that is not adjacent to any node in V_1'. Then,

$$\forall\, v_{S_j} \in V_1' \,:\, \{v_{a_i}, v_{S_j}\} \notin E_1 \;\Rightarrow\; \forall\, v_{S_j,1} \in V_{\text{opt}}^{=k} \,:\, \{v_{a_i,1}, v_{S_j,1}\} \in E$$
$$\Rightarrow\; \forall\, v_{S_j,1} \in V_{\text{opt}}^{=k} \,:\, \text{dist}_{v_{a_i,1}, v_{S_j,1}} = 1 \;\Rightarrow\; v_{a_i,1} \in \widehat{V}$$

which contradicts Proposition 3.

(b) The proof is similar to that of (a) but this time we start with a general version of SC as opposed to the restricted X3C version, and show that the reduction is approximation-preserving in an appropriate sense. In the sequel, we use the standard notation 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 (30). Given an instance formula ϕ of the standard Boolean satisfiability problem (SAT), Feige reduces ϕ to an instance $\mathcal{U}, S_1, S_2, \ldots, S_m$ of 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 $\mathsf{opt}_{\text{SC}} = \frac{n}{Q}$ or $\mathsf{opt}_{\text{SC}} > \left(\frac{n}{Q}\right)(1-\varepsilon)\ln n$.
- The reduction satisfies the following completeness and soundness properties:

  (completeness) If ϕ is satisfiable then $\mathsf{opt}_{\text{SC}} = \frac{n}{Q}$.

  (soundness) If ϕ is not satisfiable then $\mathsf{opt}_{\text{SC}} > \left(\frac{n}{Q}\right)(1-\varepsilon)\ln n$.

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 SC to MDS to ADIM$_{=k}$ 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 (30) 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 $k \leq n^\varepsilon$ 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 $\delta = \frac{\ln n}{(1-\varepsilon)\ln n_1}$. Our claim ($\star$) in the proof of (a) is now modified to

$$(\star) \qquad \begin{array}{ll} \text{(completeness)} & \text{if } \nu(G_1) = \frac{n_1}{Q} \text{ then } \mathcal{L}^{=k}_{\text{opt}} \leq \frac{n_1}{Q} \\ \text{(soundness)} & \text{if } \nu(G_1) > \left(\frac{n_1}{Q}\right)(1-\varepsilon)\ln n_1 \\ & \text{then } \mathcal{L}^{=k}_{\text{opt}} > \left(\frac{n_1}{Q}\right)(1-\varepsilon)\ln n_1 = \left(\frac{n_1}{Q}\right)\frac{1}{\delta}\ln n \end{array}$$

- Our proof of the *completeness* claim follows the "Proof of $\nu(G_1) = \frac{n_1}{3} \Rightarrow \mathcal{L}^{=k}_{\text{opt}} \leq \frac{n_1}{3}$" 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

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

and the proof of this contra-positive follows the "Proof of $\mathcal{L}_{\text{opt}}^{=k} \leq \frac{n_1}{3} \Rightarrow \nu(G_1) = \frac{n_1}{3}$" 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, $\Pi^{=}_{V \setminus V_{\text{opt}}^{=k}, -V_{\text{opt}}^{=k}}$ 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 $\Pi^{=}_{V \setminus V', -V'}$ 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 $\mathcal{L}_{\text{opt}}^{=k} = c$ if a solution exists.

3.5 Proof of Theorem 15

(a) Note that trivially $\mathcal{L}_{\text{opt}}^{=1} \leq n - 1$ and thus $V_{\text{opt}}^{=1} \neq \emptyset$. Our algorithm, shown as Algorithm V, uses the greedy logarithmic approximation of Johnson (31) for SC that selects, at each successive step, a set that contains the maximum number of elements that are still not covered.

Algorithm V: $O\left(n^3\right)$-time $(1 + \ln(n-1))$-approximation algorithm for ADIM$_{=1}$.

---

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

2. $\widehat{\mathcal{L}_{\text{opt}}^{=1}} \leftarrow \infty$; $\widehat{V_{\text{opt}}^{=1}} \leftarrow \emptyset$

3. **for** each node $v_i \in V$ **do** ($*$ we *guess* the set $\{v_i\}$ to belong to $\Pi^{=}_{V \setminus V_{\text{opt}}^{=1}, -V_{\text{opt}}^{=1}}$ $*$)

3.1 create the following instance of SC containing $n-1$ elements and $n-1$ sets:

$$\mathcal{U} = \left\{ a_{v_j} \,|\, v_j \in V \setminus \{v_i\} \right\},$$

$$S_{v_j} = \left\{a_{v_j}\right\} \cup \left\{ a_{v_\ell} \,|\, \text{dist}_{v_i, v_j} \neq \text{dist}_{v_\ell, v_j} \right\} \text{ for } j \in \{1, 2, \ldots, n\} \setminus \{i\}$$

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

3.2.1 run the greedy approximation algorithm (31) for this instance of SC giving a solution $\mathcal{I} \subseteq \{1,2,\ldots,n\} \setminus \{i\}$

3.2.2 $V' = \{\, v_j \,|\, j \in \mathcal{I} \,\}$

3.2.3 **if** $\left(\, |V'| < \widehat{\mathcal{L}^{=1}_{\text{opt}}} \,\right)$ **then** $\widehat{\mathcal{L}^{=1}_{\text{opt}}} \leftarrow |V'|$; $\widehat{V^{=1}_{\text{opt}}} \leftarrow V'$

4. **return** $\widehat{\mathcal{L}^{=1}_{\text{opt}}}$ and $\widehat{V^{=1}_{\text{opt}}}$ as our solution

---

Lemma 19 (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_ℓ for some $v_\ell \in V$. We claim that this cannot be the case since $\{v_\ell\} \in \Pi^{=}_{V\setminus V', -V'}$. Indeed, since $\mathcal{I}$ is a valid solution of the SC instance, for every $j \notin \{\ell\} \cup \mathcal{I}$, the following holds:

$$\exists\, t \in \mathcal{I} \,:\, a_{v_j} \in S_{v_t} \;\Rightarrow\; \exists\, v_t \in V' \,:\, \text{dist}_{v_\ell, v_t} \neq \text{dist}_{v_j, v_t}$$

and thus v_ℓ cannot be together with any other node in any equivalence class in $\Pi^{=}_{V\setminus V', -V'}$. ❑

Lemma 20 (Proof of approximation bound) *Algorithm V solves* ADIM$_{=1}$ *with an approximation ratio of* $1 + \ln(n-1)$.

Proof. Fix any optimal solution $V^{=1}_{\text{opt}}$. Since $\mu\left(\mathcal{D}_{V\setminus V^{=1}_{\text{opt}}, -V^{=1}_{\text{opt}}}\right) = 1$, $\{v_\ell\} \in \Pi^{=}_{V\setminus V^{=1}_{\text{opt}}, -V^{=1}_{\text{opt}}}$ for some $v_\ell \in V$. Consider the iteration of the **for** loop in Step 3 when v_i is equal to v_ℓ. We now analyze the run of *this particular iteration*, and claim that the set-cover instance created during

this iteration satisfies $\mathsf{opt}_{\text{SC}} \leq \left|V_{\text{opt}}^{=1}\right| = \mathcal{L}_{\text{opt}}^{=1}$. To see this, construct the following solution of the set-cover instance from V_{opt} containing exactly $\mathcal{L}_{\text{opt}}$ sets:

$$v_i \in V_{\text{opt}}^{=1} \equiv i \in \mathcal{I}$$

To see that this is indeed a valid solution of the set-cover instance, consider any $a_{v_j} \in \mathcal{U} = \{a_{v_1}, a_{v_2}, \ldots, a_{v_n}\} \setminus \{a_{v_\ell}\}$. Then, the following cases apply showing that a_{v_j} belongs to some set selected in our solution of SC:

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

Using the approximation bound of the algorithm of (31) it now follows that the quality of our solution $\widehat{\mathcal{L}_{\text{opt}}^{=1}}$ satisfies

$$\widehat{\mathcal{L}_{\text{opt}}^{=1}} = \left|\widehat{V_{\text{opt}}^{=1}}\right| = |\mathcal{I}| < (1 + \ln(n-1))\mathsf{opt}_{\text{SC}} \leq (1 + \ln(n-1))\mathcal{L}_{\text{opt}}^{=1}$$

❑

Lemma 21 (Proof of time complexity) *Algorithm V runs in* $O\left(n^3\right)$ *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 trivially done in $O\left(n^2\right)$ time by comparing 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 $O\left(n^2\right)$ time. ❑

(b) Let v_i be the node of degree 1. Let v_ℓ be the unique node adjacent to v_i (*i.e.* , $\{v_i, v_\ell\} \in E$). Consider the following solution of ADIM$_{=1}$: $V' = \{v_i\}$. We claim that is a valid solution of 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, $1 = \text{dist}_{v_\ell, v_i} \neq \text{dist}_{v_j, v_i}$.

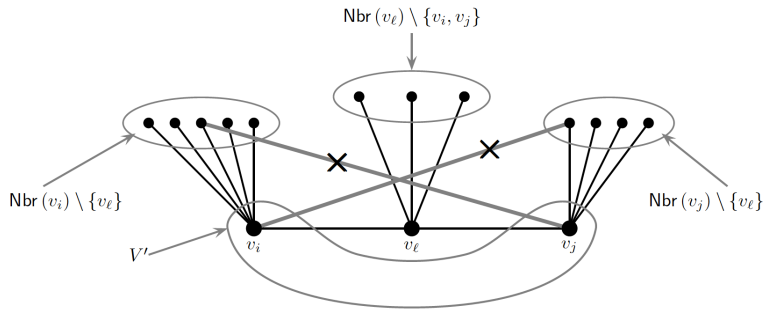


Figure 16. Illustration of the proof of Theorem 15(c). Edges marked by × cannot exist. No node in Nbr $(v_\ell) \setminus \{v_i, v_j\}$ can have an edge to *both* v_i and v_j.

(c) Since G does not contain a 4-cycle, 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_ℓ 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. Figure 16). Consider the following solution of ADIM$_{=1}$:

$V' = \{v_i, v_j\}$. Note that $v_\ell \in V \setminus V'$. We claim that this is a valid solution of $\textsc{Adim}_{=1}$ 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_ℓ in the same equivalence class of $\Pi^{=}_{V\setminus V', -V'}$) in the following manner:

- If $v_p \in \mathsf{Nbr}(v_i) \setminus \{v_\ell\}$ then $\text{dist}_{v_\ell, v_j} = 1$ but $\text{dist}_{v_p, v_j} \neq 1$ since G has no 4-cycle (see the edges marked $\times$ in Fig. Figure 16).
- If $v_p \in \mathsf{Nbr}(v_j) \setminus \{v_\ell\}$ then $\text{dist}_{v_\ell, v_i} = 1$ but $\text{dist}_{v_p, v_i} \neq 1$ since G has no 4-cycle (see the edges marked $\times$ in Fig. Figure 16).
- If $v_p \in \mathsf{Nbr}(v_\ell) \setminus \{v_i, v_j\}$ then v_p cannot be adjacent to *both* v_i and v_j since G does not contain a 4-cycle. This implies that $\text{dist}_{v_\ell, v_i} = \text{dist}_{v_\ell, v_j} = 1$ but at least one of dist_{v_p, v_i} and dist_{v_p, v_j} is not equal to 1.
- If v_p is any node not covered by the above cases, then $\text{dist}_{v_p, v_i} > 1$ but $\text{dist}_{v_\ell, v_i} = 1$.

CHAPTER 4

CONCLUSION

In this thesis we have examined various techniques for quantifying loss of privacy in networked and multi-agent systems. In the first part of the thesis we investigated approximate privacy model. We identified a protocol that provides constant *average* privacy approximation ratio for *tiling functions*. We also provided calculations of average and worst case privacy approxixation ratio of bisection protocols for non-tiling functions. There are some natural research problems for the geometric privacy model discussed that needs further investigation. Examples of such questions include:

- Can we identify other non-tiling classes of functions for which good approximate-privacy preserving protocols are possible?
- For three or more party communications, can we design good approximate-private protocols for some proper sub-classes of tiling functions?
- Can we identify and formalize the relationship between approximate privacy, differential privacy (4) and pan-privacy (32) models?

In the second part of the thesis, we formalized problems concerning a privacy measure for quantifying privacy in large networks. Prior to our work, known results for these privacy measures only included some heuristic algorithms with no provable guarantee on performances such as in (14), or algorithms for very special cases. In fact, it was not even known if any version of

these related computational problems is NP-hard. Our work provides the first non-trivial computational complexity results for effective computation of these measures. Theorem 13 shows that both ADIM and $\text{ADIM}_{\geq k}$ are *provably* computationally easier problems than $\text{ADIM}_{=k}$. In contrast, Theorem 14(a)–(b) and Theorem 15 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 conjecture that a $O(\log n)$-approximation is possible for $\text{ADIM}_{=k}$ for every fixed k.
- 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 ? We conjecture that the inapproximability factor can be further improved to $\Omega\left(n^{\varepsilon}\right)$ for some constant $0 < \varepsilon < 1$ when k is around $\sqrt{n}$.

APPENDIX

ADDITIONAL PRIVACY MEASURES

A.1 Differential Privacy

The differential privacy model was introduced by C. Dwork in (4). This model arose due to difficulties in preserving privacy in statistical databases. Consider a statistical database that contains information obtained from a survey of some population. The differntial privacy model allows a third party to learn the properties of the population but at the same time preserves the privacy of individuals that participated in the survey.

Definition 22 *(4) A randomized function $\mathcal{K}$ gives ε-differential privacy if for all data sets D_1 and D_2 differing on at most one row, and all $S \subseteq Range(\mathcal{K})$,*

$$\Pr[\mathcal{K}(D_1) \in S] \leq \mathrm{e}^{\varepsilon} \times \Pr[\mathcal{K}(D_2) \in S]$$

Differential privacy can be achieved by computing the correct answer to a query and adding a noise drawn from the so-called $\mathcal{L}aplace(f(\varepsilon))$ distribution for some appropriate function f. This approach is sufficient to handle individual queries. In (4) the author also provides a mechanism for ensuring differential privacy in case of adaptive queries.

McGregor *et al.* (33) introduced and investigated the differential privacy model in a 2-party communication setting. In such a setting the two parties want to find out the *hamming distance* between the n bit inputs that they hold. This setting is defined in (33) as give below:

- A *mechanism* M (on Σ^n) is a family of probability distributions $\{\mu_x : x \in \Sigma^n\}$ on $\mathbb{R}$. Such a mechanism M is ε-differentially private if and only if the following condition holds:

$$\forall\, x, x \in \Sigma^n : \left|x - x\right|_H = 1 \quad \text{and} \quad \text{for all measurable subsets } S \text{ of } \mathbb{R} \;:\; \mu_x(S) \leq \mathrm{e}^{\varepsilon}\mu_x(S)$$

  where the notation $\left|x - x\right|_H$ denotes the Hamming distance between x and x'.

- VIEW $\overset{A}{P}(x, y)$ is the joint probability distribution over $x, y \in \Sigma^n$, the transcript of a protocol P and the private randomness of party A (the probability space is private randomness for both parties). VIEW $\overset{B}{P}(x, y)$ is defined in a similar manner with respect to party B.

Then, a protocol P has ε-differential privacy if and only if both of the following conditions hold:

(a) For all $x \in \Sigma^n$, VIEW $\overset{A}{P}(x, y)$ is ε-differential private.

(b) For all $y \in \Sigma^n$, VIEW $\overset{B}{P}(x, y)$ is ε-differential private.

A major contrition of (33) is a *lower bound* on the *least additive error* of any differentially private protocol that is used to compute the hamming distance.

Theorem 23 *(33) Let $P(x, y)$ be a randomized protocol with ε-differential privacy for inputs $x, y \in \{0, 1\}^n$, and let $\delta > 0$. Then, with probability at least $1 - \delta$ over $x, y \in \{0, 1\}^n$ and the coin tosses of P, party B's output differs from $\langle x, y\rangle$ by at least $\Delta = \Omega\left(\dfrac{\sqrt{n}}{\log n} \times \dfrac{\delta}{e^{\varepsilon}}\right)$.*

In (33) the authors state that there is a connection between differential privacy model and pan-privacy model defined in (32).

CITED LITERATURE

1. J. Feigenbaum, A. Jaggard and M. Schapira: *Approximate privacy: Foundations and quantification.* ACM Conference on Electronic Commerce, pages 167–178, 2010.

2. M. Comi, B. DasGupta, M. Schapira and V. Srinivasan: *On communication protocols that compute almost privately.* Theoretical Computer Science, 457:45–58, 2012.

3. M. Comi, B. DasGupta, M. Schapira and V. Srinivasan: *On communication protocols that compute almost privately.* Proceedings of the 4th international conference on Algorithmic game theory, pages 44–56, 2011.

4. C. Dwork: *Differential privacy.* Proc. 33rd International Colloquium on Automata, Languages and Programming, pages 1–12, July 2006.

5. P. Samarati and L. Sweeney.: *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.* Technical report, Technical report, 1998.

6. K. Liu and E. Terzi.: *Towards identity anonymization on graphs.* In Proc. 2008 ACM SIGMOD International Conference on Management of Data, pages 93–106, New York, NY, USA, 2008.

7. B. Zhou, J. Pei and W. S. Luk.: *A brief survey on anonymization techniques for privacy preserving publishing of social network data.* SIGKDD Explorations Newsletter, 10(2):12–22, 2008.

8. A. Narayanan and V. Shmatikov.: *De-anonymizing social networks.* 30th IEEE Symposium on Security and Privacy, pages 173–187, 2009.

9. L. Zou, L. Chen and M. T. Özsu: *K-automorphism: A general framework for privacy preserving network publication.* Proc. VLDB Endowment, 2(1):946–957, 2009.

10. L. Backstrom, C. Dwork and J. Kleinberg: *Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography.* Proc. 16th International Conference on World Wide Web, pages 181–190, 2007.

11. B. Viswanath et al: *Canal: Scaling social network-based sybil tolerance schemes.* In Proc. 7th ACM European Conference on Computer Systems, pages 309–322, New York, NY, USA, 2012.

12. M. Netter, S. Herbst and G. Pernul.: *Analyzing privacy in social networks–an interdisciplinary approach.* IEEE 3rd International Conference on Privacy, Security, Risk and Trust and IEEE 3rd International Conference on Social Computing, pages 1327–1334, 2011.

13. X. Wu, X. Ying, K. Liu and L. Chen.: *A survey of privacy-preservation of graphs and social networks.* In Managing and Mining Graph Data, eds. C. C. Aggarwal and H. Wang, volume 40 of Advances in Database Systems, pages 421–453. Springer, 2010.

14. R. Trujillo-Rasua and I. G. Yero.: *k-metric antidimension: A privacy measure for social graphs.* Information Sciences, 328:403–417, 2016.

15. T. Feder, S. U. Nabar and E. Terzi: *Anonymizing graphs.* CoRR, abs/0810.5578., 2008.

16. B. Chor and E. Kushilevitz: *A zero-one law for boolean privacy.* SIAM Journal of Discrete Mathematics, 4:36–47, 1991.

17. E. Kushilevitz: *Privacy and communication complexity.* SIAM Journal of Discrete Mathematics, 5(2):273–284, 1992.

18. R. Bar-Yehuda, B. Chor, E. Kushilevitz and A. Orlitsky: *Privacy, additional information, and communication.* IEEE Transactions on Information Theory, 39:55–65, 1993.

19. E. Grigorievaa, P. J.-J. Heringsb, R. Müllera and D. Vermeulena: *The communication complexity of private value single-item auctions.* Operations Research Letters, 34:491–498, 2006.

20. E. Grigorievaa, P. J.-J. Heringsb, R. Müllera and D. Vermeulena: *The private value single item bisection auction.* Economic Theory, 30:107–118, 2007.

21. P. Berman, B. DasGupta and S. Muthukrishnan: *On the exact size of the binary space partitioning of sets of isothetic rectangles with applications.* SIAM Journal of Discrete Mathematics, 15(2):252–267, 2002.

22. F. d'Amore and P. G. Franciosa: *On the optimal binary plane partition for sets of isothetic rectangles.* Information Processing Letters, 44:255–259, 1992.

23. M. Paterson and F. F. Yao: *Efficient binary space partitions for hidden-surface removal and solid modeling.* Discrete and Computational Geometry, 5(1):485–503, 1990.

24. M. Paterson and F. F. Yao: *Optimal binary space partitions for othogonal objects.* Journal of Algorithms, 13:99–113, 1992.

25. A. Dumitrescu, J. S. B. Mitchell and M. Sharir: *Binary space partitions for axis-parallel segments, rectangles, and hyperrectangles.* Discrete & Computational Geometry, 31(2):207–227, 2004.

26. A. C. Yao.: *Some complexity questions related to distributive computing (preliminary report).* Proc. 11th ACM Symposium on Theory of Computing, pages 209–213, 1979.

27. T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein: Introduction to Algorithms, 2nd edition. The MIT Press, 2001.

28. V. Vazirani.: Approximation Algorithms. Springer-Verlag, 2001.

29. M. R. Garey and D. S. Johnson: Computers and Intractability - A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., 1979.

30. U. Feige: *A threshold for approximating set cover.* Journal of the ACM, 45:634–652, 1998.

31. D. S. Johnson: *Approximation algorithms for combinatorial problems.* Journal of Computer and System Sciences, 9:256–278, 1974.

32. C. Dwork, M. Naor, T. Pitassi, G. N. Rothblum and S. Yekhanin: *Pan-private streaming algorithms.* In Proceedings of ICS, 2010.

33. A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. Vadhan: *The limits of two-party differential privacy.* Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium, pages 81–90, 2010.

34. R. Trujillo-Rasua and I. G. Yero.: *Characterizing* 1*-metric antidimensional trees and unicyclic graphs.* The Computer Journal, 59(8):1264, 2016.

35. S. Mauw, R. Trujillo-Rasua and B. Xuan.: *Counteracting active attacks in social network graphs.* In 30th Annual IFIP WG 11.3 Working Conference on Data and Applications Security and Privacy, Trento, Italy, 2016.

36. C. Dwork: *A firm foundation for private data analysis.* Communications of the ACM, 54(1):86–95, 2010.

37. C. Dwork: *Differential privacy: A survey of results.* In Proc. 5th Intl Conf. on Theory and Applic. of Models of Comp., 2008.

VITA

VENKATAKUMAR SRINIVASAN

EDUCATION

PhD, Computer Science, ***University of Illinois at Chicago, Chicago, IL***, Dec 2016

B.E., Computer Science & Engineering, ***Bharathiyar University, Tamil Nadu, India***, April 2002

PUBLICATIONS

Marco Comi, Bhaskar DasGupta, Michael Schapira and **Venkatakumar Srinivasan**, *On Communication Protocols that Compute Almost Privately*, Theoretical Computer Science, 457, 45-58, 2012.

Bhaskar DasGupta and **Venkatakumar Srinivasan**, *A review of some approximate privacy measures of multi-agent communication protocols*, Frontiers of Intelligent Control and Information Processing, Derong Liu, Cesare Alippi, Dongbin Zhao, and Huaguang Zhang (editors), Chapter 10, 267-283, World Scientific Publishing, 2014.

Tanima Chatterjee, Bhaskar DasGupta, Nasim Mobasheri, **Venkatakumar Srinivasan** and Ismael G. Yero, *On the Computational Complexities of Three Privacy Measures for Large Networks Under Active Attack*, arXiv:1510.08779 [cs.CC]

Bhaskar DasGupta and **Venkatakumar Srinivasan**, *A Review of Several Optimization Problems Related to Security in Networked System*, to appear in Operations Research, Engineering, and Cyber Security : Trends in Applied Mathematics and Technology, N. J. Daras and Th. M. Rassias (editors), Springer Optimization and Its Applications series, Springer.

EXPERIENCE

Software Engineering Intern, ***Xaptum****, Chicago, Summer 2013, Summer 2014, Summer 2015, Summer 2016*

Software Engineering Research Intern, ***Motorola Mobility****, Chicago, May 2012 - July 2012*

Software Contractor, ***Motorola Solutions*** *&* ***Motorola Mobility****, Chicago, May 2010 - August 2011*

Software Consultant, ***Novell****, India, March 2007 - July 2008*

Senior Applications Engineer, ***Oracle****, India, August 2003 - March 2007*

Programmer Analyst, ***Cognizant Technology Solutions****, India, August 2002 - August 2003*

SPRINGER LICENSE
TERMS AND CONDITIONS

Apr 08, 2017

This Agreement between Venkatakumar Srinivasan ("You") and Springer ("Springer") consists of your license details and the terms and conditions provided by Springer and Copyright Clearance Center.

License Number	4084051500867
License date	Apr 08, 2017
Licensed Content Publisher	Springer
Licensed Content Publication	Springer eBook
Licensed Content Title	On Communication Protocols That Compute Almost Privately
Licensed Content Author	Marco Comi
Licensed Content Date	Jan 1, 2011
Type of Use	Book/Textbook
Requestor type	Publisher
Publisher	UIC Indigo
Portion	Full text
Format	Print and Electronic
Will you be translating?	No
Print run	25000
Author of this Springer article	Yes and you are the sole author of the new work
Order reference number	
Title of new book	Analysis of Privacy Measures for Multi-Agent and Networked Systems
Publisher	UIC Indigo
Author of new book	Venkatakumar Srinivasan
Expected publication date of new book	Apr 2017
Estimated size of new book (pages)	100
Requestor Location	Venkatakumar Srinivasan 1926 W Harrison Street Apt 902 CHICAGO, IL 60612 United States Attn: Venkatakumar Srinivasan
Billing Type	Invoice
Billing Address	Venkatakumar Srinivasan ,

United States
Attn: Venkatakumar Srinivasan

Total 0.00 USD

Terms and Conditions

Introduction
The publisher for this copyrighted material is Springer. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).
Limited License
With reference to your request to reuse material on which Springer controls the copyright, permission is granted for the use indicated in your enquiry under the following conditions:
- Licenses are for one-time use only with a maximum distribution equal to the number stated in your request.
- Springer material represents original material which does not carry references to other sources. If the material in question appears with a credit to another source, this permission is not valid and authorization has to be obtained from the original copyright holder.
- This permission
• is non-exclusive
• is only valid if no personal rights, trademarks, or competitive products are infringed.
• explicitly excludes the right for derivatives.
- Springer does not supply original artwork or content.
- According to the format which you have selected, the following conditions apply accordingly:
• **Print and Electronic:** This License include use in electronic form provided it is password protected, on intranet, or CD-Rom/DVD or E-book/E-journal. It may not be republished in electronic open access.
• **Print:** This License excludes use in electronic form.
• **Electronic:** This License only pertains to use in electronic form provided it is password protected, on intranet, or CD-Rom/DVD or E-book/E-journal. It may not be republished in electronic open access.
For any electronic use not mentioned, please contact Springer at permissions.springer@spi-global.com.
- Although Springer controls the copyright to the material and is entitled to negotiate on rights, this license is only valid subject to courtesy information to the author (address is given in the article/chapter).
- If you are an STM Signatory or your work will be published by an STM Signatory and you are requesting to reuse figures/tables/illustrations or single text extracts, permission is granted according to STM Permissions Guidelines: http://www.stm-assoc.org/permissions-guidelines/
For any electronic use not mentioned in the Guidelines, please contact Springer at permissions.springer@spi-global.com. If you request to reuse more content than stipulated in the STM Permissions Guidelines, you will be charged a permission fee for the excess content.
Permission is valid upon payment of the fee as indicated in the licensing process. If permission is granted free of charge on this occasion, that does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.
-If your request is for reuse in a Thesis, permission is granted free of charge under the following conditions:
This license is valid for one-time use only for the purpose of defending your thesis and with a maximum of 100 extra copies in paper. If the thesis is going to be published, permission

needs to be reobtained.
- includes use in an electronic form, provided it is an author-created version of the thesis on his/her own website and his/her university's repository, including UMI (according to the definition on the Sherpa website: http://www.sherpa.ac.uk/romeo/);
- is subject to courtesy information to the co-author or corresponding author.

Geographic Rights: Scope

Licenses may be exercised anywhere in the world.

Altering/Modifying Material: Not Permitted

Figures, tables, and illustrations may be altered minimally to serve your work. You may not alter or modify text in any manner. Abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of the author(s).

Reservation of Rights

Springer reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction and (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

License Contingent on Payment

While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Springer or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received by the date due, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Springer reserves the right to take any and all action to protect its copyright in the materials.

Copyright Notice: Disclaimer

You must include the following copyright and permission notice in connection with any reproduction of the licensed material:
"Springer book/journal title, chapter/article title, volume, year of publication, page, name(s) of author(s), (original copyright notice as given in the publication in which the material was originally published) "With permission of Springer"
In case of use of a graph or illustration, the caption of the graph or illustration must be included, as it is indicated in the original publication.

Warranties: None

Springer makes no representations or warranties with respect to the licensed material and adopts on its own behalf the limitations and disclaimers established by CCC on its behalf in its Billing and Payment terms and conditions for this licensing transaction.

Indemnity

You hereby indemnify and agree to hold harmless Springer and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

No Transfer of License

This license is personal to you and may not be sublicensed, assigned, or transferred by you without Springer's written permission.

No Amendment Except in Writing

This license may not be amended except in a writing signed by both parties (or, in the case of Springer, by CCC on Springer's behalf).

Objection to Contrary Terms

Springer hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these

terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and Springer (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

Jurisdiction

All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by arbitration, to be held in the Federal Republic of Germany, in accordance with German law.

Other conditions:

V 12AUG2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

ELSEVIER LICENSE
TERMS AND CONDITIONS

Apr 08, 2017

This Agreement between Venkatakumar Srinivasan ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4084060421069
License date	Apr 08, 2017
Licensed Content Publisher	Elsevier
Licensed Content Publication	Theoretical Computer Science
Licensed Content Title	On communication protocols that compute almost privately
Licensed Content Author	Marco Comi,Bhaskar DasGupta,Michael Schapira,Venkatakumar Srinivasan
Licensed Content Date	26 October 2012
Licensed Content Volume	457
Licensed Content Issue	n/a
Licensed Content Pages	14
Start Page	45
End Page	58
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Analysis of Privacy Measures for Multi-Agent and Networked Systems
Expected completion date	Apr 2017
Estimated size (number of pages)	100
Elsevier VAT number	GB 494 6272 12
Requestor Location	Venkatakumar Srinivasan 1926 W Harrison Street Apt 902 CHICAGO, IL 60612 United States Attn: Venkatakumar Srinivasan
Publisher Tax ID	98-0397604
Total	0.00 USD

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.
3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."
4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.
5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.
10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).
13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:
15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.
16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.
Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.
17. **For journal authors:** the following clauses are applicable in addition to the above:
Preprints:
A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).
Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.
Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.
Authors can share their accepted author manuscript:

- immediately
  - via their non-commercial person homepage or blog
  - by updating a preprint in arXiv or RePEc with the accepted manuscript
  - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
  - directly by providing copies to their students or to research collaborators for their personal use
  - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
  - via non-commercial hosting platforms such as their institutional repository
  - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.
Policies for sharing publishing journal articles differ for subscription and gold open access articles:
Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.
Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.
If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.
Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.
Please refer to Elsevier's posting policy for further information.
18. **For book authors** the following clauses are applicable in addition to the above:
Authors are permitted to place a brief summary of their work online only. You are not

allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.
19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions
You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.
Terms & Conditions applicable to all Open Access articles published with Elsevier:
Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.
The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.
If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.
Additional Terms & Conditions applicable to each Creative Commons user license:
CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.
CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.
CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0.
Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.
Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access

- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.9

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.