A Study on a Privacy Measure for Social Networks: Computational Complexity and Properties on Random Graphs

by

Congsong Zhang

B.Sc. in Communication Engineering, Nanjing Tech University, China, 2006
M.Sc. in Information and Communication Engineering, Southeast University, China, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The College of Graduate Studies

(Computer Science)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

December 2017

The following individuals certify that they have read, and recommend to the College of Graduate Studies for acceptance, a thesis/dissertation entitled:

A Study on a Privacy Measure for Social Networks: Computational Complexity and Properties on Random Graphs

submitted by Congsong Zhang in partial fulfilment of the requirements of the degree of Master of Science

Dr. Yong Gao, Irving K. Barber School of Arts and Sciences, Unit 5 - Computer Science
Supervisor

Dr. Heinz Bauschke, Irving K. Barber School of Arts and Sciences, Unit 5 - Mathematics
Supervisory Committee Member

Dr. Paramjit Gill, Irving K. Barber School of Arts and Sciences, Unit 5 - Statistics
Supervisory Committee Member

Dr. Zheng Liu, School of Engineering
University Examiner

Abstract

Since the booming of social networks, network analysis has benefited greatly from the released data. Meanwhile, the leakage of users' information is getting more and more serious. The users' personal information may be compromised even if the data are released after being anonymized. Adversaries can uniquely re-identify a user in an anonymized social network by the quasi-identifiers as their background knowledge. To measure the resistance against privacy attacks in anonymized social networks where the background knowledge of adversaries is the metric representation, R. Trujillo-Rasua et al. introduced a new measure: (k, l)-anonymity based on the notions of k-antiresolving set and k-metric antidimension in [TRY16b].

In this thesis, we prove that the problem of computing k-metric antidimension is NP-hard by a polynomial-time reduction from a well-known NP-complete problem, the exact cover by 3-sets problem (X3C problem), to a decision version of the problem of computing k-metric antidimension. With this conclusion, we prove that the (k, l)-anonymity problem is NP-complete.

Also, in the hope to get a general relation between k and l in the (k, l)-anonymity problem, we study the behaviors of k-antiresolving sets in Erdős-Rényi random graphs. We establish three bounds on the size of k-antiresolving sets in Erdős-Rényi random graphs leading to a range of k-metric antidimension where k is constant.

Preface

The results contained in this thesis have appeared in [ZG17].

Table of Contents

List of Figures

Acknowledgements

First of all, I would like to thank Unit 5 of UBC Okanagan for the two years' support. Evolving the master program of computer science for two years is quite a fantastic experience. Also, I greatly appreciate the opportunity of teaching undergraduate students as a lab instructor.

I want to give my thanks to everyone in the department I have dealed with as well. You are so friendly and have contributed an excellent experience for me.

I also would like to thank Dr. Y. Gao for his mentorship and encouragement over the duration in the last two years. His mathematical influence is invaluable for me, and his guideline of how to contribute to a solid research drives force behind the work of this thesis.

Dedication

I would like to dedicate this thesis to my wife Lingling Ji, my lovely daughter Jiyan Zhang, and my parents Shuyi Zhang and Xiuyun Cong.

My parents give me every academic opportunity I pursuit, as well as my wife. Without their support, I can not imagine that I can finish my master degree in UBC Okanagan.

Lastly, let me thank again for their endless support and love.

Chapter 1

Introduction

The Internet and big data analytics bring a significant change to our lives. We can use the online banking to make our payment conveniently and efficiently. E-commence applications, such as Taobao and Amazon, give us a new experience of shopping. Moreover, by social networks, e.g., Facebook, Twitter, and Linkedin, we can acquaint new friends around the world and exchange messages instantly. The result extracted from the analysis of big data brings us the knowledge unknown before.

But these benefits are not free. With the booming of Internet, information privacy and secrecy become more and more important. A malicious hacker can attack web servers and cause damage to the public. For example, CBC reported on April 14th, 2014, that 900 social insurance numbers were stolen [SIN]. The attacker accessed the data by exploiting a bug of OpenSSL - Heartbleed [Hea].

Even though a malicious hacker does not attack servers actively, the inappropriate leakage of user information can lead a breach of privacy. For example, A. Tockar shows an example on how to track individuals in New York using data from the 2013 NYC Taxi data release [Toc].

The concepts of information secrecy and information privacy in computer and information sciences are related, but still with some significant differences. Information secrecy is a practice of sharing information with a group of people or a person while hiding it from others, and the challenge of information privacy is to utilize information while protecting individual's information.

To achieve information security, we use many cryptographic techniques. One of them is cryptosystems. On the other hand, information privacy is an interdisciplinary topic of studying how to prevent private information being recovered from released data sets. For example, investers can use the transaction data of stock market to predict the future of stock market in order to make a profit. But, they can not extract the personal transaction information.

In the practice of protecting information privacy, there are two main frameworks: interactive and non-interactive. In the interactive framework,

data analysts inquiry data from a trusted data collector; in the non-interactive framework, the data collector uses anonymization techniques, e.g., the k-anonymity [Swe02], to get rid of identifies of data and then publishes the anonymized data.

Although anonymization techniques can get rid of identifiers of data in social networks, adversaries may compromise the privacy by using quasi-identifiers as their background knowledge.

In an anonymized network graph, let u be a user vertex and S be a subset of attacker vertices. Supposing the background knowledge of adversaries is the metric representation of u with respect to S, R. Trujillo-Rasua et al. in [TRY16b] introduced a measure of resistance against privacy attacks in anonymized social networks: (k, l)-anonymity. This notion creats a new problem in graph theory.

Let $G = (V, E)$ be a simple connected graph, S be a proper subset of V, and v be a vertex in $V \setminus S$. The metric representation of v with respect to S is a tuple formed by the shortest-path distances from v to vertices in S. The set S is a k-antiresolving set if k is the greatest integer that for any vertex $v \in V \setminus S$ there are at least $k-1$ different vertices in $V \setminus S$ having the same metric representation with respect to S as v. A k-antiresolving basis is defined to be a k-antiresolving set of minimum cardinality. The k-metric antidimension is the cardinality of a k-antiresolving basis. We denote the k-metric antidimension by $\mathrm{adim}_k(G)$. The graph G meets (k, l)-anonymity if k is the smallest positive integer that $\mathrm{adim}_k(G)$ is less than or equal to l.

From the definition of k-antiresolving set, it is clear that, if the set of controlling nodes by an adversary is a k-antiresolving set, the adversary can not uniquely re-identify other nodes with the probability that is greater than $1/k$. The $\mathrm{adim}_k(G)$ is the lower bound on the number of vertices controlled by an adversary to approach this probability.

The main results of this thesis include a proof of the NP-hardness of the problem of computing $\mathrm{adim}_k(G)$ and the bounds on the size of k-antiresolving sets in Erdős-Rényi random graphs. The NP-hardness proof is based on a reduction from a well-known NP-complete problem: the exact cover by 3-sets problem (X3C problem). The bounds on the size of k-antiresolving sets are established by making use of an observation under the case that an Erdős-Rényi random graph has the diameter less than or equal to 2. This observation helps us to overcome the difficulty brought by the dependence of the shortest-path distances between different vertices in Erdős-Rényi random graphs.

In Chapter 2, we give an overview of information privacy and information secrecy including the origins of these two topics and the development in both

areas. In Chapter 3, we give the background knowledge of this thesis, as well as our results and the related works. In Chapter 4, we give a polynomial-time reduction from X3C problem to a decision version of the problem of computing $\mathrm{adim}_k(G)$. With the reduction, we can say the decision version of the problem of computing $\mathrm{adim}_k(G)$ is also NP-complete. Thus, the problem of computing $\mathrm{adim}_k(G)$ is NP-hard. With this conclusion, we give the proof of (k, l)-anonymity is NP-complete. In Chapter 5, we give the proofs of the bounds on the size of k-antiresolving sets in Erdős-Rényi random graphs $G(n, p)$. With the bounds, we give a range of k-metric antidimension where $k \in O(1)$ when n tends to infinite. In Chapter 6, we summarize our results and mention the future direction and some open problems.

Chapter 2

Information Privacy and Information Secrecy

The study of information privacy has benefited greatly from the development of computer science. For example, graph theory and the study of social networks have supported the study of information privacy under the popularity and accessibility of online social networks in recent years.

In Section 2.1, we give an overview of information privacy. Then, we introduce the techniques of how to protect information privacy in the practice of social networks.

Information secrecy that benefits a lot from Number Theory is often confused with information privacy. Then, in Section 2.2, we also give a brief introduction on this topic.

2.1 Information Privacy

Information privacy is an interdisciplinary concept, and its meaning varies with cultures and historical stages [NHP11, Hol09, BJKL04]. S. Warren et al. defined the concept of privacy from the law perspective [WB90]. Because of the advancement of computer technology, such as big data storage and analytics, the Internet, and social networks, information privacy has become an increasing concern [JS05, CP02]. A potential danger is that most actions in our daily lives are recorded on computers somewhere, e.g., the track of what we bought in supermarkets or our medical history. Improper disclosure of such information can lead harmful effects [And96, WW96]. Another hidden danger arises from the practice of data mining and social networks analysis [Ale11]. The purposes of these two processes are to figure out some patterns in large data sets and investigate social structures in social networks. It requires us to reserve some statistical information when we do the sanitization on the original information. However, it is a trade-off between preserving statistical information and hiding personal information. If we sanitize the original information thoroughly, we will lose

statistical information. On the other hand, if we reserve statistical information too much, adversaries may compromise individual privacy just like the way of breaking deterministic cryptosystems mentioned in Subsection 2.2. Furthermore, the results extracted from the data mining or social networks analysis, such as classification rules, may make a breach on individual privacy. The results may even lead to unfair treatments, e.g., the bias or discrimination on specific groups or races, when being used on decision tasks, see [PRT08, DL17, CCSZ13].

Interactive and non-interactive frameworks are two models for privacy mechanisms in the practice of data mining and social networks analytics [Dwo06]. In the setting of the interactive model, a trusted data collector provides some interfaces of queries through which users can get data. In the setting of the non-interfaces model, a data collector publishes the collected data after getting rid of identifiers of data, such as names and social insurance numbers. The corresponding techniques are called anonymization techniques or de-identification techniques in the literature.

The results of the interactive case are powerful, see [AS00a, DMNS06]. But the non-interactive case seems to be more difficult [EGS03]. A possible reason is that it is difficult to supply the utility that has not been specified when the sanitization is carried out [Dwo06].

T. Dalenius articulated a desideratum for the statistical database in 1977 [Dal77]: nothing about an individual should be learnable from the database that can not be learned without access to the database. This desideratum means that anything that an adversary could learn from the database can also be learned without the database so that accessing to the database would not compromise individual privacy. This notion was defined by S. Goldwasser et al. as the semantic security which we will introduce in Subsection 2.2. C. Dwork proved that this type of privacy could not be achieved [Dwo06], and then C. Dwork designed a relative guarantee called differential privacy mechanism: any given disclosure will be, within a small factor, no matter whether the individual participates in the database. Below is the formal definition of differential privacy mechanism in [Dwo06].

Definition 2.1. [Dwo06](*Differential privacy*). A randomized function $\mathcal{K}$ gives ϵ-differential ($\epsilon > 0$) if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq \mathit{Range}(\mathcal{K})$,

$$\Pr[\mathcal{K}(D_1) \in S] \leq \exp(\epsilon) \times \Pr[\mathcal{K}(D_2) \in S].$$

The ϵ-differential privacy mechanism matches the relative guarantee. For example, an insurance provider consults the database to decide whether to

insure Tom, and the presence or absence of Tom in the database would not affect the result significantly. F. McSherry proved that if we query a ϵ-differential privacy mechanism t times where each query is independent, then the result is ϵt-differential privacy [McS09].

Social networking services are used widely in our lives, such as Facebook and Linkedin. The popularity of social networks enables governments or third-party institutions to collect the data of social networks which can be released for research purposes. The analysis of social networks can help us uncover previously unknown knowledge, e.g., community-based problems. In other fields, such as recommended systems, the analysis of social networks also supports substantially [SANT14].

However, these benefits are not free. An adversary can compromise the individual privacy from the published data of social networks and make the sensitive information of individuals disclosure. An approach to solve this issue is to use anonymization techniques to remove identify attributes before releasing the data, see a brief survey on anonymization techniques and anonymized data of some social networks in [ZPL08, Les]. But, due to the complex structure of social networks, an adversary still can compromise the individual privacy by the background knowledge of quasi-identifier attributes of victims, e.g., link relationship, neighborhoods, and embedded subgraphs.

In [BDK07], L. Backstrom et al. described a family of privacy attacks in anonymized social networks. W. Peng et al. gave another example called the two-stage deanonymization attack in [PLZW14]. The example shows that an adversary can first register new users with connections to the targeted users in a social network. Then the adversary creates edges between the newly registered users to construct a unique subgraph. After that, the adversary identifies the subgraph in the anonymized social network that is released so that the adversary can re-identifies the targeted users.

K. Liu et al. proposed a framework for the identity anonymization on graphs under the background knowledge of adversaries is vertex degree [LT08]. Contemporaneously, B. Zhou et al. studied how to protect privacy in social networks against neighborhood attacks [ZP08].

We refer to [NHP11, WYLC10] for further reading on the privacy-preserving publication of social graphs.

2.2 Information Secrecy

In the first century, Pliny the Elder described how the milk from a thitymallus plant could be used as invisible ink [Mat]. Unlike hiding the actual

information, croptography is a class of techniques to hide the meaning of information. Cryptography makes sure the real meaning of information could not be revealed to the wrong receiver. Indeed, cryptography is the practice and study of techniques for guaranteeing information secrecy by secure communications in the presence of the third parties called adversaries [Riv90]. In cryptography, a cryptosystem is a term referring to a set of cryptographic algorithms used for information security services [MOV01]. A cryptosystem can hide the real meaning of data from adversaries. People may firstly use cryptosystems in the military field. Julius Caesar invented the Caesar cipher to protect messages of military significance around 50 B.C. [LP87]. Another famous story about cryptosystems is that Alan Turing tackled the problem of Enigma [Cop04].

A cryptosystem has three types of algorithms: (1) the key generation algorithm; (2) the encryption algorithm; (3) the decryption algorithm. Below is the formal definition of a cryptosystem.

Definition 2.2. [Buc04](*Cryptosystem*). A cryptosystem is a tuple

$$(\mathcal{P}, \mathcal{C}, \mathcal{K}, \mathcal{E}, \mathcal{D})$$

with the following properties:

1. $\mathcal{P}$ is a set. It is called *the plaintext space*. Its elements are called *plaintexts*.

2. $\mathcal{C}$ is a set. It is called *the ciphertext space*. Its elements are called *ciphertexts*.

3. $\mathcal{K}$ is a set. It is called *the key space*. Its elements are called *keys*.

4. $\mathcal{E} = \{E_k | k \in \mathcal{K}\}$ is a family of functions $E_k : \mathcal{P} \to \mathcal{C}$. Its elements are called *encryption functions*.

5. $\mathcal{D} = \{D_k | k \in \mathcal{K}\}$ is a family of functions $D_k : \mathcal{C} \to \mathcal{P}$. Its elements are called *decryption functions*.

6. For each $e \in \mathcal{K}$, there is a $d \in \mathcal{K}$ such that $D_d(E_e(p)) = p$ for all $p \in \mathcal{P}$.

We assume that an adversary can intercept ciphertexts and compute posterior probabilities for different plaintexts. In 1949, C. Shannon introduced the definition of perfect secrecy against such adversaries for a cryptosystem [Sha49]. Informally, we say a cryptography system has perfect secrecy if the adversary could not get any information by intercepting the ciphertext. To

formalize this property mathematically, let $\Pr(X)$ be the prior probability for a message X and $\Pr(X|Y)$ be the posterior probability where Y is a ciphertext. A cryptography system has perfect secrecy if $\Pr(X|Y) = \Pr(X)$ for all plaintexts X and all ciphertexts Y.

In a cryptosystem, if encryption functions and decryption functions use the same keys, the cryptosystem is symmetric; otherwise, the cryptosystem is asymmetric. Advanced Encryption Standard (AES) and Triple Data Encryption Standard (3DES) are two well-known symmetric cryptosystems [Sta05]. In asymmetric cryptosystems, the key e of encryption functions and the key d of decryption functions are different. If a person wants to receive the encrypted messages, he could publish an encryption key e and keep the corresponding decryption key d. Anyone can use e to encrypt messages and send the ciphertexts to him, and only he could decrypt the ciphertexts by d. Therefore, asymmetric cryptosystems are also called public key cryptosystems. R. Rivest et al. firstly described a practical public key cryptosystem RSA [RSA78].

Symmetric cryptosystems are faster than asymmetric cryptosystems. Moreover, their implementations are not complicated due to their lower complexity compared to asymmetric cryptosystems. However, a significant disadvantage of symmetric cryptosystems is that symmetric cryptosystems require all participants have already been configured well with same keys through some external security channels.

Both symmetric and asymmetric cryptosystems described above are deterministic cryptosystems. We say a cryptosystem is deterministic which means this cryptosystem always produces the same ciphertext for a given plaintext and key in separate executions. Deterministic cryptosystem may leak information to an adversary. An adversary can do a statistical analysis of ciphertexts by listening to the encrypted channel and then construct a dictionary of pairs of plaintexts and ciphertexts if they have a close enough appearance frequency. For example, if plaintexts are English sentences and the appearance frequency of a ciphertext is close to the appearance frequency of the word *the* in English sentences, then the adversary could say this ciphertext corresponding to the plaintext *the*. This problem is serious in public key cryptosystems. As any part in public key cryptosystems can encrypt plaintexts by a public key, an adversary could build the dictionary of pairs of plaintexts and ciphertexts actively.

To overcome this problem, S. Goldwasser et al. proposed a probabilistic public key cryptosystem in 1984 [GM84]. The probabilistic public cryptosystem publishes a pair of integers as a public key. The sender can use the public key to encrypt a plaintext into a random ciphertext, and the receiver

can decrypt the random ciphertext by the corresponding decryption key. S. Goldwasser et al. proved that their cryptosystem is polynomial security and semantic security [GM84]. Informally, we say that a public key cryptosystem is polynomial security if, for all plaintexts with any probability distribution, an adversary could not find two plaintexts m_1 and m_2 whose ciphertexts are distinguishable in polynomial time of the length of plaintexts. Informally, we say that a public key cryptosystem is semantic security if the information that an adversary can compute from a given ciphertext could also be computed without the ciphertext. We use two games[1] to illustrate the concept of semantic security.

Game 1:

Let f be a function defined on all plaintexts. Both us and the adversary know this function. We randomly choose a plaintext m. Then we ask an adversary to guess the value of $f(m)$ without being told the value of m.

Game 2:

The adversary choose a function f_E defined on all plaintexts, and then the adversary tells us the function. After that, we randomly choose a plaintext m and do not tell the adversary this plaintext. We compute the ciphertext for this plaintext and give this ciphertext to the adversary. Then we ask the adversary to guess the value of $f_E(m)$.

We say a cryptosystem is semantic security if the adversary can not win Game 2 with higher probability than Game 1.

[1] These two games also come from [GM84]

Chapter 3

Background Knowledge and Results

In Section 3.1, we introduce the complexity classes of P, NP, NP-complete, and NP-hard. In Section 3.2, we show the random graph model and the probabilistic method used in Chapter 5. Moreover, in Section 3.3, we introduce the formal definitions of k-antiresolving set, k-antiresolving basis, k-metric antidimension, and (k, l)-anonymity. In Section 3.4, we show our results and the related works.

3.1 The Complexity Classes of Problems

If the answer to a problem is yes or no, we say this problems is a decision problem. On the other hand, if a problem asks us to get a maximum or minimum solution, we say this problem is an optimization problem. Problems can also be categorized by the level of their difficulty. To illustrate the difficulty of problems, we first introduce the concepts of tractable and intractable problems.

Given an algorithm whose input size is n, if the running time of the algorithm is bounded by $O(n^k)$, where k is a constant, we say the algorithm is a polynomial-time algorithm. Moreover, if a problem has a polynomial-time algorithm, we say the problem is tractable, and if a problem has no polynomial-time algorithm, we say the problem is intractable.

Informally, we define the class of decision problems that can be solved by a polynomial-time algorithm as the class P [CJW$^+$06]. Therefore, the class P is tractable. For example, given two integer arrays A and B, the problem of whether A is the sorted array of B belongs to the class P. We can sort B in polynomial time of the length of B and then compare the sorted array to A.

We define the class NP as the class of decision problems that given a solution whose size is a polynomial of the problem input, we can verify whether the solution is correct in polynomial time [CLRS09]. If a problem

belongs to the class NP and its hardness is as hard as any problem in the class NP, we say the problem is NP-complete. In 1971, A. Cook found the first NP-complete problem: the boolean satisfiability problem [Coo71]. Under the study of NP-complete problems, researchers found many NP-complete problems, e.g., the clique problem, 3-CNF-SAT problem, and the exact cover by 3-sets problem (X3C).

Given two problems Q and $Q^{'}$, if any instance of Q can be solved by an algorithm of a polynomial number of primitive steps and a polynomial number of calls to an algorithm for Q', we say that Q can be reduced to Q' in polynomial-time and denoted by $Q \leq_P Q^{'}$.

Below is an example of a polynomial-time reduction in [CLRS09]. Given an instance of $ax + b = 0$, we transform it to $0x^2 + ax + b = 0$. The solution of $0x^2 + ax + b = 0$ is also the solution of $ax + b = 0$.

By the definition of the polynomial-time reduction, we can explain the meaning of a problem is as hard as another problem and give the formal definition of NP-complete and NP-hard. For two NP problems A and B, if $A \leq_P B$, we say A is no more than a polynomial factor harder than B .

Definition 3.1. [CLRS09](NP-complete). A problem Q is NP-complete if

1. $Q \in$ NP, and
2. $Q^{'} \leq_P Q$ for any $Q^{'} \in$ NP.

If a problem does not satisfy Property 1, e.g., the problem is an optimization problem, but satisfies Property 2, we say the problem is NP-hard.

If a decision problem can be solved by a polynomial-time algorithm, obviously a solution to this problem can be verified in polynomial time. Therefore, the class P is a subset of the class NP. But whether every decision problem whose solution can be verified in polynomial time can also be solved in polynomial time is an open question. If the answer is true, it means P = NP, and the class NP is also tractable.

A possible way to prove P = NP is to find a polynomial-time algorithm for an NP-complete problem. However, to our best knowledge, there is no polynomial-time algorithm for any NP-complete problem, and no one has given the proof of the class NP-complete is intractable.

The problem of whether P = NP is one of seven Millennium Prize Problems. The prize to the first correct solution to this problem is one million dollars [Dev02].

Besides the complexity classes we described above, there are other complexity classes, e.g., PSPACE - the class of decision problems can be solved in

polynomial space and PSPACE-complete - the hardest problems in PSPACE [AB09].

Furthermore, some problems can not be solved by any computer, e.g., the halting problem. The halting problem is the problem of determining, from a description of an arbitrary computer program and an input, whether the program will stop or run forever. A. Turing proved that there is no general algorithm to solve the halting problem [Tur36].

In the rest of this section, to illustrate how to reduce an NP-complete problem to another NP problem, we give a revised NP-completeness proof of the X3C problem by a reduction from 3-CNF-SAT problem.

Definition 3.2. [CLRS09](*3-CNF-SAT problem*).
A *literal* in a boolean formula is an occurrence of a variable or its negation. A boolean formula is in *conjunctive normal form*, or *CNF*, if it is expressed as an AND of *clauses*, each of which is the OR of one or more literals. A boolean formula is in *3-conjunctive normal form*, or *3-CNF*, if each clause has exactly three distinct literals.

Below is the definition of X3C problem.

Definition 3.3. [GJ79](*X3C problem*).
Given a set $B = \{e_1, ..., e_{3q}\}$ and a family $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_p\}$ of 3-element subsets of B, does $\mathcal{S}$ contain a subfamily such that every element in B occurs in exactly one member of the subfamily?

Theorem 3.4. [GJ79] *X3C problem is NP-complete.*

Proof. Let us suppose that an instance of 3-CNF-SAT problem has n boolean variables: $x_1, ..., x_n$ and k clauses: $C_1, ...C_k$. We give a $O(n^2k^2)$-time reduction from this instance to an instance of X3C problem.

Reduction:

1. The elements:

   (a) For each variable x_i, we create $4k$ elements:

   $$a_{i,1}, ..., a_{i,2k} \text{ and } b_{i,1}, ..., b_{i,2k}.$$

   (b) For each clause C_j, we create 2 elements: c_j and c'_j.

   (c) Futhermore, we create $2(n-1)k$ elements:

   $$q_1, ..., q_{(n-1)k} \text{ and } q'_1, ..., q'_{(n-1)k}.$$

2. The 3-element subsets:

   (a) We suppose that a clause C_j contains the variables x_{n_1}, x_{n_2}, and x_{n_3}. If x_{n_1} appears as a positive literal in C_j, we create a subset $\{c_j, c'_j, b_{n_1,2j-1}\}$; otherwise, we create a subset $\{c_j, c'_j, b_{n_1,2j}\}$. x_{n_2} and x_{n_3} apply the same reduction.

   (b) We create $(n-1)k$ subsets: $\{q_1, q'_1, b_{i,j}\}, ..., \{q_{(n-1)k}, q'_{(n-1)k}, b_{i,j}\}$ for each $b_{i,j}$.

   (c) For $a_{i,1}, ..., a_{i,2k}$, we create $2k$ subsets: $\{a_{i,j}, a_{i,(j+1)}, b_{i,j}\}$ where $j \in \left[1, 2k-1\right]$ and $\{a_{i,2k}, a_{i,1}, b_{i,2k}\}$.

As the number of elements created is

$$4kn + 2k + 2(n-1)k$$

and the number of 3-element subsets is

$$3k + 2n(n-1)k^2 + 2nk,$$

the reduction can finish in $O(n^2k^2)$-time. For example, let us consider a 3-CNF-SAT problem instance of 4 variables and 2 clauses where $C_1 = \left(x_1 \vee \overline{x_2} \vee x_3\right)$ and $C_2 = \left(x_2 \vee \overline{x_3} \vee x_4\right)$. After the reduction, we have a set B of elements:

$$\begin{aligned} B = \{&c_1, c'_1, c_2, c'_2, \\ &a_{1,1}, a_{1,2}, a_{1,3}, a_{1,4}, \\ &b_{1,1}, b_{1,2}, b_{1,3}, b_{1,4}, \\ &a_{2,1}, a_{2,2}, a_{2,3}, a_{2,4}, \\ &b_{2,1}, b_{2,2}, b_{2,3}, b_{2,4}, \\ &a_{3,1}, a_{3,2}, a_{3,3}, a_{3,4}, \\ &b_{3,1}, b_{3,2}, b_{3,3}, b_{3,4}, \\ &a_{4,1}, a_{4,2}, a_{4,3}, a_{4,4}, \\ &b_{4,1}, b_{4,2}, b_{4,3}, b_{4,4}, \\ &q_1, q_2, q_3, q_4, q_5, q_6, \\ &q'_1, q'_2, q'_3, q'_4, q'_5, q'_6\} \end{aligned}$$

and a family $\mathcal{S}$ of 3-element subsets of B:

$$\begin{aligned}\mathcal{S} = \Big\{ & \{c_1, c_1', b_{1,1}\}, \{c_1, c_1', b_{2,2}\}, \{c_1, c_1', b_{3,1}\}, \\ & \{c_2, c_2', b_{2,3}\}, \{c_2, c_2', b_{3,4}\}, \{c_1, c_1', b_{4,3}\}, \\ & \{a_{1,1}, a_{1,2}, b_{1,1}\}, \{a_{1,2}, a_{1,3}, b_{1,2}\}, \{a_{1,3}, a_{1,4}, b_{1,3}\}, \{a_{1,4}, a_{1,1}, b_{1,4}\}, \\ & ..., \\ & \{a_{4,1}, a_{4,2}, b_{4,1}\}, \{a_{4,2}, a_{4,3}, b_{4,2}\}, \{a_{4,3}, a_{4,4}, b_{4,3}\}, \{a_{4,4}, a_{4,1}, b_{4,4}\}, \\ & \{q_1, q_1', b_{1,1}\}, \{q_1, q_1', b_{1,2}\}, \{q_1, q_1', b_{1,3}\}, \{q_1, q_1', b_{1,4}\}, \\ & \{q_1, q_1', b_{2,1}\}, \{q_1, q_1', b_{2,2}\}, \{q_1, q_1', b_{2,3}\}, \{q_1, q_1', b_{2,4}\}, \\ & \{q_1, q_1', b_{3,1}\}, \{q_1, q_1', b_{3,2}\}, \{q_1, q_1', b_{3,3}\}, \{q_1, q_1', b_{3,4}\}, \\ & \{q_1, q_1', b_{4,1}\}, \{q_1, q_1', b_{4,2}\}, \{q_1, q_1', b_{4,3}\}, \{q_1, q_1', b_{4,4}\}, \\ & ..., \\ & \{q_6, q_6', b_{1,1}\}, \{q_6, q_6', b_{1,2}\}, \{q_6, q_6', b_{1,3}\}, \{q_6, q_6', b_{1,4}\}, \\ & \{q_6, q_6', b_{2,1}\}, \{q_6, q_6', b_{2,2}\}, \{q_6, q_6', b_{2,3}\}, \{q_6, q_6', b_{2,4}\}, \\ & \{q_6, q_6', b_{3,1}\}, \{q_6, q_6', b_{3,2}\}, \{q_6, q_6', b_{3,3}\}, \{q_6, q_6', b_{3,4}\}, \\ & \{q_6, q_6', b_{4,1}\}, \{q_6, q_6', b_{4,2}\}, \{q_6, q_6', b_{4,3}\}, \{q_6, q_6', b_{4,4}\}\Big\}.\end{aligned}$$

Now we show that a 3-CNF-SAT problem is satisfiable if and only if the reduced X3C problem has an exact cover.

Lemma 3.5. *Given a 3-CNF-SAT problem instance, if the problem is satisfiable, the reduced X3C problem has an exact cover.*

Proof. As the 3-CNF-SAT problem is satisfiable, there is such an assignment of variables that makes each clause is true. If the assignment of a variable x_i is true and x_i appears as a positive literal in a clause C_j, the elements c_j, c_j' can be exactly covered by the subsets $\{c_j, c_j', b_{i,2j-1}\}$.

Similarly, if the assignment of x_i is false and x_i appears as a negative literal in a clause C_j, the elements c_j, c_j' can be exactly covered by the subsets $\{c_j, c_j', b_{i,2j}\}$.

For the same i, the values of j of the covered elements $b_{i,j}$ above are all even or odd, depending on the assignment of x_i.

After applying the above procedure for all variables, we have that all elements $c_j, c_{j'}$ are exactly covered. If not, there exists such a clause that the assignments of its variables appearing as positive literals are false, and the assignments of its variables appearing as negative literals are true. Thus, the clause is not satisfiable which is a contradiction.

Now, let us consider how to exactly cover the elements $a_{i,1}, ..., a_{i,2k}$. If $b_{i,j}$ of an even number j have been covered, $a_{i,1}, ..., a_{i,2k}$ can be exactly covered by the subsets

$$\{a_{i,1}, a_{i,2}, b_{i,1}\}, \{a_{i,3}, a_{i,4}, b_{i,3}\}, ..., \{a_{i,2k-1}, a_{i,2k}, b_{i,2k-1}\}.$$

Otherwise, $a_{i,1}, ..., a_{i,2k}$ can be exactly covered by the subsets

$$\{a_{i,2}, a_{i,3}, b_{i,2}\}, \{a_{i,4}, a_{i,5}, b_{i,4}\}, ..., \{a_{i,2k}, a_{i,1}, b_{i,2k}\}.$$

After covering $a_{i,1}, ..., a_{i,2k}$, we know that k elements $b_{i,j}$ are also covered where j are all even or odd. Then, there are $(n-1)k$ elements $b_{i,j}$ and $(n-1)k$ pairs of q_i and q'_i uncovered.

These elements can be exactly covered by the subsets $\{q_i, q'_i, b_{i',j'}\}$. As shown in the reduction, there are exactly $(n-1)k$ pairs of q_i and q'_i; for each pair of q_i and q'_i, there are $2nk$ subsets $\{q_i, q'_i, b_{i',j'}\}$. Therefore, there is an exact cover for the reduced X3C problem. □

Lemma 3.6. *Given a 3-CNF-SAT problem instance, if the reduced X3C problem has an exact cover, the 3-CNF-SAT problem is satisfiable.*

Proof. For a variable x_i, we suppose that two clauses C_j and C'_j both contain x_i with different literals. Without loss of generality, we suppose that x_i appears as a positive literal in C_j and a negative literal in $C_{j'}$. Then, we can prove that the case of $\{c_j, c'_j, b_{i,2j-1}\}$ and $\{c_{j'}, c'_{j'}, b_{i,2j'}\}$ are in the exact cover does not exist.

By the reduction, the elements $a_{i,1}, ..., a_{i,2k}$ can only be exactly covered by

$$\{a_{i,1}, a_{i,2}, b_{i,1}\}, \{a_{i,3}, a_{i,4}, b_{i,3}\}, ..., \{a_{i,2k-1}, a_{i,2k}, b_{i,2k-1}\}$$

or

$$\{a_{i,2}, a_{i,3}, b_{i,2}\}, \{a_{i,4}, a_{i,5}, b_{i,4}\}, ..., \{a_{i,2k}, a_{i,1}, b_{i,2k}\},$$

which means that we need k elements $b_{i,j}$ where j are all even or odd to exactly cover $a_{i,1}, ..., a_{i,2k}$.

If $\{c_j, c'_j, b_{i,2j-1}\}$ and $\{c_{j'}, c'_{j'}, b_{i,2j'}\}$ are in the exact cover, there are at most $k-1$ uncovered elements $b_{i,j}$ where j are all even or odd. Then, this observation leads to a contradiction that the elements $a_{i,1}, ..., a_{i,2k}$ can not be exactly covered.

Therefore, if the pair of c_j, c'_j is covered by $\{c_j, c'_j, b_{i,2j-1}\}$, the pair of $c_{j'}, c'_{j'}$ would not be covered by $\{c_{j'}, c'_{j'}, b_{i,2j'}\}$, and vice versa.

Now we show that there exists an assignment of variables such that makes each clause is true. Let x_i be true, if a pair of c_j, c'_j is covered by the subset

$\{c_j, c'_j, b_{i,2j-1}\}$; otherwise, let x_i be false if the pair is covered by the subset $\{c_j, c'_j, b_{i,2j}\}$. By the reduction, the existing of subset $\{c_j, c'_j, b_{i,2j-1}\}$ means that x_i appears as a positive literal in C_j. Therefore, the true assignment of x_i makes the clause C_j is true. Similarly, the existing of subset $\{c_j, c'_j, b_{i,2j}\}$ means that x_i appears as a negative literal in C_j, and the assignment of x_i be false makes the clause C_j is true. □

As 3-CNF-SAT problem is NP-complete [CLRS09], Lemma 3.5 and 3.6 lead to Theorem 3.4. □

3.2 Erdős-Rényi Random Graphs and the Probabilistic Method

In this section, we introduce the Erdős-Rényi random graphs. Moreover, we list the main inequalities used in the probabilistic method.

In 1959, P. Erdős and A. Rényi introduced the random graph model now named after them [ER59]. Contemporaneously, E. Gilbert introduced a closely related random graph model independently [Gil59].

Definition 3.7. [ER59, Gil59] ($G(n, M)$ model and $G(n, p)$ model). In the $G(n, M)$ model, a graph with n vertices and M edges is chosen uniformly at random from the collection of all graphs with n vertices and M edges. In the $G(n, p)$ model, a graph with n vertices is built by connecting any two vertices with the same probability independently.

In this thesis, we analyze the properties of k-antiresolving sets in the graph built by the $G(n, p)$ model.

In the following section, we introduce some inequalities in the probabilistic method. The first one is Markov's inequality [Als11].

Theorem 3.8. [Als11] (*Markov's inequality*). *If X(integrable) is a nonnegative random variable, then for any real number $a > 0$,*

$$\Pr\{X \geq a\} \leq \frac{\mathbb{E}(X)}{a}.$$

Proof.

$$\begin{aligned}\mathbb{E}(X) &= \int_0^\infty x \cdot \mathrm{d}F(X) \\ &\geq \int_a^\infty x \cdot \mathrm{d}F(X) \\ &\geq \int_a^\infty a \cdot \mathrm{d}F(X) \\ &= a \cdot \int_a^\infty \mathrm{d}F(X) \\ &= a \cdot \Pr\{X \geq a\}\end{aligned}$$

□

If X is a nonnegative integral valued random variable, then we know

$$\Pr\{X \geq 1\} = \Pr\{X > 0\}.$$

By plugging $a = 1$ into Markov's inequality, we get a special case of Markov's inequality.

Proposition 3.9. [AS00b] *If X is a nonnegative integral valued random variable, then*

$$\mathbb{E}(X) \geq \Pr\{X > 0\}.$$

By using Markov's inequality, we can get Chebyshev's inequality.

Theorem 3.10. [AS00b] (*Chebyshev's inequality*). *If X(integrable) is a random variable with finite expected value μ and finite non-zero variance σ^2, then for any real number $k > 0$*

$$\Pr\{|X - \mu| \geq k\sigma\} \leq \frac{1}{k^2}.$$

Proof.

$$\begin{aligned}\Pr\{|X - \mu| \geq k\sigma\} &= \Pr\{|X - \mu|^2 \geq k^2\sigma^2\} \\ &\leq \frac{\mathbb{E}(|X - \mu|^2)}{k^2\sigma^2} (\text{by Markov's inequality}) \\ &= \frac{\sigma^2}{k^2\sigma^2} \\ &= \frac{1}{k^2}\end{aligned}$$

□

By setting $k\sigma = \mu$, we can prove Theorem 4.3.1 in [AS00b].

Theorem 3.11. [AS00b]

$$\Pr\{X = 0\} \leq \frac{\sigma^2}{\mu^2}$$

Proof.

$$\Pr\{X = 0\} \leq \Pr\{|X - \mu| \geq k\sigma\} \leq \frac{1}{k^2} = \frac{\sigma^2}{\mu^2}.$$

□

We suppose that X is a random variable can be decomposed by $\sum_{i=1}^{n} X_i$ where X_i is the indicator random variable for an event A_i. We say $X_1, ..., X_n$ are symmetric if for every $i \neq j$ there is an automorphism of the underlying probability space that sends the event A_i to the event A_j. For indices i, j, $i \sim j$ means the events A_i and A_j are not independent. Then we have the following theorem.

Theorem 3.12. [AS00b] *Let X is a random variable can be decomposed by $\sum_{i=1}^{n} X_i$ where X_i is the indicator random variable for an event A_i, and $X_1, ..., X_n$ are symmetric. Let*

$$\Delta^* = \sum_{i \sim j} \Pr\{A_j | A_i\}.$$

Then,

$$\mathrm{Var}(X) \leq \mathbb{E}(X) \cdot (1 + \Delta^*).$$

Proof. Note that

$$\mathrm{Var}(X) = \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$$

where

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i) \cdot \mathbb{E}(X_j).$$

As X_i is the indicator random variable for the event A_i, then

$$\mathrm{Var}(X_i) = \Pr\{A_i\} \cdot (1 - \Pr\{A_i\}).$$

Clearly,

$$\mathrm{Var}(X_i) \leq \Pr\{A_i\} = \mathbb{E}(X_i).$$

Thus,

$$\sum_i \mathrm{Var}(X_i) \leq \mathbb{E}(X).$$

As

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= \Pr\{A_i A_j\} - \Pr\{A_i\}\Pr\{A_j\} \\ &= \Pr\{A_i\} \cdot \big(\Pr\{A_j|A_i\} - \Pr\{A_j\}\big),\end{aligned}$$

if A_i and A_j are independent, then

$$\mathrm{Cov}(X_i, X_j) = 0.$$

Otherwise,

$$\mathrm{Cov}(X_i, X_j) \leq \Pr\{A_i\} \cdot \Pr\{A_j|A_i\}.$$

Therefore,

$$\sum_{i \neq j} \mathrm{Cov}(X_i, X_j) \leq \Delta^* \cdot \Big(\sum_i \Pr\{A_i\}\Big) = \Delta^* \cdot \mathbb{E}(X).$$

□

Moreover, in Chapter 5, we will apply Boole's inequality (also called as the union bound in discrete mathematics) and Fréchet inequalities (or called as Boole-Fréchet inequalities).

Theorem 3.13. [LB11] (*Boole's inequality*). *For a finite set of events* $A_1, A_2, ..., A_n$,

$$\Pr\Big\{\bigcup_{i=1}^{n} A_i\Big\} \leq \sum_{i=1}^{n} \Pr\{A_i\}.$$

Theorem 3.14. [LB11] (*Fréchet inequalities*). *For a finite set of events* $A_1, A_2, ..., A_n$,

$$\Pr\Big\{\bigcap_{i=1}^{n} A_i\Big\} \leq \min_{i\in[1,n]} \big(\Pr\{A_i\}\big) \leq \max_{i\in[1,n]} \big(\Pr\{A_i\}\big).$$

Besides the above introduced inequalities, we also use Chernoff Bound and Azuma-Hoeffding inequality. We will show their descriptions and proofs in Appendix B.

3.3 k-Metric Antidimension and (k, l)-Anonymity

The definition of k-metric antidimension comes from the concept of metric dimension in graph theory. To illustrate the metric dimension, we first introduce the concept of metric representation. Below is the formal definition of metric representation in [TRY16b].

Definition 3.15. [TRY16b] (*Metric representation*). Let $G = (V, E)$ be a simple connected graph and $d_G(u, v)$ be the length of a shortest path between the vertices u and v in G. For a set $S = \{u_1, ..., u_t\}$ of vertices in V and a vertex v, we call the t-tuple $r(v|S) := (d_G(v, u_1), ..., d_G(v, u_t))$ the metric representation of v with respect to S.

Given a simple connected graph G, let S be a set of vertices in G. For any two vertices $u, v \notin S$, if they have different metric representations with respect to S, then we call S as a resolving set. The metric dimension of G is the minimum cardinality of a resolving set.

In [KRR96], S. Khuller et al. showed a proof of that the problem of finding the metric dimension of an arbitrary graph is NP-hard. In [EMRVY13], A. Estrada-Moreno et al. generalized the metric dimension by k-metric dimension. A vertex set S is a k-metric generator if for any two vertices $u, v \notin S$, they are distinguished by at least k vertices in S. The minimum cardinality of a k-metric generator is the k-metric dimension. They proved that the problem of finding k-metric dimension is NP-hard.

A resolving set is what an adversary wants to plant in a social network graph. To measure the resistance against adversaries' privacy attacks on anonymized social network graphs whose background knowledge is the metric representation, R. Trujillo-Rasua et al. introduced the concept of k-antiresolving set. Below is the formal definition of k-antiresolving set in [TRY16b].

Definition 3.16. [TRY16b] (*k-antiresolving set*). Let $G = (V, E)$ be a simple connected graph and let $S = \{u_1, ..., u_t\}$ be a subset of vertices of G. The set S is called a k-antiresolving set if k is the greatest positive integer such that for every vertex $v \in V \setminus S$, there exist at least $k - 1$ different vertices $v_1, ..., v_{k-1} \in V \setminus$ S with $r(v|S) = r(v_1|S) = ... = r(v_{k-1}|S)$, i.e., v and v_1 , ..., v_{k-1} have the same metric representation with respect to S.

With the k-antiresolving set, R. Trujillo-Rasua et al. also defined the k-metric antidimension and k-antiresolving basis in [TRY16b].

Definition 3.17. [TRY16b] (*k-metric antidimension and k-antiresolving basis*). The k-metric antidimension of a simple connected graph $G = (V, E)$

is the minimum cardinality amongst the k-antiresolving sets in G and is denoted by $\text{adim}_k(G)$. A k-antiresolving set of cardinality $\text{adim}_k(G)$ is called a k-antiresolving basis for G.

If an adversary controls some vertices in an anonymized network graph, the maximum probability that the adversary can distinguish others users by their metric representations with respect to the adversary's controlled vertices is an important measure of privacy in the graph. The formal definition of this measure is called (k, l)-anonymity [TRY16b].

Definition 3.18. [TRY16b] ((k, l)-*anonymity*). A graph G meets (k, l)-anonymity with respect to active attacks if k is the smallest positive integer such that the k-metric antidimension of G is lower than or equal to l.

3.4 Results and Related Works

We first studied the computational complexity of $\text{adim}_k(G)$ and (k, l)-anonymity. Also, we studied the behavior of k-antiresolving sets in $G(n, p)$ where k and p are constants. Our main results are: (1) a reduction from X3C problem to a decision version of the problem of computing $\text{adim}_k(G)$; (2) three bounds on the size of k-antiresolving sets in $G(n, p)$ with constant p.

The reduction proves that the decision version of the problem of computing $\text{adim}_k(G)$ is NP-complete. Thus, the problem of computing $\text{adim}_k(G)$ is NP-hard. From this conclusion, we show that the (k, l)-anonymity problem is NP-complete.

For $G(n, p)$ random graphs, we use **w.h.p.** as the abbreviation of *with high probability* to mean that the occurrence probability of an event tends to 1 when n tends to infinity. Then, we establish the following three bounds of the size of k-antiresolving sets in $G(n, p)$.

The first bound on the size of k-antiresolving sets is such an upper bound that **w.h.p.** there is no k-antiresolving set where $k \in O(1)$. The second bound is such a lower bound that **w.h.p.** there is no k-antiresolving set where $k \in \omega(1)$. The last bound is such a lower bound that **w.h.p.** there is at least one k-antiresolving set where $k \in O(1)$.

By the first and the last bound, we establish a range of the k-metric antidimension where $k \in O(1)$ in random graphs $G(n, p)$.

In [CDM+16], T. Chatterjee et al. also proved that the problem of computing $\text{adim}_k(G)$ is NP-hard independently. Moreover, they gave a $O(n^3)$-time $(1 + \ln(n - 1))$-approximation algorithm for 1-antiresolving basis.

In [MTRX16], S. Mauw et al. provided an efficient method to transform a graph G into another graph G' such that G' is not (1,1)-anonymity. The method is only based on the edge addition operation.

In [TRY16a], R. Trujillo-Rasua et al. studied how to decide whether a graph is 1-metric antidimensional. They provided characterizations for 1-metric antidimensional trees and unicyclic graphs. Futhermore, they gave efficient algorithms to decide whether these two types of graphs are 1-metric antidimensional.

Chapter 4

The Complexity of Computing $\text{adim}_k(G)$ and (k, l)-Anonymity

In this chapter, we prove that the problem of computing $\text{adim}_k(G)$ in an arbitrary simple connected graph is NP-hard. We prove this result by a polynomial-time reduction from X3C problem to a decision version of the problem of computing $\text{adim}_k(G)$. Then, we show that the (k, l)-anonymity is also NP-complete.

4.1 The Complexity of Computing k-Metric Antidimension

We first introduce some notations used in the proof. Let $G = (V, E)$ be a simple connected graph, S be a subset of V, and v be a vertex in $V \setminus S$. We define

$$N_S(v) = \{u : u \in S, u \text{ and } v \text{ are neighbors}\}.$$

For two vertices u, v in $V \setminus S$, we use $u =_S v$ to mean $r(u|S) = r(v|S)$ where $r(u|S)$ and $r(v|S)$ are the metric representations of u and v with respect to S (see Definition 3.15). Moreover, we denote

$$\{v : v \in V \setminus S, r(v|S) = r'\}$$

by $V_S[r']$ for a metric representation r' with respect to S. Then, we show two properties of k-antiresolving sets in simple connected graphs.

Proposition 4.1. *Let S' be a subset of a k-antiresolving set S in a simple connected graph G. If there is such a metric representation r with respect to S' that $0 < |V_{S'}[r]| < k$, then $V_{S'}[r] \subset S$.*

By the definition of k-antiresolving set, we know this proposition is true.

Proposition 4.2. *Let G be a simple connected graph, S be a k-antiresolving set in G with $k \geq 3$, and $P_n = \{v_{p_1}, ..., v_{p_n}\}$ be a path in G satisfying*: (1) *the degree of the vertex v_{p_n} is equal to* 1; (2) *the degree of the vertex v_{p_i} is equal to* 2 *where $i \in [2, n-1]$. Then, $P_n \subseteq S$, if any $v_{p_i} \in S$ where $i \in [2, n]$.*

Proof. By the definition of k-antiresolving set, for a vertex v in S,

$$|N_{V \setminus S}(v)| = 0$$

or

$$|N_{V \setminus S}(v)| \geq k.$$

Therefore, if a vertex $v_{p_i} \in S$ where $i \in [2, n]$, its neighborhood

$$\{v_{p_{i-1}}, v_{p_{i+1}}\} \subset S.$$

Similarly, the neighborhood of $v_{p_{i-1}}$ is a subset of S if $i - 1 > 1$, and the neighborhood of $v_{p_{i+1}}$ is a subset of S if $i + 1 < n$. By repeating this observation, we get that $P_n \subseteq S$. □

We prove that the following decision version of the problem of computing $\text{adim}_k(G)$ is NP-complete: given two integers k and m, is there a k-antiresolving set of the cardinality is less than or equal to m in a simple connected graph $G = (V, E)$?

Given a subset of vertices in V, in polynomial time of $|V|$, we can verify that whether the cardinality of the subset is less than or equal to m, and whether the subset is a k-antiresolving set. Therefore, this decision problem belongs to the class NP.

We prove the NP-completeness by a reduction from the X3C problem.

4.1.1 A Reduction from the X3C Problem

Given an instance of the X3C problem: a set $B = \{e_1, ..., e_{3q}\}$ and a family $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_p\}$ of 3-element subsets of B, we suppose $p - q \geq 12$. If not, we create such a set $B' = \{e_{3q+1}, ..., e_{3q+36}\}$ and a family $\mathcal{S}' = \{\mathcal{S}_{p+1}, ..., \mathcal{S}_{p+24}\}$ of 3-element subsets of B' that $\mathcal{S}'$ contains an exact cover for B'. Then, we get a new instance of X3C with $B \cup B'$ and $\mathcal{S} \cup \mathcal{S}'$ that satisfies our assumption. Clearly, $\mathcal{S} \cup \mathcal{S}'$ contains an exact cover for $B \cup B'$ if and only if $\mathcal{S}$ contains an exact cover for B.

Let n be such an integer that $\lfloor (p-q)/3 \rfloor < n < \lfloor (p-q)/2 \rfloor$. We construct a simple connected graph $G = (V, E)$ with $|V| = 3qn(q+2) + p$ from an instance of X3C with $3q$ elements and p subsets as follows. Note that $n < \lfloor (p-q)/2 \rfloor$. Thus, we can get G in polynomial time of p and q. Fig. 4.1 shows the gadget corresponding to a subset $S_i = \{e_a, e_b, e_c\}$.

1. Vertices
   (a) For each S_i, we create a vertex v_{S_i}.
   (b) For each e_i, we create n vertices $v_{e_{i,1}}, ..., v_{e_{i,n}}$.
   (c) For each $v_{e_{i,j}}$, we create $q+1$ vertices $v_{p_{i,j,1}}, ..., v_{p_{i,j,q+1}}$.
2. Edges
   (a) We create edges $\{v_{S_i}, v_{S_j}\}$ where $i \neq j$.
   (b) We create edges $\{v_{e_{i,j}}, v_{e_{i',j'}}\}$ where $i \neq i'$ or $j \neq j'$.
   (c) For each $v_{p_{i,j,1}}, ..., v_{p_{i,j,q+1}}$, we create a path $= \{v_{p_{i,j,1}}, ..., v_{p_{i,j,q+1}}\}$.
   (d) For each $v_{e_{i,j}}$, we create edges $\{v_{e_{i,j}}, v_{p_{i',j',1}}\}$ where $i \neq i'$ or $j \neq j'$.
   (e) If a subset $S_i = \{e_a, e_b, e_c\}$, we create edges $\{v_{S_i}, v_{e_{a,j}}\}$, $\{v_{S_i}, v_{e_{b,j}}\}$, and $\{v_{S_i}, v_{e_{c,j}}\}$ for all $j \in [1, n]$.

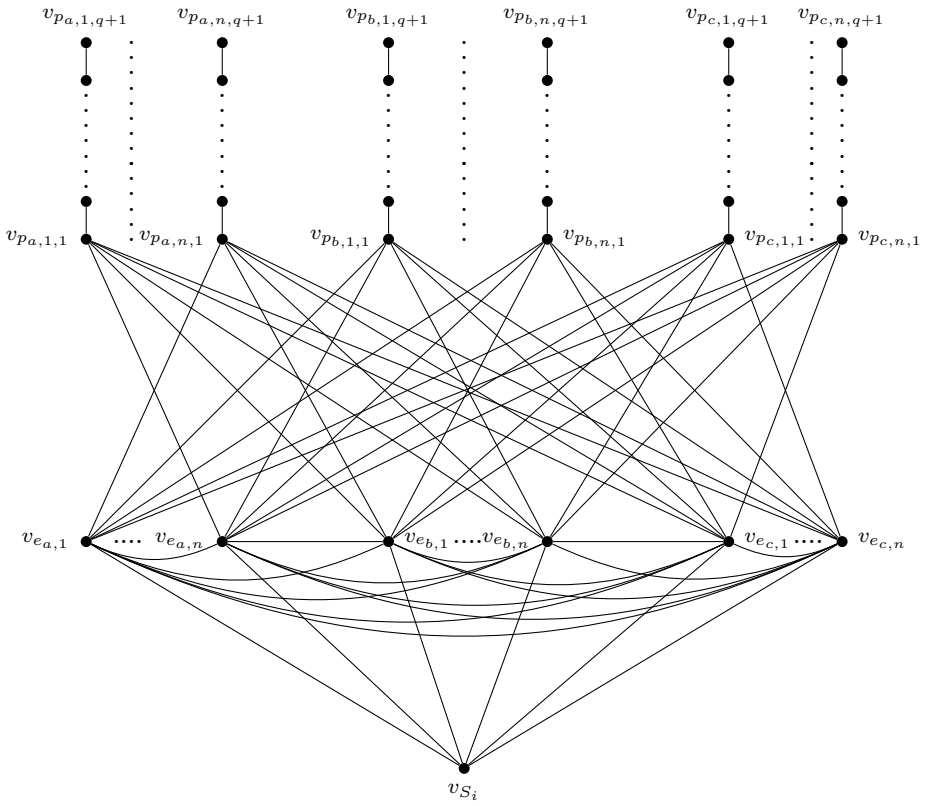

Figure 4.1: A gadget for a subset $S_i = \{e_a, e_b, e_c\}$

4.1.2 The NP-completeness Proof: Sufficient Condition

Lemma 4.3. *If the X3C problem has an exact cover $\mathcal{S}_c$, there exists such a $(p-q)$-antiresolving set S in G that $|S| \leq q$.*

Proof. Let $V_{\mathcal{S}_c}$ and $V_{\mathcal{S}}$ be the sets of vertices corresponding to the subsets in $\mathcal{S}_c$ and $\mathcal{S}$. Then, there are three types of metric representations with respect to $V_{\mathcal{S}_c}$:

1. the q-tuple $(1, ..., 1)$ for the vertex in $V_{\mathcal{S}} \setminus V_{\mathcal{S}_c}$;
2. a q-tuple where only one element is 1 and the remaining elements are 2 for the vertex $v_{e_{i,j}}$;
3. the q-tuple $(\ell+1, ..., \ell+1)$ for the vertex $v_{p_{i,j,\ell}}$.

As $\mathcal{S}_c$ is an exact cover, then

$$|V_{\mathcal{S}_c}| = q \text{ and } |V_{\mathcal{S}} \setminus V_{\mathcal{S}_c}| = p - q.$$

Thus, the number of vertices having the first type of metric representations with respect to $V_{\mathcal{S}_c}$ is $p-q$.

Clearly, for a vertex $v_{e_{i,j}}$, there are $3n-1$ different vertices $v_{e_{i',j'}}$ where $v_{e_{i',j'}} =_{V_{\mathcal{S}_c}} v_{e_{i,j}}$. Moreover, for a vertex $v_{p_{i,j,l}}$, there are $3qn-1$ different vertices $v_{p_{i',j',l}}$ where $v_{p_{i',j',l}} =_{V_{\mathcal{S}_c}} v_{p_{i,j,l}}$.

As $3qn > 3n > p-q$, then $V_{\mathcal{S}_c}$ is a $(p-q)$-antiresolving set. □

4.1.3 The NP-completeness Proof: Necessary Condition

Lemma 4.4. *If there is such a $(p-q)$-antiresolving set S in G that $|S| \leq q$, the X3C problem has an exact cover.*

Proof. We claim three facts:

1. the vertex $v_{e_{i,j}} \notin S$ and the vertex $v_{p_{i,j,l}} \notin S$;
2. $|S| = q$;
3. the metric representation of $v_{e_{i,j}}$ with respect to S has one element valued 1 and the remaining $|S|-1$ elements valued 2.

The vertex $v_{p_{i,j,l}}$ where $l \in [2, q+1]$ is not in S. If not, by Proposition 4.2, the path $\{v_{p_{i,j,1}}, ..., v_{p_{i,j,q+1}}\} \subseteq S$ due to that $p - q \geq 12$. But, the length of the path $\{v_{p_{i,j,1}}, ..., v_{p_{i,j,q+1}}\}$ is $q+1$ which leads to a contradiction that $|S| \geq q+1$.

Similarly, the vertex $v_{e_{i,j}}$ is not in S. If not, we suppose that a vertex $v_{e_{i,j}}$ is in S. The metric representation of the corresponding $v_{p_{i,j,q+1}}$ with respect to $\{v_{e_{i,j}}\}$ is $(q+2)$, and only this vertex has the metric representation $(q+2)$ with respect to $\{v_{e_{i,j}}\}$. Then, by Proposition 4.1, the vertex $v_{p_{i,j,q+1}}$ should be in S which is a contradiction.

Now, we consider two cases for the vertex $v_{p_{i,j,1}}$:

1. more than two vertices $v_{p_{i,j,1}}$ and $v_{p_{i',j',1}}$ are in S;
2. only one vertex $v_{p_{i,j,1}}$ is in S.

In Case 1, the metric representation of the corresponding $v_{p_{i,j,2}}$ with respect to $\{v_{p_{i,j,1}}, v_{p_{i',j',1}}\}$ is $(1,3)$. Only this vertex has the metric representation $(1,3)$ with respect to $\{v_{p_{i,j,1}}, v_{p_{i',j',1}}\}$. Thus, by Proposition 4.1 again, the vertex $v_{p_{i,j,2}}$ should be in S which is a contradiction.

In Case 2, as $\{v_{p_{i,j,1}}\}$ is not a $(p-q)$-antiresolving set, there is at least a vertex $v_{S_{i'}}$ in S. Similarly, only the corresponding vertex $v_{p_{i,j,2}}$ has the metric representation $(1,3)$ with respect to $\{v_{p_{i,j,1}}, v_{S_{i'}}\}$ which leads to the same contradiction as Case 1.

To prove $|S| = q$, we first suppose $|S| < q$. Then more than $p-q$ vertices $v_{S_i} \notin S$ have

$$r(v_{S_i}|S) = (1, ..., 1).$$

Moreover, $3qn$ vertices $v_{p_{i,j,l}}$ where $l \in [1, q+1]$ have

$$r(v_{p_{i,j,l}}|S) = (l+1, ..., l+1).$$

Only vertices $v_{e_{i,j}}$ may have the metric representations with respect to S different from the above two metric representations with respect to S. However, for any $v_{e_{i,j}}$, the number of vertices having the same metric representation of $v_{e_{i,j}}$ with respect to S is an integer multiple of n. Note that

$$\lfloor \frac{p-q}{3} \rfloor < n < \lfloor \frac{p-q}{2} \rfloor$$

which means an integer multiple of n is not equal to $p-q$. Hence, S is not a $(p-q)$-antiresolving set that is a contradiction. Therefore, we have $|S| \geq q$. As $|S| \leq q$, then $|S| = q$.

By the discussion in the above proof, the metric representation of $v_{e_{i,j}}$ with respect to S should not be $(1, ..., 1)$. Futhermore, if

$$r(v_{e_{i,j}}|S) \neq (2, ..., 2),$$

we can prove that there is only one element valued 1 in $r(v_{e_{i,j}}|S)$.

If not, there are such three vertices v_a, v_b, and v_c in S that

$$r(v_{e_{i,j}}|\{v_a, v_b, v_c\}) = (1, 1, 2).$$

As S is a $(p-q)$-antiresolving set, there should be such $3n-1$ different vertices $v_{e_{i',j'}}$ that

$$r(v_{e_{i',j'}}|\{v_a, v_b, v_c\}) = (1, 1, 2).$$

Because by the assumption of n, only $3n$ is greater than $p-q$. Let $\mathcal{S}_a$ and $\mathcal{S}_b$ be the 3-element subsets in $\mathcal{S}$ corresponding to v_a and v_b. By the construction of G, we get $\mathcal{S}_a = \mathcal{S}_b$ which contradicts that no subsets in $\mathcal{S}$ are equal.

Also, we prove that $r(v_{e_{i,j}}|S) \neq (2, ..., 2)$. If not, let V_e be the set of vertices $v_{e_{i,j}}$ that are adjacent to at least one vertex in S. For a vertex $v \in V_e$, $r(v|S)$ has only one element valued 1, and there should be such $3n-1$ different vertices $v' \in V_e$ that $v' =_S v$. As $|V_e| < 3qn$, at least one vertex in S is not adjacent to any vertex $v_{e_{i,j}}$. Let $\mathcal{S}'$ be the 3-element subset in $\mathcal{S}$ corresponding to this vertex in S that has no connection to any $v_{e_{i,j}}$. Then, we have that $\mathcal{S}' = \emptyset$ which contradicts that $\mathcal{S}'$ is a 3-element subsets.

Let $\mathcal{S}_c$ be the subfamily of 3-element subsets in $\mathcal{S}$ corresponding to the vertices in S. By the three claims, we know that $\mathcal{S}_c$ is an exact cover for B. □

4.1.4 The Problem of Computing k-Metric Antidimension is NP-hard

Theorem 4.5. *The problem of computing* $\mathrm{adim}_k(G)$ *is* NP-hard.

Proof. Lemma 4.3 and 4.4 complete a polynomial-time reduction from the X3C problem to the decision version of the problem of computing $\mathrm{adim}_k(G)$. As the X3C problem is NP-complete, the decision version of the problem of computing $\mathrm{adim}_k(G)$ is also NP-complete.

Clearly, the answer of the decision version problem is true if and only if $\mathrm{adim}_k(G) \leq m$. Therefore, the problem of computing $\mathrm{adim}_k(G)$ is NP-hard. □

4.2 The Computational Complexity of the (k, l)-Anonymity Problem

Corollary 4.6. *The* (k, l)*-anonymity problem is* NP-complete.

Proof. [CDM$^+$16] shows that computing $\mathrm{adim}_1(G)$ is NP-hard. Then, we give an algorithm that can compute $\mathrm{adim}_1(G)$ in $O(|V|)$-time of calling (k, l)-anonymity.

Let $\mathcal{L}$ initialize by 1. Then, the algorithm does the following loop until $\mathcal{L} = |V| - 1$. In the loop, the algorithm checks whether G meets $(1, \mathcal{L})$-anonymity. If the answer is true, the algorithm stops the loop. Otherwise, the algorithm increases $\mathcal{L}$ by 1 and repeats the loop. After the loop, the algorithm returns $\mathcal{L}$.

By the definition of the (k, l)-anonymity problem, the return value of this algorithm is $\mathrm{adim}_1(G)$. Thus, the (k, l)-anonymity problem is NP-complete. □

Chapter 5

The k-Metric Antidimension in Random Graphs

As shown in Corollary 4.6, it is hard to get an exact relationship between k and l. To estimate the relationship between these two parameters, we study the k-metric antidimension in random graphs. We wish this study could help us characterize the trade-off between the level of anonymity and the minimum cost of achieving such level of anonymity. For $k \in O(1)$ and $k \in \omega(1)$, we establish three bounds on the size of k-antiresolving sets in $G(n,p)$ where p is constant.

5.1 The Definition of Relaxed Metric Representation

Clearly, the shortest-path distance between different pairs of vertices are not mutually independent in $G(n,p)$. Thus, the analysis of k-metric antidimension is much more difficult than the analysis of the size of cliques and independent sets in [JLR00]. To overcome the difficulty from the correlation among distances, we introduce the concept of a relaxed metric representation and establish bounds on the size of k-antiresolving sets under the relaxed metric representation. Then, we convert these bounds to the bounds on the size of k-antiresolving sets under the standard metric representation, by taking into the consideration of an observation on the diameter of the random graph $G(n,p)$.

In the relaxed metric representation, we use the relaxed shortest-path distance instead of the shortest-path distance.

Definition 5.1. Given two vertices v_i, v_j in $G(n,p)$, we define the relaxed shortest-path distance d_{ij} as

$$d_{ij} = \begin{cases} 1: & v_i \text{ and } v_j \text{ are adjacent,} \\ *: & \text{otherwise.} \end{cases}$$

5.2 The First Bound on the Size of k-Antiresolving Sets

For a set S of vertices and a vertex $v \notin S$, we denote the relaxed metric representation of v with respect to S by $r^*(v|S)$. Also, we denote the family of metric representations and relaxed metric representations with respect to S by R_S and R_S^*.

Given a $G(n,p)$ where p is constant, let

$$\begin{aligned} p_m &= \min(p, 1-p), \\ C_\alpha &= \min\big[\frac{\alpha^2}{2}, (1+\alpha)\ln(1+\alpha) - \alpha\big], \\ \epsilon &= \ln\big[\frac{(2+\beta)}{C_\alpha \ln(\frac{1}{p_m})}\ln^2(n)\big] \cdot [\ln(n)]^{-1} \end{aligned}$$

where α, β are arbitrary positive constants. We have the following theorem.

Theorem 5.2. *Given a random graph $G(n,p)$ where p is constant,* **w.h.p.** *there is no k-antiresolving set S where $k \in O(1)$ satisfying*

$$|S| \leq (1-\epsilon)\log_{\frac{1}{p_m}}(n)$$

where $p_m = \min(p, 1-p)$.

Proof. Let S be such a subset of vetices in $G(n,p)$ that

$$|S| \leq (1-\epsilon)\log_{\frac{1}{p_m}}(n).$$

Given a vertex $v \notin S$ and a relaxed metric representation $r'_* \in R_S^*$, we define

$$I_v^*(r'_*, S) = \begin{cases} 1: & r^*(v|S) = r'_*, \\ 0: & \text{otherwise} \end{cases}$$

and

$$X_{r'_*}^*(S) = \sum_{v \notin S} I_v^*(r'_*, S).$$

By the definition of $G(n,p)$, the relaxed shortest-path distances between v and vertices in S are mutually independent. Therefore,

$$\begin{aligned} &\mathbb{E}(I_v^*(r'_*, S)) = p^{\beta_1(r'_*)} \cdot (1-p)^{|S|-\beta_1(r'_*)}, \\ &\mathbb{E}(X_{r'_*}^*(S)) \geq (n - |S|) \cdot p_m^{|S|} \end{aligned}$$

where $\beta_1(r'_*)$ is the number of 1 in r'_*. By the assumptions of α and ϵ,

$$(n - |S|) \cdot p_m^{|S|} \geq n^\epsilon - |S| \cdot n^{\epsilon-1},$$
$$n^\epsilon = \frac{(2+\beta)}{C_\alpha \ln(\frac{1}{p_m})} \ln^2(n).$$

Thus,

$$\min(\mathbb{E}(X^*_{r'_*}(S))) \in \Omega(\ln^2(n)).$$

We define $E^*(r'_*, S)$ as the event

$$|X^*_{r'_*}(S) - \mathbb{E}(X^*_{r'_*}(S))| \leq \alpha \mathbb{E}(X^*_{r'_*}(S)).$$

By Chernoff Bound, see Corollary A.1.14 in [JLR00],

$$\begin{aligned}\Pr\{\tilde{} E^*(r'_*, S)\} &= \Pr\{|X^*_{r'_*}(S) - \mathbb{E}(X^*_{r'_*}(S))| \geq \alpha \mathbb{E}(X^*_{r'_*}(S))\} \\ &\leq 2e^{-C_\alpha \mathbb{E}(X^*_{r'_*}(S))}.\end{aligned}$$

Let r' be a metric representation with respect to S where the corresponding relaxed representation of r' is r'_*. We define

$$I_v(r', S) = \begin{cases} 1: & r(v|S) = r', \\ 0: & \text{otherwise} \end{cases}$$

and

$$X_{r'}(S) = \sum_{v \notin S} I_v(r', S).$$

Let $E(r', S)$ be the event

$$|X_{r'}(S) - \mathbb{E}(X^*_{r'_*}(S))| \leq \alpha \mathbb{E}(X^*_{r'_*}(S)).$$

Let $D_{\leq 2}$ be such the event that the diameter of $G(n, p)$ is less than or equal to 2. We can rewrite the event $\tilde{} E(r', S)$ as

$$(\tilde{} E(r', S) \cap D_{\leq 2}) \cup (\tilde{} E(r', S) \cap \tilde{} D_{\leq 2}).$$

Note that the event $\tilde{} E(r', S) \cap D_{\leq 2}$ is exactly the event $\tilde{} E^*(r'_*, S) \cap D_{\leq 2}$. Therefore,

$$\begin{aligned}\Pr\{\tilde{} E(r', S)\} &= \Pr\{\tilde{} E^*(r'_*, S) \cap D_{\leq 2}\} + \Pr\{\tilde{} E(r', S) \cap \tilde{} D_{\leq 2}\} \\ &\leq \Pr\{\tilde{} E^*(r'_*, S)\} + \Pr\{\tilde{} D_{\leq 2}\}.\end{aligned}$$

Note that two vertices have the shortest-path distance greater than 2 if and only if they have no common neighbor. Let X be the number of pairs of vertices in $G(n,p)$ that they have no common neighbor. By Markov's inequality,

$$\Pr\{\tilde{} D_{\leq 2}\} = \Pr\{X > 0\} \leq \mathbb{E}(X) = \binom{n}{2}(1-p^2)^{n-2}.$$

Let $E(S)$ be the event

$$\bigcap_{r' \in R_S} E(r', S).$$

By the union bound,

$$\Pr\{E(S)\} \geq 1 - n^{|S|} \cdot \left[2e^{-C_\alpha(n-|S|)\cdot p_m^{|S|}} + \binom{n}{2}(1-p^2)^{n-2}\right].$$

Note that

$$\binom{n}{k} \leq \frac{n^k}{k!} \leq \frac{n^k}{\sqrt{2\pi k}(k/e)^k} \leq \frac{n^k}{(k/e)^k} = (\frac{en}{k})^k.$$

Let

$$A = |\{S : S \subset V(G), |S| \leq (1-\epsilon)\log_{\frac{1}{p_m}}(n), \tilde{} E(S)\}|.$$

Then,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}(A) &= \lim_{n\to\infty} \binom{n}{|S|} \Pr\{\tilde{} E(S)\} \\
&\leq \lim_{n\to\infty} (\frac{en}{|S|})^{|S|} \cdot n^{|S|} \cdot \left[2e^{-C_\alpha(n-|S|)\cdot p_m^{|S|}} + \binom{n}{2}(1-p^2)^{n-2}\right] \\
&\leq \lim_{n\to\infty} 2 \cdot \left\{\frac{en^2}{|S|} \cdot \left[\frac{e^{C_\alpha n^{\epsilon-1}}}{n^{2+\beta}} + \left(\binom{n}{2}(1-p^2)^{n-2}\right)^{\frac{1}{|S|}}\right]\right\}^{|S|} \\
&= 0.
\end{aligned}$$

By Markov's inequality, we get that $\lim_{n\to\infty} \Pr\{A = 0\} = 1$. □

5.3 The Second Bound on the Size of k-Antiresolving Sets

Theorem 5.3. *Given a random graph $G(n,p)$ with constant p,* **w.h.p.** *there is no k-antiresolving set S where $k \in \omega(1)$ satisfying*

$$|S| \geq \log_{\frac{1}{2p^2-2p+1}}(n).$$

Proof. We use $I_v(r', S)$ as the definition in the above theorem, as well as p_m. Let S be such a subset of vetices in $G(n, p)$ that

$$|S| \geq \log_{\frac{1}{2p^2-2p+1}}(n).$$

We consider two cases of $|S|$:

1. $|S| \in \Theta(n)$;
2. $|S| \in o(n)$.

In Case 1,

$$\Pr\{I_v(r', S) = 1\} \leq (1 - p_m)^{|S|}.$$

Let $E(r', S)$ be the event

$$\sum_{v \notin S} I_v(r', S) \in \omega(1).$$

Thus,

$$\Pr\{E(r', S)\} \leq \binom{n}{\omega(1)}(1 - p_m)^{|S|\omega(1)}.$$

Let $E(S)$ be the event that S is a k-antiresolving set where $k \in \omega(1)$. By Fréhet inequalities,

$$\Pr\{E(S)\} = \Pr\{\bigcap_{r' \in R_S} E(r', S)\} \leq \binom{n}{\omega(1)}(1 - p_m)^{|S|\omega(1)}.$$

Let

$$A = |\{S : S \subset V(G), |S| \in \Theta(n), E(S)\}|.$$

By the notation of Θ, there are such a n_0 and a positive constant c_1 that when $n > n_0$, then $c_1 n \leq |S|$. Thus,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}(A) &= \lim_{n\to\infty} \binom{n}{|S|} \Pr\{E(S)\} \\
&\leq \lim_{n\to\infty} \left[\frac{ne}{\omega(1)(\frac{1}{1-p_m})^{\frac{|S|}{2}}}\right]^{\omega(1)} \cdot \left[\frac{ne}{|S|(\frac{1}{1-p_m})^{\frac{\omega(1)}{2}}}\right]^{|S|} \\
&\leq \lim_{n\to\infty} \left[\frac{\frac{1}{c_1}e}{(\frac{1}{1-p_m})^{\frac{\omega(1)}{2}}}\right]^{|S|} = 0
\end{aligned}$$

which leads to $\lim_{n\to\infty} \Pr\{A = 0\} = 1$.

In Case 2, we claim the following fact: if $|R_S| \in \Theta(n - |S|)$, then S is a k-antiresolving set of constant k. If not, for any $r' \in R_S$, we have

$$\sum_{v \notin S} I_v(r', S) \in \omega(1).$$

Therefore, by Proposition A.7, the number of vertices $\notin S$ belongs to

$$\omega(1) \cdot \Theta(n - |S|) \in \omega(n - |S|).$$

This conclusion contradicts that the number of vertices $\notin S$ is exactly $n-|S|$.

In Case 2, we can prove that $\mathbb{E}(|R_S^*|) \in \Theta(n - |S|))$. Let r'_* be a relaxed metric representation in R_S^*. We define $I_S(r'_*)$ as

$$I_S(r'_*) = \begin{cases} 1 : \exists v \notin S \text{ where } r^*(v|S) = r'_*, \\ 0 : \text{ otherwise.} \end{cases}$$

Then,

$$\Pr\{I_S(r'_*) = 1\} = 1 - \left[1 - p^{\beta_1(r'_*)}(1-p)^{|S|-\beta_1(r'_*)}\right]^{n-|S|}. \tag{5.1}$$

We denote

$$p^{\beta_1(r'_*)}(1-p)^{|S|-\beta_1(r'_*)}$$

by Φ. By $e^x \geq 1 + x$, we have

$$(1-x)^n \leq e^{-nx}$$

for $n > 0$. By $e^{-x} \leq 1 - x + x^2/2$ for $x > 0$, we have

$$1 - (1-x)^n \geq nx - \frac{(nx)^2}{2}$$

for $x, n > 0$. Applying this inequality to (5.1), we get

$$\Pr\{I_S(r'_*) = 1\} \geq [\Phi \cdot (n - |S|)] - \frac{[\Phi \cdot (n - |S|)]^2}{2}$$

which leads to

$$\begin{aligned}
\mathbb{E}(|R_S^*|) &\geq \sum_{\beta_1(r'_*)=0}^{|S|} \binom{|S|}{\beta_1(r'_*)} \left\{ [\Phi \cdot (n - |S|)] - \frac{[\Phi \cdot (n - |S|)]^2}{2} \right\} \\
&\geq (n - |S|) \cdot \Big(\sum_{\beta_1(r'_*)=0}^{|S|} \Phi \Big) - \frac{(n - |S|)^2}{2} \cdot \Big(\sum_{\beta_1(r'_*)=0}^{|S|} \Phi^2 \Big) \\
&= (n - |S|) - \frac{(n - |S|)^2}{2} (2p^2 - 2p + 1)^{|S|} \\
&\geq (n - |S|) - \frac{1}{2} n^2 (2p^2 - 2p + 1)^{|S|} - \frac{1}{2} |S|^2 (2p^2 - 2p + 1)^{|S|}.
\end{aligned}$$

As the assumption

$$|S| \geq \log_{\frac{1}{2p^2-2p+1}}(n),$$

we get

$$\mathbb{E}(|R^*_S|) \in \Omega(n - |S|).$$

Obviously,

$$|R^*_S| \leq n - |S|.$$

Then,

$$\mathbb{E}(|R^*_S|) \in \Theta(n - |S|)).$$

To follow Azuma-Hoeffding inequality, we define mutually independent random variables

$$Z_{v_1}, ..., Z_{v_{n-|S|}}$$

where $v_1, ..., v_{n-|S|} \notin S$ and $Z_{v_i} = r^*(v_i|S)$. Also, we define a function

$$f(Z_{v_1}, ..., Z_{v_{n-|S|}}) = |R^*_S|.$$

If two vectors $z, z' \in \prod_{i=1}^{n-|S|} r^*(v_i|S)$ are different with only one coordinate,

$$|f(z) - f(z')| \leq 1.$$

Let α be an arbitary positive constant. By Azuma-Hoeffding inequality, see Corollary 2.27 in [JLR00],

$$\Pr\{\big||R^*_S| - \mathbb{E}(|R^*_S|)\big| \geq \alpha\mathbb{E}(|R^*_S|)\} \leq 2e^{-\frac{(\alpha\mathbb{E}(|R^*_S|))^2}{2(n-|S|)}}.$$

As

$$\mathbb{E}(|R^*_S|) \in \Theta(n - |S|),$$

there are such a n_0 and a positive constant c_1 that when $n > n_0$, then

$$c_1(n - |S|) \leq \mathbb{E}(|R^*_S|).$$

Let $E(S)$ be the event

$$|R_S| \in \Theta(n - |S|)$$

and $E^*(S)$ be the event

$$|R^*_S| \in \Theta(n - |S|).$$

Clearly,

$$|R_S| \geq |R^*_S|.$$

Thus,

$$\Pr\{\tilde{} E(S)\} \leq \Pr\{\tilde{} E^*(S)\}$$

which means

$$\lim_{n\to\infty} \Pr\{\tilde{} E(S)\} \leq 2e^{-C_\alpha(n-|S|)}$$

where $C_\alpha = (\alpha c_1)^2/2$. Let

$$A = |\{S : S \subset V(G), |S| \geq \log_{\frac{1}{2p^2-2p+1}}(n), |S| \in o(n), \tilde{} E(S)\}|.$$

Then,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}(A) &= \lim_{n\to\infty} \binom{n}{|S|} \Pr\{\tilde{} E(S)\} \\
&\leq \lim_{n\to\infty} 2 \cdot \left(\frac{ne}{|S|}\right)^{|S|} e^{-C_\alpha(n-|S|)} \\
&= \lim_{n\to\infty} 2 \cdot \left[\frac{e \cdot \frac{n}{|S|}}{e^{C_\alpha(\frac{n}{|S|}-1)}}\right]^{|S|} = 0.
\end{aligned}$$

Thus, we have that $\lim_{n\to\infty} \Pr\{A = 0\} = 1$. □

5.4 The Third Bound on the Size of k-Antiresolving Sets

Theorem 5.4. *Given a random graph $G(n,p)$ with constant p,* **w.h.p.** *there is at least one k-antiresolving set S where $k \in O(1)$ satisfying*

$$|S| \geq \log_{\frac{1}{p_m}}(n)$$

where $p_m = \min(p, 1-p)$.

Proof. As shown in the above theorem, we only need to consider the case

$$|S| \in \Theta(\log_{\frac{1}{p_m}}(n)).$$

We define r_*^S as such the relaxed metric representation with respect to S that

$$\Pr\{I_v^*(r_*^S, S) = 1\} = p_m^{|S|}.$$

Let $E(S)$ be the event that S is a k-antiresolving set where $k \in O(1)$ and $E'(S)$ be the event that

$$\sum_{v\notin S} I_v^*(r_*^S, S) = c$$

where c is constant. Clearly,

$$E'(S) \subseteq E(S).$$

Let

$$A = |\{S : S \subset V(G), |S| \geq \log_{\frac{1}{p_m}}(n), |S| \in \Theta(\log_{\frac{1}{p_m}}(n)), E'(S)\}|.$$

Then,

$$\mathbb{E}(A) = \binom{n}{|S|}\Pr\{E'(S)\} = \binom{n}{|S|+c} p_m^{|S|c}(1-p_m^{|S|})^{n-|S|-c}.$$

Note that

$$\frac{n}{k} \leq \frac{n-1}{k-1}$$

for $n \geq k$ that leads to

$$\binom{n}{k} \geq \left(\frac{n}{k}\right)^k.$$

As the assumption that $p_m^{|S|} \leq 1/n$, then

$$(1-p_m^{|S|})^{n-|S|-c} \geq (1-\frac{1}{n})^{n-|S|-c} \geq (1-\frac{1}{n})^n \geq \frac{1}{4}$$

for $n \geq 2$ due to $(1-\frac{1}{n})^n$ increases as n increases. Thus,

$$\mathbb{E}(A) \geq \frac{1}{4} \cdot \left(\frac{np^c}{|S|+c}\right)^{|S|} \cdot \left(\frac{n}{|S|+c}\right)^c.$$

Then,

$$\lim_{n\to\infty} \mathbb{E}(A) = \infty.$$

To prove

$$\lim_{n\to\infty} \Pr\{A = 0\} = 0,$$

we first prove

$$\mathrm{Var}(A) \in o((\mathbb{E}(A))^2).$$

Let

$$d_m = \begin{cases} 1 : p = p_m, \\ * : \text{ otherwise.} \end{cases}$$

To apply Theorem 3.16, for two subsets of vertices S and S', we consider two cases:

1. $|S| = |S'|, S \neq S', S \cap S' \neq \emptyset$;
2. $|S| = |S'|, S \neq S', S \cap S' = \emptyset$.

In Case 1, let $y = |S \cap S'|$, v_u be a vertex $\notin S' \setminus S$, and v_t be a vertex $\in S' \setminus S$, then

$$\begin{aligned}\Pr\{d_{ut} = d_m | E'(S)\} &= \Pr\{d_{ut} = d_m, v_u \in S | E'(S)\} \\ &\quad + \Pr\{d_{ut} = d_m, v_u \notin S | E'(S)\} \\ &\leq \frac{|S|}{n - |S| + y} + p_m.\end{aligned}$$

Thus,

$$\begin{aligned}\Pr\{r^*(v|S') = r_*^{S'} | E'(S)\} &\leq \Pr\{r^*(v|(S' \setminus S)) = r_*^{(S' \setminus S)} | E'(S)\} \\ &\leq (\frac{|S|}{n - |S| + y} + p_m)^{|S'| - y}.\end{aligned}$$

Let

$$f_m(y) = \frac{|S|}{n - |S| + y} + p_m$$

and z be the number of such vertices $v \notin S \cup S'$ that

$$r^*(v|S) = r_*^S \text{ and } r^*(v|S') = r_*^{S'}.$$

Then,

$$\Pr\{E'(S')|E'(S)\} \leq \binom{n}{c - z} [f_m(y)]^{c(|S'| - y)}.$$

Thus,

$$\sum_{S,S'} \Pr\{E'(S')|E'(S)\} \leq \sum_{z=0}^{c} \sum_{y=1}^{|S|-1} \binom{n}{c - z} \binom{n}{|S| - y} [f_m(y)]^{c(|S| - y)}.$$

By changing variables z to $c - z$, y to $|S| - y$, we have

$$f_m(y) = \frac{|S|}{n - y} + p_m,$$

$$\sum_{S,S'} \Pr\{E'(S')|E'(S)\} \leq \sum_{z=0}^{c} \sum_{y=1}^{|S|-1} \binom{n}{z} \binom{n}{y} [f_m(y)]^{cy}.$$

As $y \in O(\ln(n))$, there exists such a n_0 that when $n > n_0$, then

$$\binom{n}{y}[f_m(y)]^{cy} \text{ increases as } y \text{ increases.}$$

Thus,

$$\lim_{n\to\infty} \max_y \left[\binom{n}{y}[f_m(y)]^{cy}\right] = \lim_{n\to\infty} \binom{n}{|S|-1}[f_m(|S|-1)]^{c(|S|-1)}.$$

Then, as the argument preceding in Theorem 3.16, we know

$$\sum_{S,S'} \mathrm{Cov}(S,S') \leq \mathbb{E}(A) \cdot \left\{c \cdot n^c \cdot |S| \cdot \binom{n}{|S|-1} \cdot [f_m(|S|-1)]^{c(|S|-1)}\right\}.$$

Note that

$$\begin{aligned}\binom{n}{k} &= \frac{n^k}{k!} \cdot (1-\frac{1}{n}) \cdots (1-\frac{k-1}{n}) \\ &\geq \frac{n^k}{k!}(1-\frac{k-1}{n})^{k-1}.\end{aligned}$$

As $(1-(k-1)/n)^{k-1}$ decreases as k increases, if $k \leq \sqrt{n}$, then

$$(1-\frac{k-1}{n})^{k-1} \geq (1-\frac{1}{\sqrt{n}})^{\sqrt{n}} \geq \frac{1}{4}.$$

Thus, for $k \leq \sqrt{n}$, we have

$$\binom{n}{k} \geq \frac{n^k}{4k!}.$$

Therefore,

$$\mathbb{E}(A) \geq \frac{n^{|S|+c}}{16(|S|+c)!} \cdot p_m^{c|S|}.$$

Then,

$$\begin{aligned}\lim_{n\to\infty} \frac{\sum_{S,S'} \mathrm{Cov}(S,S')}{[\mathbb{E}(A)]^2} &\leq \lim_{n\to\infty} \frac{16 \cdot c \cdot |S| \cdot (|S|+c)^{1+c} \cdot (1+\frac{|S|}{n-|S|+1} \cdot \frac{1}{p_m})^{c(|S|-1)}}{p_m^c \cdot n} \\ &= 0.\end{aligned}$$

In Case 2, let V' be the set of vertices $v \notin S'$ where $r^*(v|S') = r_*^{S'}$. Under the case $V' \cap S = \emptyset$, $E'(S)$ and $E'(S')$ are independent events. Thus, the corresponding covariance of $E'(S)$ and $E'(S')$ is 0. If $V' \cap S \neq \emptyset$, we define

$$z = |V' \cap S|.$$

Then,

$$\Pr\{E'(S')|E'(S)\} \leq \sum_{z=1}^{c} \binom{|S|}{z} \binom{n-2|S|}{c-z} [f_m(0)]^{c|S|}.$$

Therefore, in Case 2,

$$\sum_{S,S'} \mathrm{Cov}(S,S') \leq \mathbb{E}(A) \cdot \Big\{ \binom{n}{|S|} \cdot c \cdot |S|^c \cdot n^{c-1} \cdot [f_m(0)]^{c|S|} \Big\}.$$

Hence,

$$\begin{aligned}
\lim_{n\to\infty} \frac{\sum_{S,S'} \mathrm{Cov}(S,S')}{[\mathbb{E}(A)]^2} &\leq \lim_{n\to\infty} \frac{\frac{n^{|S|}}{|S|!} \cdot c \cdot |S|^c \cdot n^{c-1} \cdot [f_m(0)]^{c|S|}}{\frac{n^{|S|+c}}{16(|S|+c)!} \cdot p_m^{c|S|}} \\
&\leq \lim_{n\to\infty} \frac{16c \cdot |S|^c \cdot (|S|+c)^c \cdot [1 + \frac{|S|}{n-|S|} \cdot \frac{1}{p_m}]^{c|S|}}{n} \\
&= 0.
\end{aligned}$$

The observations from Case 1 and 2 lead to the following conclusion:

$$\mathrm{Var}(A) \in o([\mathbb{E}(A)]^2).$$

Thus, by Theorem 3.15,

$$\lim_{n\to\infty} \Pr\{A = 0\} = 0.$$

As $E'(S) \subseteq E(S)$, we know

$$\begin{aligned}
\lim_{n\to\infty} \Pr\Big\{ \bigcup_S E(S) \Big\} &\geq \lim_{n\to\infty} \Pr\Big\{ \bigcup_S E'(S) \Big\} \\
&= \lim_{n\to\infty} \Pr\{A > 0\} \\
&= 1.
\end{aligned}$$

□

5.5 A Range of k-Metric Antidimension Where $k \in O(1)$

By Theorem 5.2 and 5.4, we get the following corollary.

Corollary 5.5. *Given a random graph $G(n,p)$ with constant p,* **w.h.p.** $\mathrm{adim}_k(G)$ *where $k \in O(1)$ is between*

$$(1-\epsilon)\log_{\frac{1}{p_m}}(n) \text{ and } \log_{\frac{1}{p_m}}(n)$$

where

$$p_m = \min(p, 1-p),$$
$$\epsilon = \ln\big[\frac{(2+\beta)}{C_\alpha \ln(\frac{1}{p_m})}\ln^2(n)\big]\cdot[\ln(n)]^{-1},$$
$$C_\alpha = \min\big[\frac{\alpha^2}{2}, (1+\alpha)\ln(1+\alpha)-\alpha\big],$$
$$\alpha, \beta \text{ are arbitary positive constants.}$$

Chapter 6

Conclusion

In this thesis, we prove that the problem of computing $\text{adim}_k(G)$ is NP-hard and the (k, l)-anonymity problem is NP-complete. Also, to study the relationship between k and l, we establish three bounds on the size of k-antiresolving sets in $G(n, p)$ with constant p. With the bounds, we establish a range of k-metric antidimension where $k \in O(1)$ in random graphs $G(n, p)$ with constant p when n tends to infinity. To some extent, Corollary 5.5 shows that, when the size of an anonymized social network increases, an adversary can keep a probability which is independent on the network size to re-identify other anonymized users by adding only a few controlled vertices.

Futhermore, for applications of checking whether an anonymized social network has the potential danger to leak user's privacy, the three bounds can be seen as indicators if an adversary controls such many vertices.

For the future study, we are looking for an exact algorithm better than the brute force algorithm to find a k-antiresolving basis in a simple connected graph. Besides that, a gap of $\Theta(\ln(\ln(n)))$ exists between the first bound and the third bound. We conjecture that $\log_{\frac{1}{pm}}(n)$ is the sharp threshold for the appearance of k-antiresolving sets with constant k in $G(n, p)$. For the second bound, we also conjecture that it has a lower value.

Moreover, how to efficiently add noise (e.g., fake vertices and edges) into anonymized social networks to against privacy attacks is another interesting topic.

Bibliography

[AB09] S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, New York, NY, USA, 1st edition, 2009. → pages 12

[Ale11] K. Aleksandra. Privacy violations using microtargeted ads: A case study. *Journal of Privacy and Confidentiality*, 3(1), 2011. → pages 4

[Als11] G. Alsmeyer. *Chebyshev's Inequality*, pages 239–240. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. → pages 16

[And96] R. Anderson. A security policy model for clinical information systems. In *Proceedings of the 1996 IEEE Conference on Security and Privacy*, SP'96, pages 30–43, Washington, DC, USA, 1996. IEEE Computer Society. → pages 4

[AS00a] R. Agrawal and R. Srikant. Privacy-preserving data mining. *SIGMOD Rec.*, 29(2):439–450, 2000. → pages 5

[AS00b] N. Alon and J. Spencer. *The Probabilistic Method*. Wiley Publishing, 2th edition, 2000. → pages 17, 18, 55, 56, 57, 58, 59

[BDK07] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 181–190, New York, NY, USA, 2007. ACM. → pages 6

[BJKL04] S. Bellman, E. Johnson, S. Kobrin, and G. Lohse. International differences in information privacy concerns: A global survey of consumer. *The Information Society*, 20:313–324, 2004. → pages 4

[Buc04] J. Buchmann. *Introduction to Cryptography (2nd ed.)*. Springer, 2004. → pages 7

[CCSZ13] B. Custers, T. Calders, B. Schemer, and T. Zarsky. *Discrimination and Privacy in the Information Society*, volume 3. Springer Berlin Heidelberg, 2013. → pages 5

[CDM+16] T. Chatterjee, B. DasGupta, N. Mobasheri, V. Srinivasan, and I.G. Yero. On the computational complexities of three privacy measures for large networks under active attack. *arXiv:1607.01438*, 2016. → pages 21, 29

[CJW+06] J. Carlson, A. Jaffe, A. Wiles, Clay Mathematics Institute, and American Mathematical Society. *The Millennium Prize Problems*. American Mathematical Society, 2006. → pages 10

[CLRS09] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009. → pages 10, 11, 12, 16, 51, 52, 53

[Coo71] A. Cook. The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, STOC '71, pages 151–158, New York, NY, USA, 1971. ACM. → pages 11

[Cop04] J. Copeland. *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life: Plus the Secrets of Enigma*. Oxford University Press, 2004. → pages 7

[CP02] W. Chung and J. Paynter. Privacy issues on the internet. In *Proceedings of the 35th Hawaii International Conference on System Sciences*. IEEE Computer Society Press, 2002. → pages 4

[Dal77] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15(429-444):2–1, 1977. → pages 5

[Dev02] K. Devlin. *The Millennium Problems: The Seven Greatest Unsolved Mathematical Puzzles of Our Time*. 2002. → pages 11

[DL17] D. Danks and A. London. Algorithmic bias in autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 2017. → pages 5

[DMNS06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag. → pages 5

[Dwo06] C. Dwork. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, pages 1–12, Berlin, Heidelberg, 2006. Springer-Verlag. → pages 5

[EGS03] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pages 211–222, New York, NY, USA, 2003. ACM. → pages 5

[EMRVY13] A. Estrada-Moreno, J. Rodríguez-Velázquez, and I.G. Yero. The k-metric dimension of a graph. *ArXiv e-prints*, 2013. → pages 20

[ER59] P. Erdős and A. Rényi. On random graphs i. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959. → pages 16

[Gil59] E. Gilbert. Random graphs. *Ann. Math. Statist.*, 30:1141–1144, 12 1959. → pages 16

[GJ79] M. Garey and D. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1979. → pages 12

[GM84] S. Goldwasser and S. Micali. Probabilistic encryption. *Journal of Computer and System Science*, 28(2):270–299, 1984. → pages 8, 9

[Hea] Heartbleed [online, cited June 26, 2017]. → pages 1

[Hol09] J. Holvast. History of privacy. In *IFIP Advances in Information and Communication Technology*, volume 298. Springer, Berlin, Heidelberg, 2009. → pages 4

[JLR00] S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*. Wiley, 2000. → pages 30, 32, 36, 64, 66

[JS05] H. Jones and J. Soltren. Facebook: Threats to privacy, 2005. → pages 4

[KRR96] S. Khuller, B. Raghavachari, and A. Rosenfeld. Landmarks in graphs. *Discrete Applied Mathematics*, 70(3):217–229, 1996. → pages 20

[LB11] Z. Lin and Z. Bai. *Probability Inequalities*. Springer, Berlin, Heidelberg, 2011. → pages 19

[Les] J. Leskovec. Stanford large network dataset collection [online, cited June 2, 2017]. → pages 6

[LP87] D. Luciano and G. Prichett. Cryptology: From caesar ciphers to public-key cryptosystems. *The College Mathematics Journal*, 18:2–17, 1987. → pages 7

[LT08] K. Liu and E. Terzi. Towards identity anonymization on graphs. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 93–106, New York, NY, USA, 2008. ACM. → pages 6

[Mat] J. Mathai. History of computer cryptography and secrecy systems [online, cited May 30, 2017]. → pages 6

[McL05] D.L. McLeish. *Monte Carlo Simulation and Finance*. Wiley Finance. Wiley, 2005. → pages 60, 61, 62, 63

[McS09] F. McSherry. Privacy integrated queries. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD)*. Association for Computing Machinery, Inc., June 2009. → pages 6

[MOV01] A. Menezes, P. Oorschot, and S. Vanstone. *Handbook of Applied Cryptography (5th ed.)*. CRC Press, 2001. → pages 7

[MTRX16] S. Mauw, R. Trujillo-Rasua, and B. Xuan. *Counteracting Active Attacks in Social Network Graphs*, pages 233–248. Springer International Publishing, Cham, 2016. → pages 22

[NHP11] M. Netter, S. Herbst, and G. Pernul. Analyzing privacy in social networks - an interdisciplinary approach. In *Proc. of the Third IEEE International Conference on Social Computing Workshop on Security and Privacy in Social Networks (SPSN*

at SocialCom). IEEE Computer Society Press, Boston, USA, October 2011. → pages 4, 6

[PLZW14] W. Peng, F. Li, X. Zou, and J. Wu. A two-stage deanonymization attack against anonymized social networks. *IEEE Trans. Comput.*, 63(2):290–303, 2014. → pages 6

[PRT08] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 560–568. ACM, New York, NY, USA, 2008. → pages 5

[Riv90] R. Rivest. Handbook of theoretical computer science (vol. a). chapter Cryptography, pages 617–755. MIT Press, Cambridge, MA, USA, 1990. → pages 7

[RSA78] R. Rivest, A. Shamir, and L. Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 21(2):120–126, 1978. → pages 8

[SANT14] K. Sellami, M. Ahmed-Nacer, and P. Tiako. From social network to semantic social network in recommender system. *ArXiv e-prints*, 2014. → pages 6

[Sha49] C. Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28(4):656–715, 1949. → pages 7

[SIN] Heartbleed bug: RCMP asked revenue canada to delay news of sin thefts [online, cited June 26, 2017]. → pages 1

[SRN] A. Sokol and A. Rønn-Nielsen. *Advanced Probability*. Department of Mathematical Sciences, University of Copenhagen. → pages 60

[Sta05] W. Stallings. *Cryptography and Network Security Principles and Practices, Fourth Edition*. Prentice Hall, New Jersey, 2005. → pages 8

[Swe02] L. Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002. → pages 2

[Toc] A. Tockar. Riding with the stars: Passenger privacy in the NYC taxicab dataset [online, cited June 26, 2017]. → pages 1

[TRY16a] R. Trujillo-Rasua and I.G. Yero. Characterizing 1-metric antidimensional trees and unicyclic graphs. *The Computer Journal*, 59(8):1264–1273, 2016. → pages 22

[TRY16b] R. Trujillo-Rasua and I.G. Yero. k-metric antidimension: A privacy measure for social graphs. *Information Sciences*, 328:403–417, 2016. → pages iii, 2, 20, 21

[Tur36] Alan M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2(42):230–265, 1936. → pages 12

[WB90] S. Warren and L. Brandeis. The right to privacy. *Harvard Law Review*, 4(5):193–220, December 1890. → pages 4

[WW96] L. Willenborg and T. Waal. *Statistical Disclosure Control in Practice*, volume 111. Springer-Verlag, 1996. → pages 4

[WYLC10] X. Wu, X. Ying, K. Liu, and L. Chen. *A Survey of Privacy-Preservation of Graphs and Social Networks*, pages 421–453. Springer US, Boston, MA, 2010. → pages 6

[ZG17] C. Zhang and Y. Gao. On the complexity of k-metric antidimension problem and the size of k-antiresolving sets in random graphs. In *Proceedings of Computing and Combinatorics 23rd International Conference*, pages 555–567, Hong Kong, China, 2017. Springer International Publishing AG 2017. → pages iv

[ZP08] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, ICDE '08, pages 506–515, Washington, DC, USA, 2008. IEEE Computer Society. → pages 6

[ZPL08] B. Zhou, J. Pei, and W. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *SIGKDD Explor. Newsl.*, 10(2):12–22, 2008. → pages 6

Appendix

Appendix A

The Asymptotic Notations

The asymptotic notations describe different sets of functions which help abstract away some details of functions. We use the asymptotic notations to analyze the worst-case running time of algorithms or the behavior of functions. For example, we say an^2+bn and cn^2 belong to the same set where a, b, and c are constants. Because when n is big enough, n^2 becomes the dominated factor of these two polynomials. Below is the formal definitions of asymptotic notations in Chapter 3 of [CLRS09].

A.1 Θ-Notation

Definition A.1. [CLRS09](Θ-*notation*). For a given function $g(n)$, we denote the set of funcions by $\Theta(g(n))$ where

$$\Theta(g(n)) = \{f(n) : \text{there exist such positive constants } c_1, c_2, \text{ and } n_0 \text{ that } 0 \leq c_1 g(n) \leq f(n) \leq c_2 g(n) \text{ for all } n \geq n_0\}.$$

Because $\Theta(g(n))$ is a set, we could use $f(n) \in \Theta(g(n))$ to indicate a function $f(n)$ belongs to the set $\Theta(g(n))$. If $f(n) \in \Theta(g(n))$, we know there exist such c_1, c_2, and n_0 that $f(n)$ is sandwiched between $c_1 g(n)$ and $c_2 g(n)$. Then we can say $g(n)$ is an asymptotically tight bound for $f(n)$. Moreover, Θ-notation has the symmetry property:

$$f(n) \in \Theta(g(n)) \text{ if and only if } g(n) \in \Theta(f(n)).$$

This property can be proved directly from the definition of Θ-notation.

Proof. If $f(n) \in \Theta(g(n))$, there exist such positive constants c_1, c_2, and n_0 that

$$0 \leq c_1 g(n) \leq f(n) \leq c_2 g(n)$$

when $n \geq n_0$. Then when $n \geq n_0$, we have

$$0 \leq \frac{1}{c_2} f(n) \leq g(n) \leq \frac{1}{c_1} f(n).$$

As c_1 and c_2 are constants, $1/c_1$ and $1/c_2$ are also constants. Based on the definition of Θ-notation, we know $g(n) \in \Theta(f(n))$. □

A.2 O-Notation and Ω-Notation

When describing an asymptotic upper bound, we use O-notation.

Definition A.2. [CLRS09](O-*notation*). For a given function $g(n)$, we denote the set of funcions by $O(g(n))$ where

$$O(g(n)) = \{f(n) : \text{there exist such positive constants } c \text{ and } n_0 \text{ that } 0 \leq f(n) \leq cg(n) \text{ for all } n \geq n_0\}.$$

Just like an asymptotic upper bound, we denote an asymptotic lower bound by Ω-notation.

Definition A.3. [CLRS09](Ω-*notation*). For a given function $g(n)$, we denote the set of funcions by $\Omega(g(n))$ where

$$\Omega(g(n)) = \{f(n) : \text{there exist such positive constants } c \text{ and } n_0 \text{ that } 0 \leq cg(n) \leq f(n) \text{ for all } n \geq n_0\}.$$

From the definitions of asymptotic notations Θ, O, and Ω, we can prove the following theorem in [CLRS09].

Theorem A.4. [CLRS09] *For two functions $f(n)$ and $g(n)$, we have $f(n) \in \Theta(g(n))$ if and only if $f(n) \in O(g(n))$ and $f(n) \in \Omega(g(n))$.*

Proof. If $f(n) \in \Theta(g(n))$, we know there exist such positive constants c_1, c_2, and n_0 that

$$0 \leq c_1 g(n) \leq f(n) \leq c_2 g(n)$$

when $n \geq n_0$. Then we have $f(n) \in O(g(n))$ because there exist such positive constants c_2 and n_0 that

$$0 \leq f(n) \leq c_2 g(n)$$

when $n \geq n_0$. Similarly, as there exist such positive constants c_1 and n_0 that

$$0 \leq c_1 g(n) \leq f(n)$$

when $n \geq n_0$, we know $f(n) \in \Omega(g(n))$.

If $f(n) \in O(g(n))$ and $f(n) \in \Omega(g(n))$, there exist such positive constants c_2 and n_2 that

$$0 \leq f(n) \leq c_2 g(n)$$

when $n \geq n_2$, and there exist such positive constants c_1 and n_1 that

$$0 \leq c_1 g(n) \leq f(n)$$

when $n \geq n_1$. Let $n_0 = max(n_1, n_2)$, then

$$0 \leq c_1 g(n) \leq f(n) \leq c_2 g(n)$$

when $n \geq n_0$. □

A.3 *o*-Notation and ω-Notation

In some cases, O-notation may be not asymptotically tight. As the example in [CLRS09], $O(n^2)$ is asymptotically tight for $2n^2$, but it is not asymptotically tight for $2n$. We denote an upper bound but not an asymptotically upper bound by o-notation.

Definition A.5. [CLRS09](o-*notation*). For a given function $g(n)$, we denote the set of funcions by $o(g(n))$ where

$$o(g(n)) = \{f(n) : \text{for any positive constant } c, \text{ there exists such a positive constant } n_0 \text{ that } 0 \leq f(n) < cg(n) \text{ for all } n \geq n_0\}.$$

If $f(n) \in o(g(n))$, the function $f(n)$ becomes insignificant compared with $g(n)$ when n tends to infinity. By the definition of a limit in preliminary calculus, we get

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0.$$

Similarly, we use ω-notation to denote a lower bound but not an asymptotically lower bound.

Definition A.6. [CLRS09](ω-*notation*). For a given function $g(n)$, we denote the set of funcions by $\omega(g(n))$ where

$$\omega(g(n)) = \{f(n) : \text{for any positive constant } c, \text{ there exists such a positive constant } n_0 \text{ such that } 0 \leq cg(n) < f(n) \text{ for all } n \geq n_0\}.$$

If $f(n) \in \omega(g(n))$, we know

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = \infty.$$

The asymptotical notations have the following property.

Proposition A.7. *If a function* $f(n) = f_1(n) \cdot f_2(n)$ *where* $f_1(n) \in \omega(1)$ *and* $f_2(n) \in \Theta(g(n))$*, then* $f(n) \in \omega(g(n))$*.*

Proof. By $f_2(n) \in \Theta(g(n))$, we have that there exist such positive constants c_1 and n_1 that when $n \geq n_1$, $f_2(n) \geq c_1 \cdot g(n)$. For any positive constant c, we can get a new constant $c_2 = c/c_1$. Then by $f_1(n) \in \omega(1)$, there exists such a positive constant n_2 that when $n \geq n_2$, $f_1(n) > c_2$.

Let $n_0 = max(n_1, n_2)$. Then, for any positive constant c there exists such a positive constant n_0 that when $n \geq n_0$,

$$f(n) = f_1(n) \cdot f_2(n) > c \cdot g(n).$$

By the definition of ω-notation, we have $f(n) \in \omega(g(n))$. □

Appendix B

The Probabilistic Method

In this chapter, we introduce Chernoff Bound and Azuma-Hoeffding Inequality and give the proofs.

B.1 Chernoff Bound

We suppose that the distribution X follows the following assumption: $X_1, ..., X_n$ are mutually independent indicator random variable with

$$\Pr[X_i = 1 - p_i] = p_i$$

$$\Pr[X_i = -p_i] = 1 - p_i$$

where

$$p_1, ..., p_n \in [0, 1].$$

We set

$$p = \frac{p_1 + ... + p_n}{n}$$
$$X = X_1 + ... + X_n.$$

Lemma B.1. [AS00b] *For real numbers α, β where $|\alpha| \leq 1$,*

$$\cosh(\beta) + \alpha \sinh(\beta) \leq e^{\beta^2/2 + \alpha\beta}.$$

Proof. Clearly, if $\alpha = 1$ or $\alpha = -1$, the above inequality is true. Note that

$$e^{\beta^2/2 - |\beta|} \leq e^{\beta^2/2 + \alpha\beta}$$

and

$$\cosh(\beta) + \alpha \sinh(\beta) \leq 2e^{|\beta|}.$$

Thus, if

$$|\beta| \geq 100,$$

the inequality is true.

Also, we know that

$$f(0,\beta) \leq 0,$$
$$f(\alpha,0) = 0$$

where

$$f(\alpha,\beta) = \cosh(\beta) + \alpha\sinh(\beta) - e^{\beta^2/2+\alpha\beta}.$$

We suppose that when $|\alpha| \leq 1, \alpha \neq 0$ and $|\beta| \leq 100, \beta \neq 0$, there exist such α, β that $f(\alpha,\beta) > 0$. Therefore, $f(\alpha,\beta)$ has a positive global maximum in the range

$$R = \{(\alpha,\beta)||\alpha| \leq 1, \alpha \neq 0, |\beta| \leq 100, \beta \neq 0\}.$$

Setting the partial derivatives equal to zero, we get

$$\frac{\partial f(\alpha,\beta)}{\partial\alpha} = \sinh(\beta) - \beta e^{\beta^2/2+\alpha\beta} = 0$$

$$\frac{\partial f(\alpha,\beta)}{\partial\beta} = \sinh(\beta) + \alpha\cosh(\beta) - (\alpha+\beta)e^{\beta^2/2+\alpha\beta} = 0.$$

Thus,

$$\tanh(\beta) = \beta$$

that means

$$\beta = 0$$

which is a contradiction. □

Corollary B.2. [AS00b] *For real numbers θ, λ where $\theta \in [0,1]$,*

$$\theta e^{\lambda(1-\theta)} + (1-\theta)e^{-\lambda\theta} \leq e^{\lambda^2/8}.$$

Proof. Setting

$$\theta = \frac{1+\alpha}{2}$$

and

$$\lambda = 2\beta,$$

we get Corollary B.2 from Lemma B.1. □

Theorem B.3. [AS00b] *For positive real numbers a,*

$$\Pr\{|X| \geq a\} \leq 2e^{-2a^2/n}.$$

Proof. From Corollary B.2, we know

$$\mathbb{E}(e^{\lambda X_i}) = p_i e^{\lambda(1-p_i)} + (1-p_i)e^{-\lambda p_i} \leq e^{\lambda^2/8}.$$

Thus,

$$\mathbb{E}(e^{\lambda X}) = \prod_{i=1}^{n} \mathbb{E}(e^{\lambda X_i}) \leq e^{\lambda^2 n/8}.$$

For $\lambda > 0$, we know

$$\Pr\{X \geq a\} = \Pr\{e^{\lambda X} \geq e^{\lambda a}\}.$$

Applying Markov's inequality, we get

$$\Pr\{e^{\lambda X} \geq e^{\lambda a}\} \leq \frac{\mathbb{E}(e^{\lambda X})}{e^{\lambda a}} \leq e^{\lambda^2 n/8 - \lambda a}.$$

Setting $\lambda = 4a/n$ to optimize the inequality, we get

$$\Pr\{X \geq a\} \leq e^{-2a^2/n}.$$

By symmetry, we get

$$\Pr\{X \leq -a\} \leq e^{-2a^2/n}.$$

□

Lemma B.4. [AS00b]

$$\mathbb{E}(e^{\lambda X}) \leq e^{-\lambda pn}[pe^{\lambda} + (1-p)]^n.$$

Proof.

$$\begin{aligned}\mathbb{E}(e^{\lambda X}) &= \prod_{i=1}^{n} \mathbb{E}(e^{\lambda X_i}) = \prod_{i=1}^{n} [p_i e^{\lambda(1-p_i)} + (1-p_i)e^{-\lambda p_i}] \\ &= e^{-\lambda pn} \prod_{i=1}^{n} [p_i e^{\lambda} + (1-p_i)].\end{aligned}$$

With λ fixed, the function

$$f(x) = \ln(xe^{\lambda} + 1 - x) = \ln[x(e^{\lambda} - 1) + 1]$$

is concave. Thus, by Jensen's Inequality,

$$\sum_{i=1}^{n} f(p_i) \leq nf(p).$$

Exponentiating both sides, we have

$$\prod_{i=1}^{n}[p_i e^{\lambda} + (1 - p_i)] \leq [pe^{\lambda} + (1 - p)]^n.$$

□

Corollary B.5. [AS00b] *For positive real numbers* a, λ*,*

$$\Pr\{X \geq a\} \leq e^{-\lambda pn}[pe^{\lambda} + (1 - p)]^n e^{-\lambda a}.$$

Proof.

$$\Pr\{X \geq a\} = \Pr\{e^{\lambda X} \geq e^{\lambda a}\}.$$

By Markov's inequality,

$$\Pr\{e^{\lambda X} \geq e^{\lambda a}\} \leq \frac{\mathbb{E}(e^{\lambda X})}{e^{\lambda a}}.$$

Now applying Lemma B.4, we get Corollary B.5. □

Setting

$$\lambda = \ln(1 + a/pn),$$

we get Corollary B.6 by using the fact that

$$e^a = e^{(a/n)n} \geq (1 + a/n)^n,$$

Corollary B.6.

$$\Pr\{X \geq a\} \leq e^{a - pn\ln(1+a/pn) - a\ln(1+a/pn)}.$$

Plugging $a = (\beta - 1)pn$ into Corollary B.6, we get

Corollary B.7.

$$\Pr\{X \geq (\beta - 1)pn\} \leq (e^{\beta - 1}\beta^{-\beta})^{pn}.$$

Lemma B.8. *For positive real numbers* a*,*

$$\Pr\{X \leq -a\} \leq e^{-a^2/2pn}.$$

Proof. Let $\lambda > 0$. By the argument preceding in Lemma B.4, we get

$$\mathbb{E}(e^{-\lambda X}) \leq e^{\lambda pn}[pe^{-\lambda} + (1-p)]^n.$$

Thus,

$$\Pr\{X \leq -a\} \leq e^{\lambda pn}[pe^{-\lambda} + (1-p)]^n e^{-\lambda a}.$$

We apply the inequality

$$1 + \mu \leq e^{\mu}$$

valid for all μ, and then we get

$$pe^{-\lambda} + (1-p) = 1 + p(e^{-\lambda} - 1) \leq e^{p(e^{-\lambda}-1)}.$$

Thus,

$$\Pr\{X \leq -a\} \leq e^{\lambda pn + np(e^{-\lambda}-1) - \lambda a} = e^{np(e^{-\lambda}-1+\lambda) - \lambda a}.$$

We employ the inequality

$$e^{-\lambda} \leq 1 - \lambda + \lambda^2/2$$

for $\lambda > 0$. Therefore,

$$\Pr\{X \leq -a\} \leq e^{\frac{np\lambda^2}{2} - \lambda a}.$$

We set $\lambda = a/np$ to optimize the above inequality and get

$$\Pr\{X \leq -a\} \leq e^{-a^2/2pn}$$

as claimed. □

Note that $Y = X + pn$ can be interpreted as the number of successes in n independent trials when the probability of success in the i-th trial is p_i. Clearly,

$$\mathbb{E}(X_i) = \mathbb{E}(X) = 0.$$

Thus,

$$\mathbb{E}(Y) = \mathbb{E}(X) + np = np.$$

Then, we can get the following result.

Theorem B.9. [AS00b] *Let Y be the sum of mutually independent indicator random variables, $\mu = \mathbb{E}(Y)$. For all constants $\epsilon > 0$,*

$$\Pr\{|Y - \mu| > \epsilon\mu\} < 2e^{-c_\epsilon \mu},$$

where c_ϵ is a constant only depending on ϵ.

Proof. Applying Corollary B.7 and Lemma B.8 with

$$c_\epsilon = \min(-\ln(e^\epsilon(1+\epsilon)^{-(1+\epsilon)}), \epsilon^2/2),$$

we get Theorem B.9. □

B.2 Azuma-Hoeffding Inequality

Given a probability space $(\Omega, \mathcal{F}, P)$, we define $\mathcal{G}$ as a subset of $\mathcal{F}$ and suppose X as a random variable on $\mathcal{F}$. There are two equivalent versions of the definition of the conditional expectation of X with respect to $\mathcal{G}$ and denoted by $\mathbb{E}(X|\mathcal{G})$.

Definition B.10. [SRN] If

1. Y is measurable to $\mathcal{G}$ and
2. $\mathbb{E}(YI_A) = \mathbb{E}(XI_A)$ for all $A \in \mathcal{G}$,

then

$$Y = \mathbb{E}(X|\mathcal{G}).$$

Definition B.11. [McL05] Assume $\mathbb{E}(X^2) < \infty$. Then, a $\mathcal{G}$-measurable random variable Y is the conditional expectation of X with respect to $\mathcal{G}$ if

$$\mathbb{E}[(X - Y)^2] = \inf_{Z} \mathbb{E}(X - Z)^2$$

where the infimum (infimum=greatest lower bound) is over all $\mathcal{G}$-measurable random variables.

Theorem B.12. [McL05] *There exists an almost surely unique* $\mathbb{E}(X|\mathcal{G})$.

We do not show the proof of the above theorem in here; in the following Appendix B.2, we show some properties of $\mathbb{E}(X|\mathcal{G})$.

Proposition B.13. [McL05] *If* X *is* $\mathcal{G}$*-measurable, then* $\mathbb{E}(X|\mathcal{G}) = X$.

Proof. Note that

$$\mathbb{E}(X - Z)^2 \geq \mathbb{E}(X - X)^2 = 0.$$

Then, the minimizing Z is X. □

Proposition B.14. [McL05] *If* $\mathcal{G} = \{\emptyset, \Omega\}$, *then* $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$.

Proof. As $\mathcal{G} = \{\emptyset, \Omega\}$, we know that only a constant random variable is measurable with respect to $\mathcal{G}$. Let c be a constant, minimizing

$$\mathbb{E}[(X - c)^2] = \mathrm{Var}(X) + [\mathbb{E}(X) - c]^2$$

leads to $c = \mathbb{E}(X)$. □

Proposition B.15. [McL05] *For any square integrable $\mathcal{G}$-measurable random variable Z,*

$$\mathbb{E}(ZX) = \mathbb{E}(Z\mathbb{E}(X|\mathcal{G})).$$

Proof. We define a function of λ by

$$g(\lambda) = \mathbb{E}[(X - \mathbb{E}(X|\mathcal{G}) - \lambda Z)^2].$$

By Definition B.11, at $\lambda = 0$, $g(\lambda)$ is minimized at all real values of λ. Thus, $g'(0) = 0$. Plugging $\lambda = 0$ into the equation $g'(\lambda) = 0$, we get

$$\mathbb{E}[Z(X - \mathbb{E}(X|\mathcal{G})] = 0$$

which leads to

$$\mathbb{E}(ZX) = \mathbb{E}[Z\mathbb{E}(X|\mathcal{G})].$$

□

Setting $Z = 1$, we get

Proposition B.16. [McL05]

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})).$$

Proposition B.17. [McL05] *If a $\mathcal{G}$-measurable random variable Z satisfies $\mathbb{E}[(X - Z)Y] = 0$ for all other $\mathcal{G}$-measurable random variables Y, then*

$$Z = \mathbb{E}(X|\mathcal{G}).$$

Proof. As Z satisfies $\mathbb{E}[(X - Z)Y] = 0$ for all other $\mathcal{G}$-measurable random variables Y, we consider $Y = \mathbb{E}(X|\mathcal{G}) - Z$.

We define a function of λ

$$\begin{aligned} g(\lambda) &= \mathbb{E}[(X - Z - \lambda Y)^2] \\ &= \mathbb{E}[(X - Z)^2 - 2\lambda\mathbb{E}[(X - Z)Y] + \lambda^2\mathbb{E}(Y^2)] \\ &= \mathbb{E}((X - Z)^2) + \lambda^2\mathbb{E}(Y^2) \\ &\geq \mathbb{E}((X - Z)^2) = g(0). \end{aligned}$$

We know

$$g(1) = \mathbb{E}[(X - \mathbb{E}(X|\mathcal{G}))^2]$$

should be the minimum of $g(\lambda)$ which means $g(0) = g(1)$. The uniqueness shown in Theorem B.12 leads to $Z = \mathbb{E}(X|\mathcal{G})$. □

Proposition B.18. [McL05] *If Z is $\mathcal{G}$-measurable, then*

$$\mathbb{E}(ZX|\mathcal{G}) = Z\mathbb{E}(X|\mathcal{G}).$$

Proof. Let Y be an arbitary $\mathcal{G}$-measurable random variable. As Y and Z are $\mathcal{G}$-measurable random variables, ZY is $\mathcal{G}$-measurable. By Proposition B.15,

$$\mathbb{E}(ZYX) = \mathbb{E}[ZY\mathbb{E}(X|\mathcal{G}))]$$

which means

$$\mathbb{E}[(ZX - Z\mathbb{E}(X|\mathcal{G}))Y] = 0.$$

Thus, by Proposition B.17,

$$\mathbb{E}(ZX|\mathcal{G}) = Z\mathbb{E}(X|\mathcal{G}).$$

□

Proposition B.19. [McL05]

$$\mathbb{E}(X+Y|\mathcal{G}) = \mathbb{E}(X|\mathcal{G}) + \mathbb{E}(Y|\mathcal{G}),$$
$$\mathbb{E}(cX+d|\mathcal{G}) = c\mathbb{E}(X|\mathcal{G}) + d \text{ where } c \text{ and } d \text{ are constants}.$$

Proof. Let Z be an arbitary $\mathcal{G}$-measurable random variable. By Proposition B.15,

$$\begin{aligned}\mathbb{E}[Z(X+Y-\mathbb{E}(X|\mathcal{G})-\mathbb{E}(Y|\mathcal{G}))] &= \mathbb{E}[Z(X-\mathbb{E}(X|\mathcal{G}))] - E[Z(Y-\mathbb{E}(Y|\mathcal{G}))] \\ &= 0 - 0 = 0\end{aligned}$$

Then, by Proposition B.17,

$$\mathbb{E}(X+Y|\mathcal{G}) = \mathbb{E}(X|\mathcal{G}) + \mathbb{E}(Y|\mathcal{G}).$$

With the similar argument, we can prove

$$\mathbb{E}(cX+d|\mathcal{G}) = c\mathbb{E}(X|\mathcal{G}) + d.$$

□

Proposition B.20. [McL05] *If $\mathcal{H} \subset \mathcal{G}$, then*

$$\mathbb{E}[\mathbb{E}(X|\mathcal{G})|\mathcal{H}] = \mathbb{E}(X|\mathcal{H}).$$

Proof. Let Z be an arbitary $\mathcal{H}$-measurable random variable. As $\mathcal{H} \subset \mathcal{G}$, then Z, $\mathbb{E}(X|\mathcal{H})$, and $\mathbb{E}[\mathbb{E}(X|\mathcal{G})|\mathcal{H}]$ are $\mathcal{G}$-measurable.

$$\begin{aligned}\mathbb{E}[Z(\mathbb{E}(X|\mathcal{H}) - \mathbb{E}[\mathbb{E}(X|\mathcal{G})|\mathcal{H}])] &= \mathbb{E}[Z\mathbb{E}(X|\mathcal{H}) - Z\mathbb{E}[\mathbb{E}(X|\mathcal{G})|\mathcal{H}]] \\ &= \mathbb{E}[\mathbb{E}(ZX|\mathcal{H}) - \mathbb{E}[Z\mathbb{E}(X|\mathcal{G})|\mathcal{H}]] \\ &= \mathbb{E}[\mathbb{E}(ZX - Z\mathbb{E}(X|\mathcal{G})|\mathcal{H})] \\ &= \mathbb{E}(ZX - Z\mathbb{E}(X|\mathcal{G})) \\ &= 0.\end{aligned}$$

Then, by Proposition B.13 and B.17,

$$\begin{aligned}\mathbb{E}[\mathbb{E}(X|\mathcal{H})|\mathcal{G}] &= \mathbb{E}[\mathbb{E}(X|\mathcal{G})|\mathcal{H}], \\ \mathbb{E}[\mathbb{E}(X|\mathcal{H})|\mathcal{G}] &= \mathbb{E}(X|\mathcal{H}).\end{aligned}$$

□

Proposition B.21. [McL05] *If $X \leq Y$, then $\mathbb{E}(X|\mathcal{G}) \leq \mathbb{E}(Y|\mathcal{G})$.*

Proof. We suppose that there exist such $\omega \in \Omega$ that

$$\{\omega : \mathbb{E}(Y(\omega) - X(\omega)|\mathcal{G}) < 0\} \subset \mathcal{G}.$$

Let

$$\epsilon > \max_{\omega}(\mathbb{E}(Y(\omega) - X(\omega)|\mathcal{G}))$$

where

$$\omega \in \{\omega : \mathbb{E}(Y(\omega) - X(\omega)|\mathcal{G}) < 0\}$$

be a negative constant. Then, for those ω, we give a new assignment of $\mathbb{E}(Y(\omega) - X(\omega)|\mathcal{G})$ where

$$\mathbb{E}(Y(\omega) - X(\omega)|\mathcal{G}) = \epsilon.$$

Note that $\mathbb{E}(Y(\omega) - X(\omega)|\mathcal{G})$ is still $\mathcal{G}$-measurable. Then, we get a lower

$$\mathbb{E}[(Y - X) - \mathbb{E}(Y - X|\mathcal{G})]^2$$

which is a contradiction. □

In the rest of Appendix B.2, we introduce Azuma-Hoeffding inequality. We begin by the definition of a martingale.

Definition B.22. [JLR00] Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an increasing sequence of sub-ω-fields

$$\mathcal{F}_0 = \{\emptyset, \omega\} \subseteq \mathcal{F}_1 \subseteq ... \subseteq \mathcal{F}_n,$$

a sequence of random variables $X_0, X_1, ..., X_n$ (with finite expections) is called a martingale if for each $k = 0, 1, ..., n-1, \mathbb{E}(X_{k+1}|\mathcal{F}_k) = X_k$.

Proposition B.23. [JLR00] $\mathbb{E}(X_{k+1}) = \mathbb{E}(X_k)$.

Proof. Note that X_k is $\mathcal{F}_k$-measurable,

$$\mathbb{E}(X_k) = \mathbb{E}[\mathbb{E}(X_k|\mathcal{F}_k)].$$

As $\mathbb{E}(X_{k+1}|\mathcal{F}_k) = X_k$,

$$\mathbb{E}(X_k) = \mathbb{E}[\mathbb{E}(X_{k+1}|\mathcal{F}_k)].$$

Thus,

$$\mathbb{E}[\mathbb{E}(X_{k+1} - X_k|\mathcal{F}_k)] = 0.$$

By Proposition B.16, we get

$$\mathbb{E}(X_{k+1} - X_k) = 0.$$

□

Theorem B.24. [JLR00] *If $(X_k)_0^n$ is a martingale with $X_n = X$ and $X_0 = \mathbb{E}(X)$, and there exist such constants $c_k > 0$ that*

$$|X_k - X_{k-1}| \leq c_k$$

for each $k \leq n$, then, for every $t > 0$,

$$\Pr\{X \geq \mathbb{E}(X) + t\} \leq \exp(-\frac{t^2}{2\sum_{k=1}^n c_k^2}),$$
$$\Pr\{X \leq \mathbb{E}(X) - t\} \leq \exp(-\frac{t^2}{2\sum_{k=1}^n c_k^2}).$$

Proof. We set

$$Y_k = X_k - X_{k-1},$$
$$S_k = \sum_{i=1}^{k} Y_i = X_k - X_0.$$

For any $u > 0$, by Markov's inequality,

$$\Pr\{X \geq \mathbb{E}(X) + t\} = \Pr\{S_n \geq t\} \leq e^{-ut}\mathbb{E}(e^{uS_n}).$$

Because S_{n-1} is a $\mathcal{F}_{n-1}$-measurable random variable,

$$\mathbb{E}(e^{uS_n}) = \mathbb{E}[\mathbb{E}(e^{uS_n}|\mathcal{F}_{n-1})] = \mathbb{E}[e^{uS_{n-1}}\mathbb{E}(e^{uY_n}|\mathcal{F}_{n-1})]. \tag{B.1}$$

Now we use the fact: if a random variable Y satisfies $|Y| \leq a$ for some positive a, then, for any u, by the convexity of e^{uY},

$$e^{uY} \leq \frac{a+Y}{2a}e^{ua} + \frac{a-Y}{2a}e^{-ua} = \cosh(ua) + \frac{Y}{a}\sinh(ua).$$

Hence, by $\cosh(\beta) \leq e^{\beta^2/2}$,

$$e^{uY} \leq e^{u^2a^2/2} + \frac{Y}{a}\sinh(ua).$$

Thus,

$$\mathbb{E}(e^{uY_n}|\mathcal{F}_{n-1}) \leq e^{u^2{c_n}^2/2} + \frac{\mathbb{E}(Y_n|\mathcal{F}_{n-1})}{c_n}\sinh(uc_n).$$

By the definition of a martingale,

$$\begin{aligned}\mathbb{E}(Y_n|\mathcal{F}_{n-1}) &= \mathbb{E}(X_n - X_{n-1}|\mathcal{F}_{n-1}) \\ &= \mathbb{E}(X_n|\mathcal{F}_{n-1}) - \mathbb{E}(X_{n-1}|\mathcal{F}_{n-1}) \\ &= \mathbb{E}(X_n|\mathcal{F}_{n-1}) - X_{n-1} \\ &= 0.\end{aligned}$$

Thus,

$$\mathbb{E}(e^{uY_n}|\mathcal{F}_{n-1}) \leq e^{u^2{c_n}^2/2}.$$

Plugging it back to (B.1), we get

$$\mathbb{E}(e^{uS_n}) \leq e^{u^2{c_n}^2/2}\mathbb{E}(e^{uS_{n-1}}).$$

Iterating this inequality n times, we get

$$\mathbb{E}(e^{uS_n}) \leq e^{u^2\sum_{i=1}^n {c_i}^2/2}.$$

Thus,

$$\Pr\{X \geq \mathbb{E}(X) + t\} \leq e^{-ut}e^{u^2\sum_{i=1}^n {c_i}^2/2}.$$

Setting $u = t/\sum_{i=1}^{n} {c_i}^2$, we get

$$\Pr\{X \geq \mathbb{E}(X) + t\} \leq \exp(-\frac{t^2}{2\sum_{k=1}^{n} c_k^2}).$$

By the symmetry, we get

$$\Pr\{X \leq \mathbb{E}(X) - t\} \leq \exp(-\frac{t^2}{2\sum_{k=1}^{n} c_k^2}).$$

□

Corollary B.25. [JLR00] *Let $Z_1, ..., Z_N$ be independent random variables, with Z_k taking values in a set Λ_k. Assume a function f :*

$$\Lambda_1 \times \Lambda_2 \times ... \times \Lambda_N \to \mathbb{R}$$

satisfies the following condition for some constants c_k :

if two vectors $z, z' \in \prod_{i=1}^{N} \Lambda_i$ differ only in the kth coordinate, then $|f(z) - f(z')| \leq c_k$.

Then, the random variable $X = f(Z_1, ...Z_N)$ satisfies, for any $t \geq 0$,

$$\Pr\{X \geq \mathbb{E}(X) + t\} \leq \exp(-\frac{t^2}{2\sum_{k=1}^{n} c_k^2}),$$
$$\Pr\{X \leq \mathbb{E}(X) - t\} \leq \exp(-\frac{t^2}{2\sum_{k=1}^{n} c_k^2}).$$

Proof. Let $\mathcal{F}_k$ be the σ-fields generated by $Z_1, ...Z_K$. We define $X_k = \mathbb{E}(f(Z_1, ...Z_N)|\mathcal{F}_k), k = 0, 1, ..., N$. Then,

$$\begin{aligned}
\mathbb{E}(X_{k+1}|\mathcal{F}_k) &= \mathbb{E}(\mathbb{E}(f(Z_1, ...Z_N)|\mathcal{F}_{k+1})|\mathcal{F}_k) \\
&= \mathbb{E}(\mathbb{E}(f(Z_1, ...Z_N)|\mathcal{F}_k)|\mathcal{F}_{k+1}) \\
&= \mathbb{E}(X_k|\mathcal{F}_{k+1}) \\
&= X_k.
\end{aligned}$$

Thus, $(X_k)_0^N$ is a martingale, and $X_0 = \mathbb{E}(X), X_N = X$. To apply Theorem B.24, we should prove that $|X_{k+1} - X_k| \leq c_{k+1}$. Let Z'_{k+1} be an independent

copy of Z_{k+1} and $X' = f(Z_1, ..., Z'_{k+1}, ..., Z_N)$. Then,

$$\begin{aligned}
|X_{k+1} - X_k| &= |\mathbb{E}(X|\mathcal{F}_{k+1}) - \mathbb{E}(X|\mathcal{F}_k)| \\
&= |\mathbb{E}(X|\mathcal{F}_{k+1}) - \mathbb{E}(X'|\mathcal{F}_k)| \text{ (by } \mathbb{E}(X|\mathcal{F}_k) = \mathbb{E}(X'|\mathcal{F}_k)) \\
&= |\mathbb{E}(X|\mathcal{F}_{k+1}) - \mathbb{E}(X'|\mathcal{F}_{k+1})| \text{ (by } \mathbb{E}(X'|\mathcal{F}_k) = \mathbb{E}(X'|\mathcal{F}_{k+1})) \\
&= |\mathbb{E}(X - X'|\mathcal{F}_{k+1})| \\
&\leq c_{k+1} \text{ (by } |X - X'| \leq c_{k+1})
\end{aligned}$$

The corollary now follows Theorem B.24. □