Geodesic-Based Properties in Complex Networks

by

Nasim Mobasheri
B.S., Sharif University of Technology, Tehran, Iran, 2012

Thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2018

Chicago, Illinois

Defense Committee:
Bhaskar Dasgupta, Chair and Advisor
Robert Sloan
Reka Albert
Bing Liu
Piotr Gmytrasiewicz

To my parents,

and my husband.

ACKNOWLEDGMENTS

I would like to express gratitude to my PhD advisor, Professor Bhaskar Dasgupta for guiding me through past six years in my journey as a PhD student. Without his supervision the achievements and completion of this thesis was not possible.

I would also like to thank the members of my defense committee, Professor Robert Sloan, Professor Reka Albert from Pennsylvania State University, Professor Bing Liu, and Professor Piotr Gmytrasiewicz for their valuable advice and support.

PREFACE

This thesis is based on the following publications:

- Albert, Rèka, Bhaskar DasGupta, and Nasim Mobasheri. *Topological implications of negative curvature for biological and social networks.* Physical Review E 89.3 (2014): 032811.
- DasGupta, Bhaskar, and Nasim Mobasheri. *On optimal approximability results for computing the strong metric dimension.* Discrete Applied Mathematics 221 (2017): 18-24.
- Bhaskar DasGupta, Nasim Mobasheri and Ismael G. Yero, *On analyzing and evaluating privacy measures for social networks under active attack*, 2017, [under submission].

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

sdim	Strong Metric Dimension
UGC	Unique Game Conjecture
ETH	Exponential Time Hypothesis
MNC	Minimum Node Cover
diam	Diameter

SUMMARY

In the modern era and age of the Internet, complex networks are part of people's everyday life. From social networks revolutionizing social behavior, marketing, and information diffusion to biological networks and their valuable influence in modern day medicine and biology, to traffic network and aviation paths, these phenomena are playing increasingly important roles in our life. It is natural that the scientific community study and analyze these networks to gain better understanding about their structure and behavior. As a result, network measures that reflect the most salient properties of complex large-scale networks are in high demand in the network research community.

In this thesis, we look into three geodesic-based measures in complex networks. In the first part, we adapt a combinatorial measure of negative curvature (also called hyperbolicity) to parameterized finite networks, and show that a variety of biological and social networks are hyperbolic. This property has strong implications on the higher-order connectivity and other topological properties of these networks.

In the second part, we look into the complexity of another geodesic-based property known as strong metric dimension. We show the problem of calculating the strong metric dimension of a graph with n nodes admits polynomial-time 2-approximation, admits a $O^*(2^{0.287n})$-time exact computation algorithm, admits a $O(1.2738^k + nk)$-time exact computation algorithm if the strong metric dimension is at most k. We also prove three inapproximability results for calculating the strong metric dimension of a graph.

In the final part of this thesis, we investigate a geodesic-based property closely related to strong metric dimension, known as (k, ℓ)-anonymity which indicates the privacy violation in large–networks under active attacks. Our theoretical result provides some insight regarding prevention of privacy violation and designing topology of networks. Our empirical results shed light on privacy violation properties of real social networks as well as a large number of synthetic networks generated by both the classical Erdös-Rényi model and the scale-free random networks generated by the Barabasi-Albert preferential-attachment model.

CHAPTER 1

INTRODUCTION

Graphs are the best and most powerful tools to model and study network-like systems. Graph theory, introduced by Euler in 1736, is a branch of mathematics that studies the properties of pairwise relations in network structures. Despite its continuous growth through the years, it was only bounded by limited application and covered small group of graphs. Progress in different fields led to the discovery of larger and more complicated networks, and the availability of computers and development in computational techniques allowed scientists to gather and analyze large-scale networks in the real world. The demand for analyzing the structure of these networks expanded the graph theory and tied it to statistical physics, while new models were created to reflect the properties of these newly discovered networks. These networks with non-trivial topological features became known as complex networks, and the field of network science was formed to study and analyze these structures. Complex networks can demonstrate a wide variety of phenomenon for different disciplines in natural and social sciences. Metabolic networks, signaling networks, food webs, Internet, World Wide Web, power grids, neural networks, and social networks such as Facebook and Twitter are all examples of complex networks. The importance of studying complex networks lies in the simple fact that predicting network performance, behavior, robustness, and scalability requires an understanding of the underlying structure of the network. Complex network analysis and networks science is an interdisciplinary field of algorithms and methods developed based on graph theory and statistical physics which has received a lot of attention recently. On one hand, the emergence of social networks like Facebook and Twitter caused a revolution in the flow of information and news, and started to influence and shape users behavior and actions. This led to concerns over privacy and information flow and how these networks are affecting our behavior as humans on many aspects.

On the other hand, the successful mapping of many biological networks helped scientists to gain better

understanding of many biological phenomena and diseases and opened a path for discovering new solutions once deemed unsolvable problems in medical fields.

1.1 Study of Complex Networks

Since the 1950s, complex network were described by random graphs as the simplest and most straightforward model. Paul Erdös and Alfred Rènyi, who were the first mathematicians that studied random graphs, introduced the ER model for displaying and analyzing complex networks. This model, which is still widely in use today, starts with n nodes and connects every pair of nodes with probability p. The computerization of data acquisition and a noticeable increase in computing power allowed us to study giant networks with millions of nodes and build large databases on their topology. This spectacular progress in network science in the past few years raised a lot of questions about the random nature of real world networks and enlightened the fact that the ER model cannot capture many of the remarkable aspects of these networks.
These developments led to proposing new concepts, measures, and a fair amount of investigation on networks. Among them three concepts received the most attention: Watts and Strogatzs discovering small-world phenomenon in real world networks that led to introducing the WS model (1), Barabasi and Alberts investigating the preferential nature of network evolution which leads to scale-free real world networks (2) and detection of community structures in real world networks (3). Based on the literature, the studies on complex networks are mostly focused in two areas; analysis and structure. Of course there has been a substantial amount of research on the visualization and organization of complex networks, but for the most part the popular criteria of structure and analysis remains the same.

Discovering metrics and measures to gain insight and investigate network's behavior and statistical properties is the main focus of *analysis*. So far, metrics such as degree distribution, clustering coefficient, and average path length have been extensively used to create network models and categorize them. More research on non-trivial and combinatorial metrics can identify new properties and thus help the scientific

community better understand, explain, and predict the behavior of networks.

The area of modeling real world networks is mostly known as *structure*. Properties like small world phenomena and scale free networks inspired models that captured the nature of many real world networks. These models will be discussed in details in the next chapter. These models help researchers understand the structure and evolution of complex networks.

1.2 Thesis Outline

In this thesis, we look into several geodesic-based properties in complex real world networks and try to resolve common problems with respect to properties in complex network analysis, and help further enriching the research work in this new and interdisciplinary field.

This thesis consist of four chapters. In the next chapter, we look into the preliminaries of network science and provide an overview of historical background and introduction to basic concepts and terminologies. The chapter includes basics of graph theory, an overview of basic network properties, and a brief introduction to most well known network models currently being used by research community.

Chapter 3 is dedicated to investigating an interesting topological property known as hyperbolicity in real world complex networks. This geodesic-based property provides an interesting insight to the topology of many biological and social networks and we discuss the implications of hyperbolicity on the behavior of many real-world networks. In fact many interesting behavior we witness in networks can be discussed through the scope of network hyperbolicity:

- In biological networks, network motifs are often nested.
- In biological regulatory networks, paths mediating up- or down-regulation of a target node starting from the same regulator node often have many small cross-talk paths.
- An eavesdropper with limited sensor ranges can often intercept communications between nodes very far apart from it.

▹ In traffic networks, congestion can happen in a node that is not a hub.

Although each of these phenomena can be studied on its own, it is desirable to have a network measure reflecting salient properties of complex large-scale networks that can explain all these phenomena at one shot.

In chapter 4 we focus on another geodesic-based property known as strong metric dimension in networks and look deeper into algorithmic aspect of calculating strong metric dimension. In particular, we show the reduction of this problem to another well known problem within an additive logarithmic factor, thereby settling the computational complexity questions for this measure completely.

Chapter 5's focus is on a geodesic-based property known as (k, ℓ)-anonymity, which is an extension of strong metric dimension. This property is a privacy measure in networks under active attacks and provides insight into the robustness of a network against an attempted active attack. We describe our Investigation on this property in real world networks, which leads to important insights about active attacks and privacy in complex networks.

CHAPTER 2

PRELIMINARIES

In this chapter we discuss the preliminaries, terminologies, and mathematical models behind complex network analysis. This will include a brief introduction to essential graph theory concepts and an overview of important statistical properties and models for complex networks.

2.1 Mathematical Foundation

2.1.1 Graph Theory

The initial idea of graphs was presented by Leonard Euler the famous Swiss mathematician. He developed the idea of graphs in an attempt to solve the problem of crossing bridges in the town of Knigsberg. The problem was whether it was possible to walk through town in such way that each bridge was passed once and only once. He represented each land mass as a node (point) and each bridge as a line connecting two points. The following figure shows the actual map of Knigsberg and Eulers representation of it as a graph [1]. Euler argued that for every land mass there needs to be a way to get in and a way to get out. Thus each land mass requires an even number of bridges as oppose to existing odd numbers. Therefore, not possible to take a walk without passing a bridge more than once. This problem and Eulers approach to it shaped the foundation of graph theory as one of the essentials of discrete mathematics. Below we will briefly look into some key terminologies vital to graph theory.

Graph: A set of objects that are connected to each other through meaningful links. A graph $G = (V, E)$ contains a set of vertices V and set of edges E. It is useful to note that terms node and vertex and also edge and link are often used interchangeably. Each edge connects two nodes creating an *adjacent* pair or neighbors. An edge is usually depicted as pair of nodes it connects, $(u, v) \in E$ where u belongs to V and v

[1] Source: History of Mathematics Archive http://www-history.mcs.st-andrews.ac.uk/Extras/Konigsberg.html

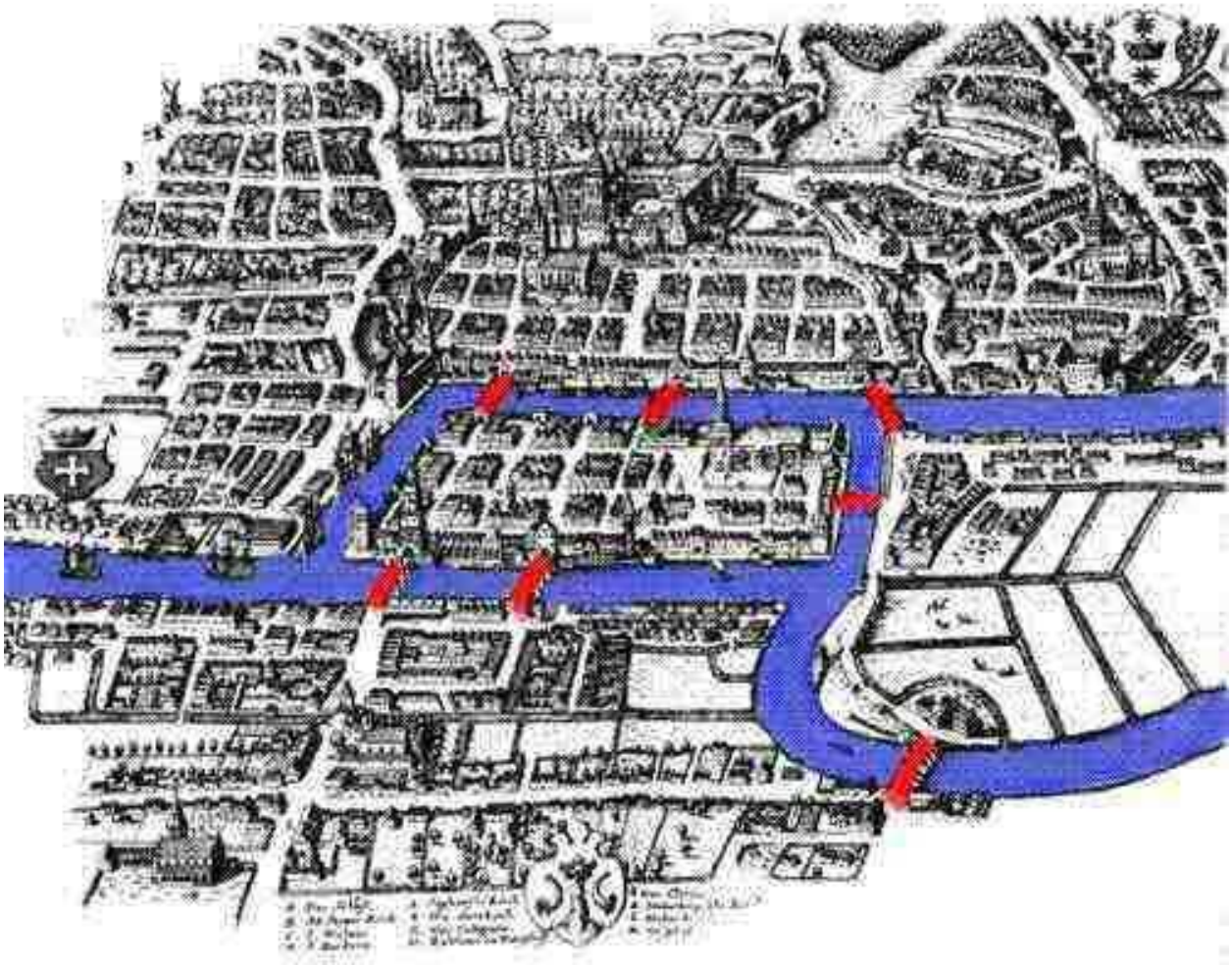


Figure 1: Image of city of Konigsberg and its seven bridges

belongs to V. A *degree* of a node is number of neighbors it has and is shown by $d(n)$.

Undirected and Directed Graphs: In an undirected graph orientation does not matter and there is no order to pairs. In other words, $(u, v) \in E \iff (v, u) \in E$. In an undirected graph each edge means both incoming and outgoing link and thus there is no distinction between $d_{outgoing}(v)$ and $d_{incoming}(v)$. In contrast, in a directed graph order matters as $(u, v) \in E \nLeftrightarrow (v, u) \in E$, and each edge has a source and a target which shows the flow and direction between two pairs. In a directed graph, nodes have two degrees, $d_{incoming}$ and $d_{outgoing}$. For a node v, the incoming degree ($d_{incoming}$) is number of edges where v is the target and outgoing

degree ($d_{outgoing}$) is number of edges where v is the source.

Simple and Multigraph: If a graph allows multiple edges between two nodes or self-loops for nodes then the graph is called a multigraph. A simple graph has one edge between each pair of nodes and self-loops are not allowed.

Weighted and Unweighted Graphs: If there are values assigned to every edge then graph is a weighted. These numbers can reflect different properties in a graph such as capacity, cost, or length. An unweighted graph is a graph where values assigned to each edge is 1.

Complete Graph: a graph that has an edge between every distinct pair of nodes is called a complete graph and shown by K_n where n is number of nodes.

Regular Graph: a graph where every node has the same number of neighbors and therefore degree. In a *regular directed graph*, number of incoming and outgoing edges for every nodes should be equal.

Weakly and Strongly Connected Graph: A connected graph is an undirected graph where there exists a path between every pair of nodes. For directed graphs, if for every $(u, v) \in E$ we have $(v, u) \in E$ then we call the graph strongly connected. A directed graph is weekly connected if we remove all direction from edges and end up with a connected undirected graph.

Bipartite Graph: A graph is bipartite if we can b divide the nodes into two disjoint sets U and V such that every node in U is connected to a node in V. A bipartite graph does not contain any odd-length cycles.

Subgraph: A subgraph of graph $G = (V, E)$ is a a graph $G' = (V', E')$ such that $V' \subset V$ and $E' \subset E$.

Path: A finite or infinite sequence of edges that connects a set of distinct nodes.

If the start node in path is the same as the end node, we call it a **cycle**.

Connected Component: A subgraph of an undirected graph such that for every distinct pair of nodes there exist a path in subgraph. In complex network analysis and network science, largest connected component is usually a desirable focus of the study and investigation.

Tree: A tree is a graph without a cycle where every two node are connected by exactly one path.

Cut: A partition of nodes into to disjoint sets is known as a *cut*. The set of edges whose end points are in different subsets of the partition is called *cutset*. Weight of a cut is defined as the sum of weight of all edges crossing the cut.

Clique: A subset of nodes in an undirected graph such that every two nodes in the subset are connected by an edge.

2.2 Properties of Networks

In this section, we will introduce some of the most popular and important properties in a network. Understanding and analyzing these properties provide insight into network behavior and prediction. Moreover, these properties are key factors in defining network models. Investigating such properties is important in network classification and topology comparison.

Diameter: The diameter of a network is the length of the longest shortest path, or the longest geodesics. Note that for disconnected networks, networks with more than one connected component, this definition only applies to reachable nodes. In some cases, the diameter of a disconnected network is ∞.

Density: Density measures the completeness of a network. In other words, it measures the ratio of the number of existing edges in E with respect to number of all possible edges. Density of a simple undirected network is

$$D = \frac{2|E|}{|V|(|V|-1)}$$

Average Degree: Number of edges connected to a node is denoted by its degree k. The average degree of a network is defined as

$$k = \frac{\sum_{i=1}^{N} k_i}{N} = \frac{2E}{N}$$

where k_i is the degree of node i and $N = |V| - 1$.

The next two properties were proposed based on works of Watts and Strogatz (1) as intuitive and important metrics. These two metrics known as Average path length $\mathcal{L}$ and Clustering Coefficient C, alongside degree distribution are often the pillars of network modeling and analysis and provide gainful insight and information about the structural behavior of a network.

Average Path Length: The average path length captures the idea of how far apart two nodes are, on average. The metric was inspired by the work of Milgram (4), which we will discuss further in network modeling. For a more formal definition, we need to get familiar with the notion of geodesic paths. *Geodesic* is the shortest path between two nodes in a network. The average path length is the average of all *geodesics*.If we show the geodesic between nodes i and j with $\mathrm{d_{i,j}}$ and define $\mathbb{D} = \{\mathrm{d_{i,j}} | i, j \in V, i \neq j\}$ as the set of all geodesics paths then we can define the average path length $\mathcal{L}$ of a network as follows:

$$\mathcal{L} = \frac{1}{\mathbb{D}} \sum_{i,j} \mathrm{d_{i,j}}$$

In real world networks, the possibility of disconnection and existence of multiple connected components is high. Therefore, network science community mostly focuses on the largest connected component which guarantees $\mathrm{d_{i,j}} < \infty$ for every $(u_i, u_j) \in V$.

Clustering Coefficient: This property is very common in social networks, and it indicates the tendency of a network to form cliques, a subset where everyone knows each other. Formally, the clustering coefficient of a node is the ratio of existing links connected to its neighbors over the maximum number of links that could possibly exist. This variation of clustering coefficient is a local property of a node i and is calculated as follows:

$$\mathrm{c_i} = \frac{2 * |e_{i,j}|}{k_i(k_i - 1)}$$

The other variation of clustering coefficient is a global metric that counts for all possible cliques and is shown as:

$$C_{glob} = \frac{3*numberoftriangles}{numberofalltriplets}$$

2.3 Network Models

Across different fields, networks are emerging as complex relational data and depicted using graphs. Network models establish a system to integrate and use mathematical and analytical tools and methods to capture network properties. They also provide great insight for predicting network behavior. Below, we take a look at three well-known models that gain a lot of attention and build the proper domain for extensive research in network science.

2.3.1 Erdös Rènyi Model

One of the first models to study networks was random graphs proposed by Erdös and Rènyi (5). This model describes a graph by a probability distribution that generates them. Due to their random nature, ER modeled graphs gained a lot of popularity and have been extensively used in complex network analysis to capture a typical graph. The model to generate a random graph with n nodes and m edges starts by having all n nodes disconnected and separate, then randomly selecting two nodes and connecting them until the desired number of edges is reached (6). It is obvious that there are $\binom{\frac{n(n-1)}{2}}{m}$ with this combination of nodes and edges and the outcome graph is just one realization.

Another model for ER random graphs is a procedure where every pair of nodes are connected with the probability $0 < p < 1$. This approach makes the total number of edges a random variable, which makes the probability of a graph like $G_0(n, p)$ as follows:

$$P(G_0(n, p)) = p^m(1p)^{\frac{n(n-1)}{2}m}$$

An interesting characteristic of ER graphs is graph evolution. By increasing p, obtained graphs evolve from low density, tree-like graphs to high link density, fully connected graphs.

The degree distribution in a random ER model follows a binomial distribution so we can simply show that degree distribution with parameters n and p as:

$$P(k) = C_k^{n-1} p^k (1-p)^{n-1-k}$$

In the above formula the probability of k edges existing is represented by p^k, while the probability of absence of other additional edges is $(1-p)^{n-1-k}$. If n is very large, as in many real networks,$< k > \simeq p.n$ and we can simplify the degree distribution to a Poisson distribution:

$$P(k) \simeq \frac{k^{<k>}}{k!} e^{-<k>}$$

For slightly large value of p, the diameter of random graphs tends to be small. Multiple studies on diameter of random graphs show that for most graphs with same n and p, diameter remains the same and is usually formulated by $\mathbb{D} = \frac{\ln n}{\ln pn}$. This is due to a general spread that takes place during the evolution step of the random graph. This spread also causes the average path length to follow a similar pattern. In random graphs, the average path length is $\mathcal{L}_{rand} = \frac{\ln n}{\ln pn}$.

Considering the procedure random graph evolution, we can state that the probability that two neighboring nodes of node i are connected is equal to the probability of connection between any two random nodes. In other words, the probability for any edge is p, which means the probability that two neighbors of a node are connected by an edge is p as well, so we can show the clustering coefficient with $\mathrm{C_{rand}} = p = \frac{<k>}{n}$.

2.3.2 Watts Strogatz Model

Stanley Milgram, an American psychologist introduced the notion of small world in his work six degrees of separation (4). He conducted series of experiments to find degree of separation in a network. In these experiments, he gave number of letters to participants living in Boston and Omaha, with delivery instructions

and target receiver. He asked the participant to mail the letters to another person they considered closest to the target receiver. Milgram experiments showed that on average the chain of people the letters go through to reach the target is about six. This groundbreaking research depicted a society far closer than what it was previously expected with short path lengths among two random strangers. Recall, what we are discussing here is a realization of average path length in a real world network. The phenomena captured in this study was later introduced by Watts and Strogatz as *small world* property (1).

The small world property shows that most real world networks, in particular social networks, are highly clustered despite their large sizes. In other words, most nodes are reachable from every other node through few numbers of steps. Watts and Strogatz introduced the formal definition of *small world* concept in 1999 by proposing a model of graphs that share the small average shortest path length with random graphs, but have relatively high clustering coefficient. This model which is a cross over between random graphs and regular lattice, is obtained through following steps:

Start with n nodes where each node has a degree of k or k edges connected to it. In order to have a sparse graph that is safe from the danger of becoming disconnected we need to insure that $n \gg k \gg \ln n \gg 1$. Build a normal ring lattice with n nodes, where each node is connected to k neighbors, $\frac{k}{2}$ on each side. Reattach each edge randomly with probability p, avoid self-loops and duplicate edges. By following the steps above, we introduce non-lattice edges, that are long-range and reduce the average path length by connecting distant nodes, resulting in small-world property, while the lattice structure keeps the locally clustered property.

It was observed (1) that for small values of p the rewired network display small path lengths and high clustering.

In WS model if $p = 0$, the degree distribution is a Direc delta function centered at k. By increasing p we introduce disturbance in the network, which change the degree distribution but keeps the average degree equal to K (7). Although the existence of small world property has been discovered in many real world networks such a food web, the World Wide Web, power grid networks, biological and social networks, the

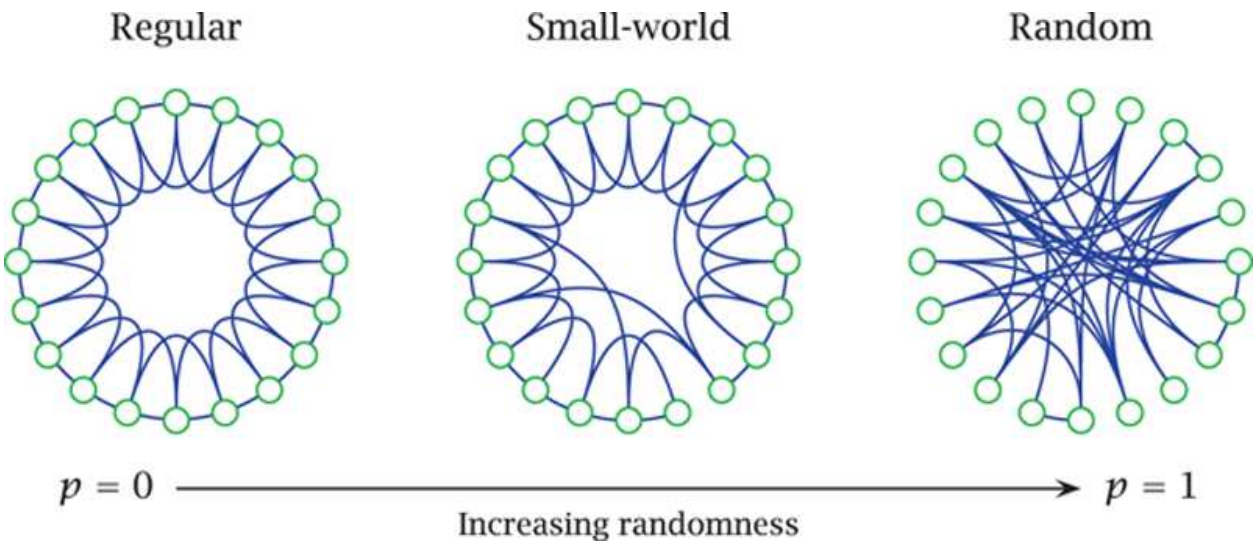


Figure 2: The process of random rewiring for interpolating between ring lattices and random graphs. The initial state is a ring of n nodes, connected to k nearest neighbors.

other properties of Watts and Strogatz model like degree distribution display a considerable difference from real world networks (7; 8). The limitation in evolution process of model and not accurate degree distribution, motivated network scientists to work on a model that can better capture real world networks.

2.3.3 Scale-free Networks

The previously discussed models were successful attempts in formulating some properties of real world networks but they failed to produce properties that capture two common aspects of real networks; network growth and preferential attachment.

From very first cases of investigating real world networks, a power-law degree distribution was observed and reported(9; 10; 11; 12). In 1999, Barabasi and Albert studied a portion of World Wide Web and discovered some nodes, known as hubs, had far more connections compare to others and the network itself has a power law degree distribution and therefore is free of scale. After discovering the same patterns in other networks, they call this type of networks scale-free networks(2). In recent years, dramatic increase in computational power and data collection have led to several large network datasets and simplified data analyzing, revealed

that in many networks degree distribution follows a power law for large k, i.e. $P(k) \sim c.k^{-\gamma}$ where c is a constant and γ is a positive value between 2 and 3. In fact even the networks with exponential-tail $P(k)$, degree distribution does not fall within Poisson distribution category.

In both ER and WS models we assume that we have a fixed number of nodes n, and these nodes are randomly connected or reconnected while n remains the same. However, most real world networks are open to adding new nodes trough their life cycle. e.g. World Wide Web grows exponentially by adding new web pages, or the research literature network that grows constantly by publication of new papers. Moreover, in both ER and WS model it is assumed that the probability that connects (or reconnects in WS model) every pair of nodes is independent from node degree, which means we randomly distribute new edges. On the contrary, in real world network we observe preferential attachment; which means it is more likely for nodes with higher degrees to receive new links. Consider our previous examples, the World Wide Web and research literature network, a new web page will more likely contains links to popular and well known pages, and a newly published paper is more likely to cite a well known paper.

In their 1999 paper Emergence of scaling in random networks , Barabasi and Albert took a new approach that was different from previous modeling attempts. In previous approaches, all the efforts were concentrated on building a graph that reflects correct topological features of the network. Barabasi and Albert proposed a new method that captured network dynamics by following the construction/evolution process of a network. Indeed, in dynamic networks it is assumed that by capturing the process that assembled the networks we see today, one can obtain their topology correctly as well.

The Barabasi-Albert model includes network growth and preferential attachment ingredients, present a class of networks known as Scale-free networks that are constructed as follows (7) start with m_0 isolated nodes, at each time step $t = 1, 2, 3, ..$ a new node with m connections where $m \leq m_0$ is added to the network. . The probability that a link will connect the new node to an existing node like i is linearly proportional to the actual degree of i:

$$\prod_{l \to i} = \frac{k_i}{\sum_j k_j}$$

After t_n time steps, the primary network evolves to a larger network with $n = t_n + m_0$ nodes and mt_n edges. Eventually, this network evolves into a network with a degree distribution $P(k) \sim k^{-\gamma}$.

Barabasi-Albert model was developed considering network growth and preferential attachment, to verify this model they developed two variants: variant A contained growth attribute but it was assumed that every new edge is assigned randomly (with equal probability for each node). In this case the probability that each node has for getting the edge is constant and leads to a degree distribution of form $P(k) = exp(-\beta k)$, which indicates scale-free property has been eliminated. variant B had preferential attachment and fixed number of nodes. This model started with n nodes and no edges, and at each time step a random node was selected and connected to node i with preferential attachment. After n^2 time steps the network evolves to a complete network, where every pair of nodes is connected. The failure of these two variants shows that network growth and preferential attachment are essential in power-law network development.

Barabasi-Albert model displays some limitations in capturing wide range of real-world networks. There are some assumptions in baseline of the model that cause these limitations. First, the model assumes that the preferential attachment is linear, $\prod k \sim k$. However, some networks show non-linear preferential attachment, which introduces deviations from power law degree distribution by limiting the size of hubs for ($\alpha < 1$) or creating super hubs for ($\alpha > 1$) (13). Second, empirical results showed γ obtained for different networks is distributed between [2.14]. Finally, they assumed that a network evolves only by adding new edges, while in some real networks the evolution process contains adding (or removing) links between already existed nodes.

Nevertheless, Barabasi-Albert model is one of the best and most used models in network science and complex network analysis field.

CHAPTER 3

HYPERBOLICITY

3.1 Introduction

Complex systems, ranging from the World-Wide Web to metabolic networks, representation as a parameterized network and graph theoretical analysis of this network have led to many useful insights (14; 7). Complex networks have been the center of much studies in recent years. In addition to established network measures such as the average degree, clustering coefficient or diameter, the complicated and highly interconnected structure of these networks has led researchers to propose and evaluate number of novel network measures (15; 16; 17; 18). We considered and studied a combinatorial measure of negative curvature known as hyperbolicity in parameterized finite networks and the implications of negative curvature on the higher-order connectivity and topological properties of these networks.

There are many ways in which the (positive or negative) curvature of a continuous surface or other similar spaces can be defined depending on whether the measure is to reflect the local or global properties of the underlying space. The specific notion of negative curvature that we use is an adoption of the hyperbolicity measure for a infinite metric space with bounded local geometry as originally proposed by Gromov (19) using a so-called "4-point condition". We adopt this measure for parameterized finite discrete metric spaces induced by a network via all-pairs shortest paths and apply it to biological and social networks. Recently, there has been a surge of empirical works measuring and analyzing the hyperbolicity of networks defined in this manner, and many real-world networks were observed to be hyperbolic in this sense. For example, preferential attachment networks were shown to be scaled hyperbolic in (20; 21), networks of high power transceivers in a wireless sensor network were empirically observed to have a tendency to be hyperbolic in (22), communication networks at the IP layer and at other levels were empirically observed to be hy-

perbolic in (23; 24), extreme congestion at a very limited number of nodes in a very large traffic network was shown in (25) to be caused due to hyperbolicity of the network together with minimum length routing, and the authors in (26) showed how to efficiently map the topology of the Internet to a hyperbolic space. Gromov's hyperbolicity measure adopted on a shortest-path metric of networks can also be visualized as a measure of the "closeness" of the original network topology to a tree topology (27). Another popular measure used in both the bioinformatics and theoretical computer science literature is the *treewidth* measure first introduced by Robertson and Seymour (28). Many NP-hard problems on general networks admit efficient polynomial-time solutions if restricted to classes of networks with bounded treewidth (29), just as several routing-related problems or the diameter estimation problem become easier if the network has small hyperbolicity (30; 31; 32; 33). However, as observed in (27), the two measures are quite different in nature: "the treewidth is more related to the least number of nodes whose removal changes the connectivity of the graph in a significant manner whereas the hyperbolicity measure is related to comparing the geodesics of the given network with that of a tree". Other related research works on hyperbolic networks include estimating the distortion necessary to map hyperbolic metrics to tree metrics (34) and studying the algorithmic aspects of several combinatorial problems on points in a hyperbolic space (35).

3.2 Hyperbolicity-related Definitions and Measures

Let $G = (V, E)$ be a *connected* undirected graph of $n \geq 4$ nodes. Consider the following notations:

- $u \overset{\mathcal{P}}{\rightsquigarrow} v$ denotes a path $\mathcal{P} \equiv (u = u_0, u_1, \ldots, u_{k-1}, u_k = v)$ from node u to node v and $\ell(\mathcal{P})$ denotes the *length* (number of edges) of such a path.
- $u_i \overset{\mathcal{P}}{\rightsquigarrow} u_j$ denotes the sub-path $\left(u_i, u_{i+1}, \ldots, u_j\right)$ of $\mathcal{P}$ from u_i to u_j.
- $u \overset{\text{S}}{\rightsquigarrow} v$ denotes a shortest path from node u to node v of length $d_{u,v} = \ell(u \overset{\text{S}}{\rightsquigarrow} v)$.

Now we can introduce the hyperbolicity measures via the 4-node condition as originally proposed by Gromov. Consider a quadruple of distinct nodes[1] u_1, u_2, u_3, u_4, and let $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ be a permutation of $\{1, 2, 3, 4\}$ denoting a rearrangement of the indices of nodes such that

$$S_{u_1,u_2,u_3,u_4} = d_{u_{\pi_1},u_{\pi_2}} + d_{u_{\pi_3},u_{\pi_4}} \leq M_{u_1,u_2,u_3,u_4} = d_{u_{\pi_1},u_{\pi_3}} + d_{u_{\pi_2},u_{\pi_4}} \leq L_{u_1,u_2,u_3,u_4} = d_{u_{\pi_1},u_{\pi_4}} + d_{u_{\pi_2},u_{\pi_3}}$$

and let $\delta^+_{u_1,u_2,u_3,u_4} = \frac{L_{u_1,u_2,u_3,u_4} - M_{u_1,u_2,u_3,u_4}}{2}$. Considering all combinations of four nodes in a graph one can define a worst-case hyperbolicity(19) as

$$\delta^+_{\text{worst}}(G) = \max_{u_1,u_2,u_3,u_4} \left\{ \delta^+_{u_1,u_2,u_3,u_4} \right\}$$

and an average hyperbolicity as

$$\delta^+_{\text{ave}}(G) = \frac{1}{\binom{n}{4}} \sum_{u_1,u_2,u_3,u_4} \delta^+_{u_1,u_2,u_3,u_4}$$

Note that $\delta^+_{\text{ave}}(G)$ is the expected value of $\delta^+_{u_1,u_2,u_3,u_4}$ when the four nodes u_1, u_2, u_3, u_4 are picked independently and uniformly at random from the set of all nodes. Both $\delta^+_{\text{worst}}(G)$ and $\delta^+_{\text{ave}}(G)$ can be trivially computed in $O(n^4)$ time for any graph G. A graph G is called δ-hyperbolic if $\delta^+_{\text{worst}}(G) \leq \delta$. If δ is a small constant independent of the parameters of the graph, a δ-hyperbolic graph is simply called a hyperbolic graph. It is easy to see that if G is a tree then $\delta^+_{\text{worst}}(G) = \delta^+_{\text{ave}}(G) = 0$. Thus all trees are hyperbolic graphs. The hyperbolicity measure δ^+_{worst} that was introduced for a metric space was originally used by Gromov in the context of group theory (19) by observing that many results concerning the fundamental group of a Riemann surface hold true in a more general context. δ^+_{worst} is trivially infinite in the standard (unbounded) Euclidean space. Intuitively, a metric space has a finite value of δ^+_{worst} if it behaves metrically in the large

[1] If two or more nodes among u_1, u_2, u_3, u_4 are identical, then $\delta^+_{u_1,u_2,u_3,u_4} = 0$ due to the metric's triangle inequality; thus it suffices to assume that the four nodes are distinct.

scale as a negatively curved Riemannian manifold, and thus the value of δ^+_{worst} can be related to the standard scalar curvature of a Hyperbolic manifold. For example, a simply connected complete Riemannian manifold whose sectional curvature is below $\alpha < 0$ has a value of δ^+_{worst} that is $\mathrm{O}\left(\left(\sqrt{-\alpha}\right)^{-1}\right)$ (see (36)). We first show that a variety of biological and social networks are hyperbolic. then, formulate and prove bounds on the existence of path-chords and on the distance among shortest or approximately shortest paths in hyperbolic networks. We determine the implications of these bounds on *regulatory* networks, *i.e.*, directed networks whose edges correspond to regulation or influence. This category includes all the biological networks that we studied. We also discuss the implications of these results on the region of influence of nodes in social networks. Some of the proofs of these theoretical results are adaptation of corresponding arguments in the continuous hyperbolic space.

3.3 Hyperbolicity of Real Networks

We analyzed twenty well-known biological and social networks. The 11 biological networks shown in Table XVI include 3 transcriptional regulatory, 5 signalling, 1 metabolic, 1 immune response and 1 oriented protein-protein interaction networks. Similarly, the 9 social networks shown in Table XVIII range from interactions in dolphin communities to the social network of jazz musicians. The hyperbolicity of the biological and directed social networks was computed by ignoring the direction of edges. The hyperbolicity values were calculated by writing codes in C using standard algorithmic procedures. As shown on Table XVI and Table XVIII, the hyperbolicity values of almost all networks are small. If $\mathcal{D} = \max_{u,v}\{d_{u,v}\}$ is the diameter of the graph, then it is easy to see that $\delta^+_{\text{worst}}(G) \leq \mathcal{D}/2$, and thus small diameter indeed implies a small value of worst-case hyperbolicity. As can be seen on Table XVI and Table XVIII, $\delta^+_{\text{worst}}(G)$ varies with respect to its worst-case bound of $\mathcal{D}/2$ from 25% of $\mathcal{D}/2$ to no more than 89% of $\mathcal{D}/2$, and there does not seem to be a systematic dependence of $\delta^+_{\text{worst}}(G)$ on the number of nodes (which ranges from 18 to 786), edges (from 42 to 2742), or on the value of the diameter $\mathcal{D}$. For all the networks $\delta^+_{\text{ave}}(G)$ is one or two orders of magnitude smaller than $\delta^+_{\text{worst}}(G)$. Intuitively, this suggests that the value of $\delta^+_{\text{worst}}(G)$ may be a rare devi-

ation from typical values of $\delta^{+}_{u_1,u_2,u_3,u_4}$ that one would obtain for most combinations of nodes $\{u_1, u_2, u_3, u_4\}$. We additionally performed the following rigorous tests for hyperbolicity of our networks.

3.3.1 Checking hyperbolicity via the scaled hyperbolicity approach

An approach for testing hyperbolicity for finite graphs was introduced and used via "scaled" Gromov hyperbolicity in (21; 23) for hyperbolicity defined via thin triangles and in (53) for for hyperbolicity defined via the four-point condition as used in this work. The basic idea is to "scale" the values of $\delta^{+}_{u_1,u_2,u_3,u_4}$ by a suitable scaling factor, say μ_{u_1,u_2,u_3,u_4}, such that there exists a constant $0 < \varepsilon < 1$ with the following property:

- the maximum achievable value of $\frac{\delta^{+}_{u_1,u_2,u_3,u_4}}{\mu_{u_1,u_2,u_3,u_4}}$ is ε in the standard hyperbolic space or in the Euclidean space, and
- $\frac{\delta^{+}_{u_1,u_2,u_3,u_4}}{\mu_{u_1,u_2,u_3,u_4}}$ goes beyond ε in positively curved spaces.

We use the notation $\mathcal{D}_{u_1,u_2,u_3,u_4} = \max\limits_{i,j\in\{1,2,3,4\}} \left\{d_{u_i,u_j}\right\}$ to indicate the diameter of the subset of four nodes u_1, u_2, u_3 and u_4. By using theoretical or empirical calculations, the authors in (53) provide the bounds shown in Table III.

We adapt the criterion proposed by Jonckheere, Lohsoonthorn and Ariaei (53) to designate a given finite graph as hyperbolic by requiring a *significant* percentage of all possible subset of four nodes to satisfy the ε bound. More formally, suppose that G has t connected components containing $n_1, n_2, \ldots, n_t$ nodes, respectively ($\sum_{j=1}^{t} n_j = n$). Let $0 < \eta < 1$ be a sufficiently high value indicating the confidence level in declaring the graph G to be hyperbolic. Then, we call our given graph G to be (scaled) hyperbolic if and only if

TABLE I: Hyperbolicity and diameter values for biological networks.

Network *id*	reference	Average degree	$\delta^+_{\text{ave}}(G)$	$\delta^+_{\text{worst}}(G)$	$\mathcal{D}$	$\frac{\delta^+_{\text{worst}}(G)}{\mathcal{D}/2}$
1. *E. coli* transcriptional	(37)	1.45	0.132	2	10	0.400
2. Mammalian Signaling	(38)	2.04	0.013	3	11	0.545
3. *E. Coli* transcriptional	♯	1.30	0.043	2	13	0.308
4. T LGL signaling	(39)	2.32	0.297	2	7	0.571
5. *S. cerevisiae* transcriptional	(40)	1.56	0.004	3	15	0.400
6. *C. elegans* Metabolic	(9)	4.50	0.010	1.5	7	0.429
7. *Drosophila* segment polarity	(41)	1.69	0.676	4	9	0.889
8. ABA signaling	(42)	1.60	0.302	2	7	0.571
9. Immune Response Network	(43)	2.33	0.286	1.5	4	0.750
10. T Cell Receptor Signalling	(44)	1.46	0.323	3	13	0.462
11. Oriented yeast PPI	(45)	3.11	0.001	2	6	0.667

♯ (37, updated version)

see www.weizmann.ac.il/mcb/UriAlon/Papers/networkMotifs/coli1_1Inter_st.txt

TABLE II: Hyperbolicity and diameter values for social networks.

Network *id*	reference	Average degree	$\delta^+_{\text{ave}}(G)$	$\delta^+_{\text{worst}}(G)$	$\mathcal{D}$	$\frac{\delta^+_{\text{worst}}(G)}{\mathcal{D}/2}$
1. Dolphins social network	(46)	5.16	0.262	2	8	0.750
2. American College Football	(47)	10.64	0.312	2	5	0.800
3. Zachary Karate Club	(48)	4.58	0.170	1	5	0.400
4. Books about US Politics	‡	8.41	0.247	2	7	0.571
5. Sawmill communication	(49)	3.44	0.162	1	8	0.250
6. Jazz musician	(50)	27.69	0.140	1.5	6	0.500
7. Visiting ties in San Juan	(51)	3.84	0.422	3	9	0.667
8. World Soccer data, 1998	†	3.37	0.270	2.5	12	0.286
9. Les Miserable	(52)	6.51	0.278	2	14	0.417

‡ V. Krebs, `www.orgnet.com`,

† Dagstuhl seminar: *Link Analysis and Visualization*, Dagstuhl 1-6, 2001;
`vlado.fmf.uni-lj.si/pub/networks/data/sport/football.htm`

TABLE III: Various scaled Gromov hyperbolicities.

Name	Notation	μ_{u_1,u_2,u_3,u_4}	ε	Method for determining ε
diameter-scaled hyperbolicity	$\delta^{\mathcal{D}}$	$\mathcal{D}_{u_1,u_2,u_3,u_4}$	0.2929	empirical
L-scaled hyperbolicity	δ^{L}	L_{u_1,u_2,u_3,u_4}	$\frac{\sqrt{2}-1}{2\sqrt{2}}$ ≈ 0.1464	mathematical
$(L+M+S)$-scaled hyperbolicity	$\delta^{\mathrm{L+M+S}}$	L_{u_1,u_2,u_3,u_4} $+ M_{u_1,u_2,u_3,u_4}$ $+ S_{u_1,u_2,u_3,u_4}$	0.0607	mathematical

$$\Delta^{\mathrm{Y}}(G) = \frac{\begin{array}{c}\text{number of subset of four nodes } \{u_i,u_j,u_k,u_\ell\} \\ \text{such that } \delta^{\mathrm{Y}}_{u_i,u_j,u_k,u_\ell} > \varepsilon\end{array}}{\begin{array}{c}\text{number of all possible combinations of four nodes} \\ \text{that contribute to hyperbolicity}\end{array}}$$

$$= \frac{\begin{array}{c}\text{number of subset of four nodes } \{u_i,u_j,u_k,u_\ell\} \\ \text{such that } \delta^{\mathrm{Y}}_{u_i,u_j,u_k,u_\ell} > \varepsilon\end{array}}{\sum_{1\le j\le t:\ n_j>3}\binom{n_j}{4}} < 1-\eta$$

The values of $\Delta^{\mathrm{Y}}(G)$ for our networks are shown in Table IV and Table V. It can be seen that, for all scaled hyperbolicity measures and for all networks, the value of $1-\eta$ is very close to zero.

We next tested the statistical significance of the $\Delta^{\mathrm{Y}}(G)$ values by computing the statistical significance values (commonly called p-values) of these $\Delta^{\mathrm{Y}}(G)$ values for each network G with respect to a null hypothesis model of the networks. We use a standard method used in the network science literature (e.g. see (17; 37)) for such purpose. For each network G, we generated 100 randomized versions of the network using a Markov-chain algorithm (54) by swapping the endpoints of randomly selected pairs of edges until 20% of the edges was changed. We computed the values of $\Delta^{\mathrm{Y}}(G_{\mathrm{rand}_1})$, $\Delta^{\mathrm{Y}}(G_{\mathrm{rand}_2})$, ..., $\Delta^{\mathrm{Y}}(G_{\mathrm{rand}_{100}})$. We

TABLE IV: $\Delta^{\mathrm{Y}}(G)$ values for biological networks for $Y \in \{\mathcal{D},\, L,\, L+M+S\}$.

Network *id*	$\Delta^{\mathcal{D}}(G)$	$\Delta^{\mathrm{L}}(G)$	$\Delta^{\mathrm{L+M+S}}(G)$
1. *E. coli* transcriptional	0.0014	0.0018	0.0015
2. Mammalian Signaling	0.0021	0.0018	0.0022
3. *E. Coli* transcriptional	0.0006	0.0006	0.0007
4. T LGL signaling	0.0228	0.0221	0.0318
5. *S. cerevisiae* transcriptional	0.0031	0.0032	0.0033
6. *C. elegans* Metabolic	0.0020	0.0018	0.0019
7. *Drosophila* segment polarity	0.0374	**0.0558**	0.0750
8. ABA signaling	0.0343	0.0285	0.0425
9. Immune Response Network	**0.0461**	0.0552	**0.0781**
10. T Cell Receptor Signalling	0.0034	0.0045	0.0056
11. Oriented yeast PPI	0.0013	0.0009	0.0012
maximum	0.0461	0.0558	0.0781

TABLE V: $\Delta^{\mathrm{Y}}(G)$ values for social networks for $Y \in \{\mathcal{D},\, L,\, L+M+S\}$.

Network *id*	$\Delta^{\mathcal{D}}(G)$	$\Delta^{\mathrm{L}}(G)$	$\Delta^{\mathrm{L+M+S}}(G)$
1. Dolphins social network	0.0115	0.0120	0.0168
2. American College Football	**0.0435**	**0.0395**	**0.0577**
3. Zachary Karate Club	0.0195	0.0249	0.0284
4. Books about US Politics	0.0106	0.0074	0.0116
5. Sawmill communication	0.0069	0.0068	0.0085
6. Jazz musician	0.0097	0.0117	0.0124
7. Visiting ties in San Juan	0.0221	0.0242	0.0275
8. World Soccer data, 1998	0.0145	0.0155	0.0212
9. Les Miserable	0.0032	0.0034	0.0049
maximum	0.0435	0.0395	0.0577rigorous tests

then used an (unpaired) one-sample student's t-test to determine the probability that $\Delta^{\mathrm{Y}}(G)$ belongs to the same distribution as $\Delta^{\mathrm{Y}}\left(G_{\mathrm{rand}_1}\right)$, $\Delta^{\mathrm{Y}}\left(G_{\mathrm{rand}_2}\right)$, $\ldots$, $\Delta^{\mathrm{Y}}\left(G_{\mathrm{rand}_{100}}\right)$. The p-values, tabulated in Table VI and Table VII, clearly show that all social networks and all except two biological networks can be classified

TABLE VI: p-values for the $\Delta^{Y}(G)$ values for biological networks for $Y \in \{\mathcal{D},\ L,\ L+M+S\}$. In general, a p-value less than 0.05 (shown in **boldface**) is considered to be statistically significant, and a p-value above 0.05 is considered to be *not* statistically significant.

		Network *id*										
		1. *E. coli*	**2**. Mammalian Signaling	**3**. *E. Coli* transcriptional	**4**. T LGL signaling	**5**. *S. cerevisiae* transcriptional	**6**. *C. elegans* Metabolic	**7**. *Drosophila* segment polarity	**8**. ABA signaling	**9**. Immune Response Network	**10**. T Cell Receptor Signalling	**11**. Oriented yeast PPI
p values	$\Delta^{\mathcal{D}}$	**0.0018**	**< 0.0001**	**< 0.0001**	**< 0.0001**	0.3321	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**
	Δ^{L}	**< 0.0001**	**< 0.0001**	**< 0.0001**	0.011	0.3434	**< 0.0001**	**< 0.0001**	0.9145	**< 0.0001**	**< 0.0001**	**< 0.0001**
	Δ^{L+M+S}	0.5226	**< 0.0001**	**< 0.0001**	**< 0.0001**	0.3424	**< 0.0001**	**< 0.0001**	0.3342	**< 0.0001**	**< 0.0001**	**< 0.0001**

as hyperbolic in a statistically significant manner, implying that the topologies of these networks are close to a "tree topology". Indeed, for biological networks, the assumption of chain-like or tree-like topology is frequently made in the traditional molecular biology literature (55). Independent current observations also provide evidence of tree-like topologies for various biological networks, e.g. the average in/out degree of transcriptional regulatory networks (37; 56) and of a mammalian signal transduction network (38) is close to 1, so cycles are very rare.

TABLE VII: p-values for the $\Delta^{Y}(G)$ values for social networks for $Y \in \{\mathcal{D},\ L,\ L+M+S\}$. In general, a p-value less than 0.05 (shown in **boldface**) is considered to be statistically significant, and a p-value above 0.05 is considered to be *not* statistically significant.

		Network *id*								
		1. Dolphins social network	**2**. American College Football	**3**. Zachary Karate Club	**4**. Books about US Politics	**5**. Sawmill communication	**6**. Jazz musician	**7**. Visiting ties in San Juan	**8**. World Soccer data, 1998	**9**. Les Miserable
p values	$\Delta^{\mathcal{D}}$	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**
	Δ^{L}	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	0.0779	**< 0.0001**	**< 0.0001**
	Δ^{L+M+S}	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**

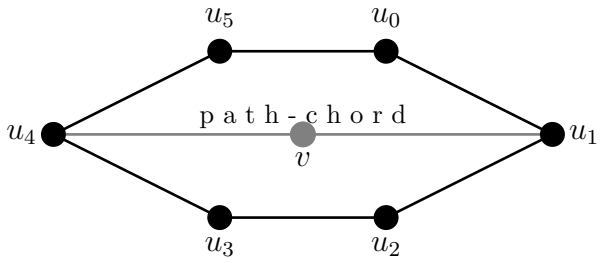


Figure 3: Path-chord of a cycle $C = (u_0.u_1, u_2, u_3, u_4, u_5, u_0)$.

3.3.2 Hyperbolicity and crosstalk in regulatory networks

Let $C = (u_0, u_1, \ldots, u_{k-1}, u_0)$ be a cycle of $k \geq 4$ nodes. A *path-chord* of C is defined to be a path $u_i \overset{\mathcal{P}}{\rightsquigarrow} u_j$ between two distinct nodes $u_i, u_j \in C$ such that the length of $\mathcal{P}$ is less than $(i - j) \pmod k$ (see Fig. Figure 3). A path-chord of length 1 is simply called a chord. We find that large cycles without a path-chord imply large lower bounds on hyperbolicity (see Theorem 3 in Section A.1 of the Appendices). In particular, G does not have a cycle of more than $4\,\delta^+_{\text{worst}}(G)$ nodes that does not have a path-chord. Thus, for example, if $\delta^+_{\text{worst}}(G) < 1$ then G has no chordless cycle, *i.e.*, G is a chordal graph. The intuition behind the proof of Theorem 3 is that if G contains a long cycle without a path-chord then we can select four *almost* equidistant nodes on the cycle and these nodes give a large hyperbolicity value. This general result has the following implications for regulatory networks:

- If a node regulates itself through a long feedback loop (e.g. of length at least 6 if $\delta^+_{\text{worst}}(G) = 3/2$) then this loop *must* have a path-chord. Thus it follows that there exists a *shorter* feedback cycle through the same node.

- A chord or short path-chord can be interpreted as *crosstalk* between two paths between a pair of nodes. With this interpretation, the following conclusion follows. If one node in a regulatory network regulates another node through two sufficiently long paths, then there must be a crosstalk path between

these two paths. For example, assuming $\delta^+_{\text{worst}}(G) = 3/2$, there must be a crosstalk path if the sum of lengths of the two paths is at least 6. In general, the number of crosstalk paths between two paths increases *at least* linearly with the total length of the two paths. The general conclusion that can be drawn is that independent linear pathways that connect a signal to the same output node (e.g. transcription factor) are rare, and if multiple pathways exist then they are interconnected through cross-talks.

3.3.3 Shortest-path triangles and crosstalk paths in regulatory networks

(*a*) Result related to triplets of shortest paths Originally, the hyperbolicity measure was introduced for infinite continuous metric spaces with negative curvature via the concept of the "thin" and "slim" triangles (e.g. see (57)). For finite discrete metric spaces as induced by an undirected graph, one can analogously define a *shortest-path triangle* (or, simply a *triangle*) $\Delta_{\{u_0,u_1,u_2\}}$ as a set of three distinct nodes u_0, u_1, u_2 with a set of three shortest paths $\mathcal{P}_\Delta(u_0, u_1), \mathcal{P}_\Delta(u_0, u_2), \mathcal{P}_\Delta(u_1, u_2)$ between u_0 and u_1, u_0 and u_2, and u_1 and u_2, respectively. As illustrated on Fig. Figure 4, in hyperbolic networks we are guaranteed to find short paths[1] between the nodes that make up $\mathcal{P}_\Delta(u_0, u_1), \mathcal{P}_\Delta(u_0, u_2), \mathcal{P}_\Delta(u_1, u_2)$. This is formally stated in Theorem 5 in Section A.2 of the Appendices. Moreover, as Corollary 6 (in Section A.2 of the Appendices) states, we can have a small Hausdorff distance between these shortest paths. This result is a *proper* generalization of our previous result on path-cords. Indeed, in the special case when u_1 and u_2 are the same node the triangle becomes a shortest-path cycle involving the shortest paths between u_0 and u_1 and the short-cord result is obtained. A proof of Theorem 5 is obtained by appropriate modification of a known similar bound for infinite continuous metric spaces. The implications of this result for regulatory networks can be summarized as follows:

---

[1] By a short path here, we mean a path whose length is at most a constant times $\delta^+_{\Delta_{\{u_0,u_1,u_2\}}}$ (note that $\delta^+_{\Delta_{\{u_0,u_1,u_2\}}} \leq \delta^+_{\text{worst}}(G)$).

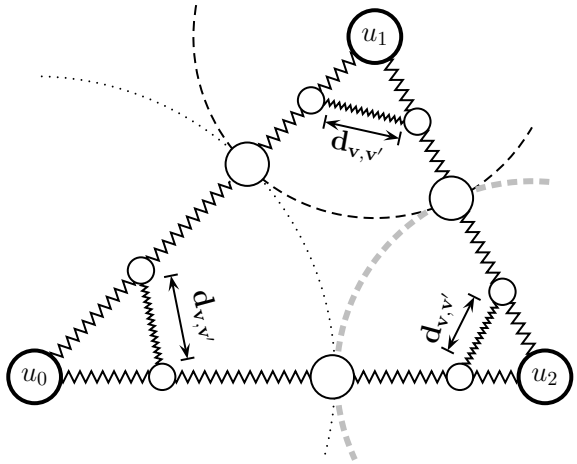


Figure 4: An informal and simplified pictorial illustration of the claims in Section 3.3.3(a).

If we consider a feedback loop (cycle) or feed-forward loop formed by the shortest paths among three nodes, we can expect short cross-talk paths between these shortest paths. Consequently, the feedback or feed-forward loop will be *nested* with "additional" feed-back or feed-forward loops in which one of the paths will be slightly longer.

The above finding is empirically supported by the observation that network motifs (e.g. feed-forward or feed-back loops composed of three nodes and three edges) are often nested (58).

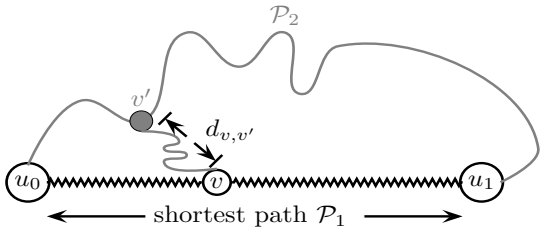


Figure 5: An informal and simplified pictorial illustration of the claims in Section 3.3.3(b).

(*b*) **Results related to the distance between two exact or approximate shortest paths between the same pair of nodes** It is reasonable to assume that, when up- or down-regulation of a target node is mediated by two or more short paths [1] starting from the *same* regulator node, additional very long paths between the same regulator and target node do *not* contribute significantly to the target node's regulation. We refer to the short paths as relevant, and to the long paths as irrelevant. Then, our finding can be summarized by saying that:

> *almost all* relevant paths between two nodes have crosstalk paths between each other.

Formal Justifications and Intuitions (see Theorem 7 and Corollary 8 in Section A.2.1 and Theorem 9 and Corollary 10 in Section A.2.2 of the Appendices)

We use the following two quantifications of "approximately" short paths:

- A path $u_0 \overset{\mathcal{P}}{\leftrightsquigarrow} u_k = (u_0, u_1, \ldots, u_k)$ is μ-approximate short provided $\ell(u_i \overset{\mathcal{P}}{\leftrightsquigarrow} u_j) \leq \mu\, d_{u_i,u_j}$ for all $0 \leq i < j \leq k$,
- A path $u_0 \overset{\mathcal{P}}{\leftrightsquigarrow} u_k$ is ε-additive-approximate short provided $\ell(\mathcal{P}) \leq d_{u_0,u_k} + \varepsilon$.

A mathematical justification for the claim then is provided by two separate theorems and their corollaries:

- Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be a shortest path and an arbitrary path, respectively, between two nodes u_0 and u_1. Then, Theorem 7 and Corollary 8 implies that, for every node v on $\mathcal{P}_1$, there exists a node v' on $\mathcal{P}_2$ such that $d_{v,v'}$ depends linearly on $\delta^+_{\text{worst}}(G)$, only logarithmically on the length of $\mathcal{P}_2$ and does *not* depend on the size or any other parameter of the network.

---

[1] Here by short paths we mean either a shortest path or an approximately shortest path whose length is not too much above the length of a shortest path, *i.e.*, a μ-approximate short path or a ε-additive-approximate short path, as defined in the subsequent "Formal Justifications and Intuitions" subsection, for small μ or small ε, respectively.

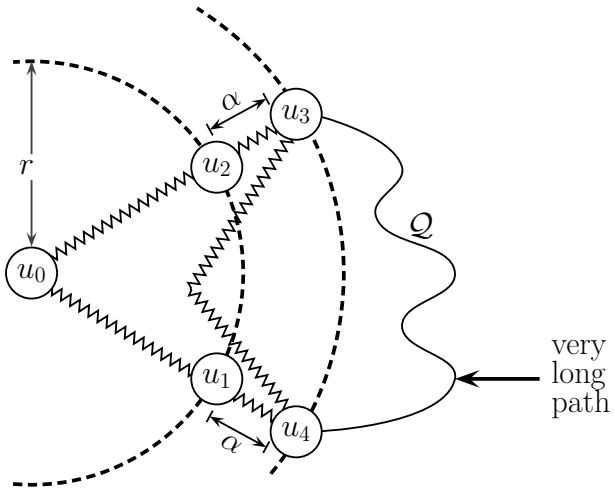


Figure 6: An informal and simplified pictorial illustration of claim ($\star$) in Section 3.3.4. As the nodes u_3 and u_4 move further away from the center node u_0, the shortest path between them bends more towards u_0 and any path between them that does not involve a node in the ball $\cup_{r' \leq r} B_{r'}(u_0)$ is long enough.

To obtain this type of bound, one needs to apply Theorem 5 on u_0, u_1 and the middle node of the path $\mathcal{P}_2$ and then use the same approach recursively on a part of the path $\mathcal{P}_2$ containing at most $\lceil \ell(\mathcal{P}_2)/2 \rceil$ edges. The depth of the level of recursion provides the logarithmic factor in the bound.

- If $\mathcal{P}_1$ and $\mathcal{P}_2$ are two short paths between u_0 and u_1 then Theorem 9 and Corollary 10 imply that the Hausdorff distance between $\mathcal{P}_1$ and $\mathcal{P}_2$ depends on $\delta^+_{\text{worst}}(G)$ only and does *not* depend on the size or any other parameter of the network.

  Intuitively, Theorem 9 and Corollary 10 can be thought of as generalizing and improving the bound in Theorem 7 for approximately short paths.

3.3.4 Identifying essential edges in the regulation between two nodes

For a given $\xi > 0$ and a node u, let $\mathcal{B}_\xi(u) = \{ v \,|\, d_{u,v} = \xi \}$ denote the "boundary of the ξ-neighborhood" of u, *i.e.*, the set of all nodes at a distance of precisely ξ from u. Our two findings in the present context are as stated in **(I)** and **(II)** below.

(I) Identifying relevant paths between a source and a target node Suppose that we pick a node v and consider the *strict* ξ-neighborhood of v

$$N_\xi^+(v) = \bigcup_{r \le \xi} \mathcal{B}_{r'}(v) \setminus \{u \,|\, \text{ degree of } u \text{ is one } \}$$

(*i.e.*, the set of all nodes, excluding nodes of degree 1, that are at a distance at most ξ from u) for a sufficiently large ξ. Consider two nodes u_1 and u_2 on the boundary of this neighborhood, *i.e.*, at a distance ξ from v. Then, the following holds:

> ($\star$) the relevant (short) regulatory paths between u_1 and u_2 do *not* leave the neighborhood, *i.e.*,
>
> *all* the edges in the relevant regulatory paths are in the neighborhood.

Thus, *only* the edges inside the neighborhood are relevant to the regulation among this pair of nodes. This result can be adapted to find the most relevant paths between the input node u_{source} and output node u_{target} of a signal transduction network. In many situations, for example when the signal transduction network is inferred from undirected protein-protein interaction data, a large number of paths can potentially be included in the signal transduction network as the protein-protein interaction network has a large connected component with a small average path length (58). There is usually *no* prior knowledge on which of the existing paths are relevant to the signal transduction network. A hyperbolicity-based method is to first find a central node u_{central} which is at equal distance between u_{source} and u_{target}, and is on the shortest, or close to shortest, path between u_{source} and u_{target}. Then one constructs the neighborhood around u_{central} such that

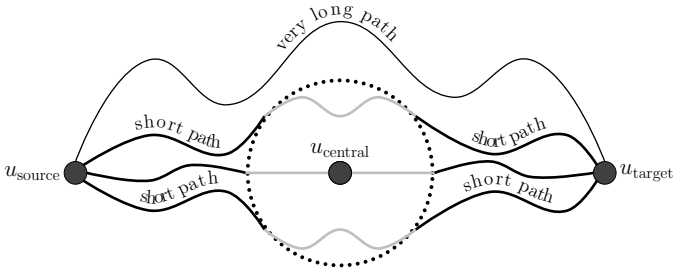


Figure 7: An informal and simplified pictorial illustration of claim (⋆⋆) in Section 3.3.4. Knocking out the nodes in a small neighborhood of u_{central} cuts off all relevant (short) regulation between u_{source} and u_{target}.

u_{source} and u_{target} are on the boundary of this neighborhood. Applying this result, the paths relevant to the signal transduction network are inside the neighborhood, and the paths that go out of the neighborhood are irrelevant. See Fig. Figure 6 for a pictorial illustration of this implication.

(II) Finding essential nodes Again, consider an input node u_{source} and output node u_{target} of a signal transduction network, and let u_{central} be a central node which is on the shortest path between them and at approximately equal distance between u_{source} and u_{target}. Our results show that[1]

> (⋆⋆) if one constructs a small ξ-neighbourhood around u_{central} with $\xi = \mathrm{O}\left(\delta^{+}_{\text{worst}}(G)\right)$, then *all* relevant (short or approximately short) paths between u_{source} and u_{target} must include a *node* in this ξ-neighborhood. Therefore, "knocking out" the nodes in this ξ-neighborhood cuts off all relevant regulatory paths between u_{source} and u_{target}.

---

[1] O and Ω are the standard notations used in analyzing asymptotic upper and lower bounds in the computer science literature: given two functions $f(n)$ and $g(n)$ of a variable n, $f(n) = \mathrm{O}(g(n))$ (respectively, $f(n) = \Omega(g(n))$ provided there exists two constants $n_0, c > 0$ such that $f(n) \leq c\, g(n)$ (respectively, $f(n) \geq c\, g(n)$) for $n \geq n_0$.

See Fig. Figure 7 for a pictorial illustration of this implication. Note that the size ξ of the neighborhood depends only on $\delta^+_{\text{worst}}(G)$ which, as our empirical results indicate, is usually a small constant for real networks.

Formal Justifications and Intuitions for ($\star$) **and** ($\star\star$) (see Theorem 12 and Corollary 13 in Section A.2.3 of the Appendices)

Suppose that we are given the following:

- three integers $\kappa \geq 4$, $\alpha > 0$,
  $r > \left(\frac{\kappa}{2} - 1\right)\left(6\,\delta^+_{\text{worst}}(G) + 2\right)$,
- five nodes u_0, u_1, u_2, u_3, u_4 such that
  - $u_1, u_2 \in B_r(u_0)$ with $d_{u_1,u_2} \geq \frac{\kappa}{2}\left(6\,\delta^+_{\text{worst}}(G) + 2\right)$,
  - $d_{u_1,u_4} = d_{u_2,u_3} = \alpha$.

Then, ($\star$) and ($\star\star$) are implied by following type of *asymptotic* bounds provided by Theorem 12 and Corollary 13:

> For a suitable positive value $\lambda = \mathrm{O}(\delta^+_{\text{worst}}(G))$, if $d_{u_1,u_4} = d_{u_2,u_3} = \alpha > \lambda$ then one of the following is true for any path Q between u_3 and u_4 that does not involve a node in $\cup_{r' \leq r}\mathcal{B}_{r'}(u_0)$:
>
> - Q does not exist (*i.e.*, $\ell(Q) \geq n$), or
> - Q is much longer than a shortest path between the two nodes, *i.e.*, if Q is a μ-approximate short path or a ε-additive-approximate short path then μ or ε is large.
>
> A pessimistic estimate shows that a value of λ that is about $6\,\delta^+_{\text{worst}}(G) + 2$ suffices. As we subsequently observe, for real networks the bound is much better, about $\lambda \approx \delta^+_{\text{worst}}(G)$.

TABLE VIII: Effect of the prescribed neighborhood in claim ($\star$) on all edges in relevant paths.

$\mathcal{SP}$: shortest path between u_{source} and u_{target}
$\mathcal{SP}^{+1}$: paths between u_{source} and u_{target} with one extra edge than $\mathcal{SP}$ (1-additive-approximate short path)
$\mathcal{SP}^{+2}$: paths between u_{source} and u_{target} with two extra edges than $\mathcal{SP}$ (2-additive-approximate short path)
$N_{\xi}^{+}(u_{\text{central}})$: strict $\xi = d_{u_{\text{source}}, u_{\text{target}}}$ neighborhood of u_{central}
n : size (number of nodes) of the network
$N_{\xi}^{+}(u_{\text{central}})/n$: fraction of strict $\xi = d_{u_{\text{source}}, u_{\text{target}}}$ neighborhood of u_{central} with respect to the size of the network

Network name	u_{source}	u_{target}	$d_{u_{\text{source}}, u_{\text{target}}}$	u_{central}	$\frac{N_{\xi}^{+}(u_{\text{central}})}{n}$	% of $\mathcal{SP}$ with every edge in the neighborhood of claim ($\star$)	% of $\mathcal{SP}^{+1}$ with every edge in the neighborhood of claim ($\star$)	% of $\mathcal{SP}^{+2}$ with every edge in the neighborhood of claim ($\star$)
Network 1: *E. coli* transcriptional	fliAZY	arcA	4	CaiF	0.20	100%	100%	18%
				crp	0.27	100%	100%	70%
	fecA	aspA	6	crp	0.43	100%	100%	100%
				sodA	0.28	100%	100%	62%
Network 4: T-LGL signaling	IL15	Apoptosis	4	GZMB	0.37	100%	66%	40%
	PDGF	Apoptosis	6	IL2, NKFB	0.72,0.59	100%	100%	100%
				Ceramide	0.60	80%	64%	36%
				MCL1	0.59	80%	88%	93%
	stimuli	Apoptosis	4	GZMB	0.37	100%	100%	100%

Empirical evaluation of ($\star$)

We empirically investigated the claim in ($\star$) on relevant paths passing through a neighborhood of a central node for the following two biological networks:

Network 1: *E. coli* transcriptional, and

Network 4: T-LGL signaling.

For each network we selected a few biologically relevant source-target pairs. For each such pair u_{source} and u_{target}, we found the shortest path(s) between them. For each such shortest path, a central node u_{central} was identified. We then considered the ξ-neighborhood of u_{central} such that both both u_{source} and u_{target} are on the boundary of the neighborhood, and for each such neighborhood we determined what percentage of shortest

TABLE IX: The effect of the size of the neighborhood in mediating short paths.

$\mathcal{SP}$: shortest path between u_{source} and u_{target}
$\mathcal{SP}^{+1}$: paths between u_{source} and u_{target} with one extra edge than $\mathcal{SP}$ (1-additive-approximate short path)
$\mathcal{SP}^{+2}$: paths between u_{source} and u_{target} with two extra edges than $\mathcal{SP}$ (2-additive-approximate short path)

Network name	u_{source}	u_{target}	$d_{u_{\text{source}}, u_{\text{target}}}$	u_{central}	% of $\mathcal{SP}$ with a node in ξ-neighborhood		% of $\mathcal{SP}^{+1}$ with a node in ξ-neighborhood		% of $\mathcal{SP}^{+2}$ with a node in ξ-neighborhood	
Network 1: *E. coli* transcriptional $\delta^{+}_{\text{worst}}(G) = 2$	fliAZY	arcA	4	CaiF	$\xi = 1$	100%	$\xi = 1$	71%	$\xi = 1$	59%
				crp	$\xi = 1$	100%	$\xi = 1$	100%	$\xi = 1$	100%
	fecA	aspA	6	crp	$\xi = 1$	100%	$\xi = 1$	100%	$\xi = 1$	100%
				sodA	$\xi = 1$	100%	$\xi = 1$	100%	$\xi = 1$	100%
Network 4: T-LGL signaling $\delta^{+}_{\text{worst}}(G) = 2$	IL15	apoptosis	4	GZMB	$\xi = 1$	100%	$\xi = 1$	100%	$\xi = 1$	100%
	PDGF	apoptosis	6	IL2	$\xi = 1$	80%	$\xi = 1$	82%	$\xi = 1$	93%
					$\xi = 2$	100%	$\xi = 2$	100%	$\xi = 2$	100%
				NFKB	$\xi = 1$	80%	$\xi = 1$	86%	$\xi = 1$	76%
					$\xi = 2$	100%	$\xi = 2$	100%	$\xi = 2$	100%
				Ceramide	$\xi = 1$	40%	$\xi = 1$	23%	$\xi = 1$	40%
					$\xi = 2$	100%	$\xi = 2$	100%	$\xi = 2$	100%
				MCL1	$\xi = 1$	60%	$\xi = 1$	47%	$\xi = 1$	73%
					$\xi = 2$	100%	$\xi = 2$	100%	$\xi = 2$	100%
	Stimuli	apoptosis	4	GZMB	$\xi = 1$	100%	$\xi = 1$	100%	$\xi = 1$	100%

or approximately short path (with one or two extra edges compared to shortest paths) between u_{source} and u_{target} had *all* edges in this neighborhood. The results, tabulated in Table VIII, support ($\star$).

Empirical evaluation of ($\star\star$)

We empirically investigated the size ξ of the neighborhood in claim ($\star\star$) for the same two biological networks and the same combinations of source, target and central nodes as in claim ($\star$). We considered the ξ-neighborhood of u_{central} for $\xi = 1, 2, \ldots$, and for each such neighborhood we determined what percentage of shortest or approximately short path (with one or two extra edges compared to shortest paths) between u_{source} and u_{target} involved a node in this neighborhood (not counting u_{source} and u_{target}). The results, tabulated in Table IX, show that removing the nodes in a $\xi \leq \delta^{+}_{\text{worst}}(G)$ neighborhood around the central nodes disrupts all the relevant paths of the selected networks. As $\delta^{+}_{\text{worst}}(G)$ is a small constant for all of our biological networks, this implies that the central node and its neighbors within a small distance are the essential nodes in the signal propagation between u_{source} and u_{target}.

3.3.5 Effect of hyperbolicity on structural holes in social networks

For a node $u \in V$, let $\mathsf{Nbr}(u) = \{\, v \,|\, \{u, v\} \in E \,\}$ be the set of neighbors of (*i.e.*, nodes adjacent to) u. To quantify the useful information in a social network, Ron Burt in (59) defined a measure of the *structural holes* of a network. For an undirected unweighted connected graph $G = (V, E)$ and a node $u \in V$ with degree larger than 1, this measure $\mathfrak{M}_u$ of the structural hole at u is defined as (59; 60):

$$\mathfrak{M}_u \quad \stackrel{\text{def}}{=\!=} \quad \sum_{v \in V} \left(\frac{a_{u,v} + a_{v,u}}{\max\limits_{x \neq u} \left\{ a_{u,x} + a_{x,u} \right\}} \left[1 - \sum_{\substack{y \in V \\ y \neq u,v}} \left(\frac{a_{u,y} + a_{y,u}}{\sum\limits_{x \neq u} (a_{u,x} + a_{x,u})} \right) \left(\frac{a_{v,y} + a_{y,v}}{\max\limits_{z \neq y} \{a_{v,z} + a_{z,v}\}} \right) \right] \right)$$

where $a_{p,q} = \begin{cases} 1, & \text{if } \{p, q\} \in E \\ 0, & \text{otherwise} \end{cases}$ are the entries in the standard adjacency matrix of G. By observing that $a_{p,q} = a_{q,p}$ and $\max\limits_{x \neq u} \{a_{u,x} + a_{x,u}\} = \max\limits_{z \neq y} \{a_{v,z} + a_{z,v}\} = 2$, the above equation for $\mathfrak{M}_u$ can be simplified to

$$\mathfrak{M}_u = \left| \mathsf{Nbr}(u) \right| - \frac{\sum\limits_{v,y \in \mathsf{Nbr}(u)} a_{v,y}}{\left| \mathsf{Nbr}(u) \right|} \tag{3.1}$$

Thus high-degree nodes whose neighbors are not connected to each other have high $\mathfrak{M}_u$ values. For an intuitive interpretation and generalization of (Equation 3.1), the following definition of weak and strong dominance will prove useful (*cf.* dominating set problem for graphs (61) and point domination problems in geometry (62)). A pair of distinct nodes v, y is weakly (ρ, λ)-dominated (respectively, **strongly** (ρ, λ)-dominated) by a node u provided (see Fig. Figure 8):

(a) $\rho < d_{u,v}, d_{u,y} \leq \rho + \lambda$, and

(b) for at least one shortest path $\mathcal{P}$ (respectively, **for every shortest path** $\mathcal{P}$) between v and y, $\mathcal{P}$ contains a node z such that $d_{u,z} \leq \rho$.

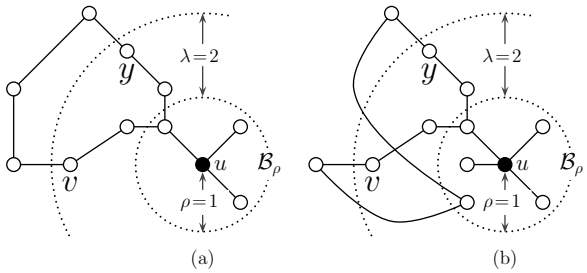


Figure 8: Illustration of weak and strong domination. (a) v, y is weakly (ρ, λ)-dominated by u since only one shortest path between v and y intersects $\mathcal{B}_\rho(u)$. (b) v, y is strongly (ρ, λ)-dominated by u since all the shortest path between v and y intersect $\mathcal{B}_\rho(u)$.

Let $\{\mathbf{v}, \mathbf{y}\} \prec^{\rho,\lambda}_{\text{weak}} \mathbf{u}$ (respectively, $\{\mathbf{v}, \mathbf{y}\} \prec^{\rho,\lambda}_{\text{strong}} \mathbf{u}$)

$$= \begin{cases} 1, & \text{if } v, y \text{ is weakly (respectively, } \textbf{strongly}\text{)} \\ & (\rho, \lambda)\text{-dominated by } u \\ 0, & \text{otherwise} \end{cases}$$

Since $\mathcal{B}_1(u) = \bigcup_{0 < j \leq 1} \mathcal{B}_j(u) = \mathsf{Nbr}(u)$, it follows that

$$\mathfrak{M}_u = \left| \cup_{0 < j \leq 1} \mathcal{B}_j(u) \right| - \frac{\sum_{v,y \in \cup_{0 < j \leq 1} \mathcal{B}_j(u)} \left(1 - \{\mathbf{v}, \mathbf{y}\} \prec^{\mathbf{0},\mathbf{1}}_{\text{weak}} \mathbf{u}\right)}{\left| \cup_{0 < j \leq 1} \mathcal{B}_j(u) \right|}$$

$$= \mathbb{E}\left[\begin{array}{l|l} \text{number of pairs of nodes} & v \text{ is selected uni-} \\ v, y \text{ such that } v, y \text{ is} & \text{formly randomly from} \\ \textbf{weakly } (0,1)\text{-dominated} & \bigcup_{0 < j \leq 1} \mathcal{B}_j(u) \\ \text{by } u & \end{array} \right]$$

$$\geq \mathbb{E}\left[\begin{array}{l|l} \text{number of pairs of nodes } v,y \text{ such that } v,y \text{ is } \textbf{strongly } (0,1)\text{-dominated by } u & v \text{ is selected uniformly randomly from } \bigcup_{0<j\leq 1}\mathcal{B}_j(u) \end{array}\right]$$

and a generalization of $\mathfrak{M}_u$ is given by (replacing $0, 1$ by ρ, λ):

$$\mathfrak{M}_{u,\rho,\lambda} = \left|\cup_{\rho<j\leq\lambda}\mathcal{B}_j(u)\right| - \frac{\sum_{v,y\in\cup_{\rho<j\leq\lambda}\mathcal{B}_j(u)}\left(1-\{\mathbf{v},\mathbf{y}\}\prec^{\rho,\lambda}_{\text{weak}}\mathbf{u}\right)}{\left|\bigcup_{\rho<j\leq\lambda}\mathcal{B}_j(u)\right|}$$

$$= \mathbb{E}\left[\begin{array}{l|l} \text{number of pairs of nodes } v,y \text{ such that } v,y \text{ is } \textbf{weakly } (\rho,\lambda)\text{-dominated by } u & v \text{ is selected uniformly randomly from } \cup_{\rho<j\leq\lambda}\mathcal{B}_j(u) \end{array}\right]$$

$$\geq \mathbb{E}\left[\begin{array}{l|l} \text{number of pairs of nodes } v,y \text{ such that } v,y \text{ is } \textbf{strongly } (\rho,\lambda)\text{-dominated by } u & v \text{ is selected uniformly randomly from } \cup_{\rho<j\leq\lambda}\mathcal{B}_j(u) \end{array}\right]$$

When the graph is hyperbolic (e.g $\delta^+_{\text{worst}}(G)$ is a constant), for moderately large λ, weak and strong dominance are essentially identical and therefore weak domination has a much stronger implication. Recall that n denotes the number of nodes in the graph G.

Our finding can be succinctly summarized as (see Fig. Figure 9 for a visual illustration):

($\star\star\star$) If $\lambda \geq \left(6\,\delta^+_{\text{worst}}(G)+2\right)\log_2 n$ then, assuming v is selected uniformly randomly from $\cup_{\rho<j\leq\lambda}\mathcal{B}_j(u)$ for any node u, the expected number of pair of nodes v, y that are **weakly** (ρ,λ)-dominated by u is precisely the same as the expected number of pair of nodes that are **strongly** (ρ,λ)-dominated by u.

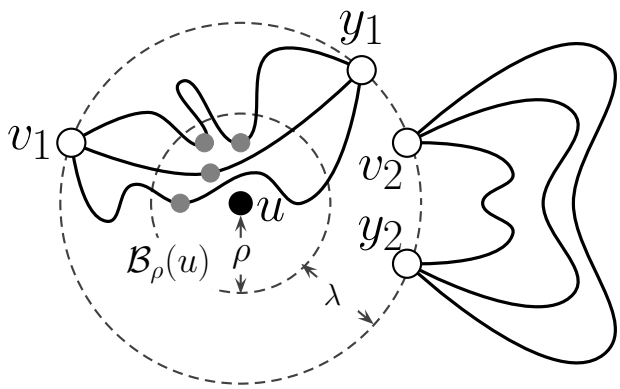


Figure 9: Visual illustration: either all the shortest paths are completely inside or all the shortest paths are completely outside of $\mathcal{B}_{\rho+\lambda}(u)$.

A mathematical justification for the claim ($\star\star\star$) is provided by Lemma 14 in Section A.2.4 of the Appendices.

An implication of ($\star\star\star$)

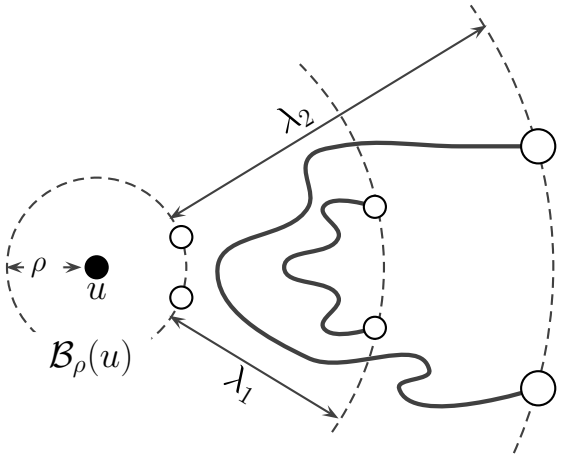


Figure 10: For hyperbolic graphs, the further we move from the central (black) node, the more a shortest path bends inward towards the central node.

TABLE X: Weak domination leads to strong domination for social networks. u is the index of the central node and

$$\nu = \frac{n_2}{n_1} = \frac{\left|\left\{(v,y) \in \mathcal{B}_{\rho+\lambda}(u) \,\middle|\, \{\mathbf{v},\mathbf{y}\} \prec^{\rho,\lambda}_{\text{strong}} \mathbf{u} = 1\right\}\right|}{\left|\left\{(v,y) \in \mathcal{B}_{\rho+\lambda}(u) \,\middle|\, \{\mathbf{v},\mathbf{y}\} \prec^{\rho,\lambda}_{\text{weak}} \mathbf{u} = 1\right\}\right|}$$

Network name	u	ρ	λ	$\left\|\mathcal{B}_{\rho+\lambda}(u)\right\|$	ν
Network 1: Dolphin social network	14	4	1	5	**80%**
	37	4	1	3	**100%**
Network 4: Books about US politics	8	4	1	4	**83%**
	3	3	1	5	**90%**
Network 7: Visiting ties in San Juan	34	4	1	4	**50%**
	9	3	1	5	**90%**

If $\lambda \geq \left(6\,\delta^{+}_{\text{worst}}(G) + 2\right)\log_2 n$ and $\mathfrak{M}_{u,\rho,\lambda} \approx \left|\mathcal{B}_{\rho+\lambda}(u)\right|$, then *almost all* pairs of nodes are strongly (ρ,λ)-dominated by u, *i.e.*, for almost all pairs of nodes $v, y \in \mathcal{B}_{\rho+\lambda}(u)$, every shortest path between v and y contains a node in $\mathcal{B}_{\rho}(u)$.

A visual illustration of this implication is in Fig. Figure 10 showing that *as λ increases the shortest paths tend to bend more and more towards the central node u for a hyperbolic network.*

Empirical verification of (⋆⋆⋆)

We empirically investigated the claim in (⋆⋆⋆) for the following three social networks from Table XVIII:

Network 1 Dolphin social network,

Network 4 Books about US politics, and

Network 7 Visiting ties in San Juan.

For each network we selected a (central) node u such that there are sufficiently many nodes in the boundary of the ξ-neighborhood $\mathcal{B}_{\xi}(u)$ of u for an appropriate $\xi = \rho + \lambda$. We then set λ to a very small value of 1, and calculated the following quantities.

- We computed the number n_1 of all pairs of nodes from $\mathcal{B}_{\xi}(u)$ that are weakly (ρ, λ)-dominated by u.
- We computed the number n_2 of all pairs of nodes from $\mathcal{B}_{\xi}(u)$ that are strongly (ρ, λ)-dominated by u.

Table X tabulates the ratio $\nu = {}^{n_2}/_{n_1}$, and shows that a large percentage of the pair of nodes that were weakly dominated were also strongly dominated by u.

CHAPTER 4

STRONG METRIC DIMENSION

4.1 Introduction

Generators of metric spaces are sets of points that every point of the space is uniquely determined by the distances from their elements. Motivated by the problem of uniquely determining the position of an intruder in a network, the concept of metric generator was introduced independently by Slater (63) and Harary and Melter (64), and has been widely investigated afterwards. It arises in many diverse areas including network discovery and verification, geographical routing protocols, robot navigation,connected joints in graphs, chemistry, etc. In this chapter we investigate a more restrictive variant of this geodesic-based property known as strong metric dimension. The strong metric dimension of graph was introduced in (65), and has been investigated in several research papers such as (66; 67; 68). The strong metric generator of a graph G can uniquely determine it by building distances between all nodes in G with respect to corresponding vectors of metric coordinates.

4.2 Strong Metric Dimension Definitions and Notations

Let $G = (V, E)$ be a given undirected graph of n nodes. To define the strong metric dimension, we will use the following notations and terminologies:

- $\mathsf{Nbr}(u) = \{ v \,|\, \{u, v\} \in E \}$ is the set of neighbors of (i.e., nodes adjacent to) a node u.
- $u \overset{\mathcal{P}}{\rightsquigarrow} v$ denotes a shortest path from between nodes u and v of length (number of edges). $d_{u,v} = \ell(u \overset{\mathsf{S}}{\rightsquigarrow} v)$.
- diam(G)= $\max_{u,v \in V} \{d_{u,v}\}$ denotes the diameter of a graph G.

- A shortest path $u \overset{\mathcal{P}}{\leftrightsquigarrow} v$ is maximal if and only if it is not properly included inside another shortest path, i.e., if and only if

$$(\forall x \in \mathsf{Nbr}(u)\ d(x,v) \leq\ d(u,v)) \wedge (\forall y \in \mathsf{Nbr}(v)\ d(y,u) \leq\ d(u,v))$$

- A node x strongly resolves a pair of nodes u and v, denoted by $x \rhd u, v$ if and only if either v is on a shortest path between x and u or either u is on a shortest path between x and v.
- A set of nodes $V' \subseteq V$ is a strongly resolving set for G, denoted by $V' \rhd G$, if and only if every distinct pair of nodes of G is strongly resolved by some node in V' , i.e., if and only if

$$\forall (u, v \in V, u \neq v)\, \exists x \in V' : x \rhd u, v$$

Then, the problem of computing the string metric dimension of a graph is defined as follows:
For an undirected graph $G = (V, E)$, the strong metric set is the smallest set of nodes $V' \subseteq V$ such that $V' \rhd G$. The strong metric dimension (sdim(G)) is $|V'|$.

4.3 Overview of Basic Concepts

First let us start with some standard definitions and assumptions for familiarizing readers with analysis of approximation algorithms.
An algorithm for a minimization problem is said to have an approximation ratio of ρ (or simply called a ρ-approximation) provided the algorithm runs in polynomial time in the size of the input and produces a solution with an objective value no larger than ρ times the value of the optimum. A computational problem P is said to be ρ-inapproximable under a complexity-theoretic assumption of $\mathbb{A}$ provided, assuming $\mathbb{A}$ to be true, there exists no ρ-approximation for P. The (standard) Boolean satisfiability problem when every clause has exactly k literals will be denoted by k-Sat. Finally, for two functions $f(n)$ and $g(n)$ of n, we say

$f(n) = O^*(g(n))$ if $f(n) = O(g(n)n^c)$ for some positive constant c.

There are three well known complexity-theoretic assumptions that we will use in this chapter, here we provide a brief introduction to these assumptions:

The $P \neq NP$ assumption Starting with the famous Cooks theorem (69) in 1971 and Karps subsequent paper in 1972 (70), the $P \neq NP$ assumption is the central assumption in structural complexity theory and algorithmic complexity analysis.

The Unique Games Conjecture (UGC) The Unique Games Conjecture, formulated by Khot in (71), is one of the most important open question in computational complexity theory. Informally speaking, the conjecture states that, assuming $P \neq NP$, a type of constraint satisfaction problems does not admit a polynomial time algorithm to distinguish between instances that are almost satisfiable from instances that are almost completely unsatisfiable. There is a large body of research works showing that the conjecture has many interesting implications and many researchers routinely assume UGC to prove non-trivial inapproximability results.

The Exponential Time Hypothesis (ETH) In an attempt to provide a rigorous evidence that the complexity of k-Sat increases with increasing k, Impagliazzo and Paturi in (72) formulated the Exponential Time Hypothesis (ETH) in the following manner. Letting $s_k = \inf\delta$: there exists$O^*(2^{\delta n})$algorithm for solving k-Sat, ETH states that $s_k > 0$ for all $k \geq 3$, i.e., k-Sat does not admit a sub-exponential time (i.e., of time $O^*(2^{O(n)})$) algorithm[1] . Eth has significant implications for worst-case time-complexity of exact solutions of search problems.

---

[1]For two functions $f(x)$ and $g(x)$ of x, $f = O(g)$ provided $\lim_{x\to\infty} f(x)/g(x) = 0$

4.4 Results and Discussion

Let $G = (V, E)$ be the given graph. It is easy to see following the approach in Khuller etal. (73) that the problem of computing the strong metric dimension sdim(G) can be reduced to an instance of the (unweighted) set-cover problem giving a $O(\log |V|)$-approximation. In this chapter, we show further improved results as summarized by the following theorems:

4.4.1 Theorem 1

Theorem 1 *(a) STR-MET-DIM admits the following type of algorithms:*

- *polynomial time 2-approximation,*
- $O^*(2^{0.287n})$*-time exact computation algorithm, and*
- $O(1.2738^k + nk)$*-time exact computation algorithm where* $sdim(G) \leq k$.

(b) Assuming the unique games conjecture [1] *(UGC)is true, STR-MET-DIM does not admit any polynomial time (2-*ϵ*)-approximation for any constant* $0 < \epsilon \leq 1$ *even if the given graph is restricted in the sense that*

(i) $diam(G) \leq 2$*, or*

(ii) G is bipartite and $diam(G) \leq 4$.

(c) Assuming P $\neq$ *NP, STR-MET-DIM does not admit any polynomial time (*$10\sqrt{5}-21-\epsilon$*)-approximation for any constant* $0 < \epsilon \leq 10\sqrt{5} - 22$ *even if the given graph is restricted in the sense that*

---

[1] *See (71) for further information on the unique games conjecture.*

(i) diam$(G) \leq 2$*, or*

(ii) G is bipartite and diam$(G) \leq 4$.

(d)) Assuming the exponential time hypothesis (Eth) is true, the following results hold for a graph G of n nodes:

(i)) there is no $O^*(2^{O(n)})$*-time algorithm for exactly computing sdim*(G)*, and*

(ii) i) if sdim$(G) \leq k$ *then there is no* $O^*(n^{O(k)})$*-time algorithm for exactly computing sdim*(G)

4.4.2 Proof of Theorem 1

This proof uses Theorem 2 whose proof is implicit in (66). However, it is not the case that Theorem 2 can be simply plugged in to get a proof of our inapproximability results. Just because a problem can be written as a node cover problem does not necessarily mean that it has the same inapproximability property for node cover since, for example, non-trivial special cases of node cover do admit efficient polynomial time solution. To show inapproximability we need to reduce appropriate hard instances of the node cover problem to that of computing *sdim*(G) (i.e., a reduction in the opposite direction) and moreover such a polynomial-time reduction must be gap-preserving in an appropriate way.

The standard minimum node cover (MNC) problem for a graph is defined as follows:

Instance:	an undirected graph $G = (V, E)$.
Valid Solution:	a set of nodes $V' \subseteq V$ such that $V' \cap u, v \neq \emptyset$ for every edge $u, v \in E$.
Goal:	minimize $\|V'\|$.
Related notation:	MNC(G) = $\min_{\forall u,v \in E: V' \cap u,v \neq \emptyset} \|V'\|$

Let $G = (V, E)$ denote the input graph of n nodes. Let $\hat{G}$ and $\tilde{G}$ be two graphs obtained from G in the following manner:

- $\hat{G} = (V, \hat{E})$ where

$$\{u, v\} \in \hat{E} \equiv u \neq v \textit{ and their distance is a maximal shortest path in } G$$

- $\tilde{G} = (\tilde{V}, \tilde{E})$ be the graph from G built in the following manner:
  - Let $u_1, u_2, ..., u_\kappa$ be the nodes in G such that, for every u_i $(1 \leq i \leq \kappa)$, there is a node $v_i \neq u_i$ in G with the property that $\mathsf{Nbr}(u_i) = \mathsf{Nbr}(v_i)$
  - Let $\bar{G} = (V, \bar{E})$ be the (edge) of G, i.e., $\{u, v\} \in \bar{E} \equiv \{u, v\} \notin E$ Then $\tilde{G}$ is constructed as follows:
  - $\tilde{V} = V \cup \{x_1, x_2, ..., x_k, y\}$ where $x_1, x_2, ..., x_k, y \notin V$
  - $\tilde{E} = \bar{E} \bigcup (\cup_{j=1}^{k} \left\{x_i, x_j\right\}) \bigcup (\cup_{\bar{y} \in \tilde{V}/\{y\}} \{y', y\})$

We recall the following result from (66)

4.4.2.1 Theorem 2

Theorem 2 ***(a)*** *sdim(G) = $MNC(\hat{G})$ and $V' \subseteq V$ is a valid solution of STR-MET-DIM on G if and only if V' is a valid solution of MNC on $\hat{G}$.*

(b) *diam($\tilde{G}$)=2 and sdim($\tilde{G}$)= $\kappa + MNC(G)$*

For further information regarding proof of Theorem 2, please read the Appendix.

4.4.2.2 Proof of Theorem 1(a)

Since sdim(G) = $MNC(\hat{G})$, and both G and $\hat{G}$ have the same number of nodes, the claim follows by applying known algorithms for node cover on $\hat{G}$. More precisely,

- the 2-approximation follows from a well known 2-approximation algorithm for MNC (74),

- the $O^*(2^{0.287n})$-time exact solution algorithm follows from the $O^*(2^{0.287n})$-time exact algorithm for maximum independent set4 problem in (75), and
- the $O^*(1.2738^k + nk)$-time exact computation algorithm follows from the $O^*(1.2738^k + nk)$-time exact algorithm for minimum node cover of $\hat{G}$ provided $MNC(\hat{G}) \leq k$(76).

4.4.2.3 Proof of Theorem 1(b)

Consider the standard Boolean satisfiability problem (SAT) and let ϕ be an input instance of SAT. Our starting point is the following inapproximability result proved Khot and Regev (77):

> [setting k=2] Assuming UGC is true, there exists a polynomial time algorithm that transforms a given instance ϕ of SAT to an input instance graph G = (V,E) of MNC with n nodes such that, for any arbitrarily small constant $0 < \epsilon < \frac{1}{4}$, the following holds:

$(\star)$

(YES case) if ϕ is satisfiable then MNC(G)$\leq (\frac{1}{2} + \epsilon)$n, and

(NO case) if ϕ is not satisfiable then MNC(G)$\geq$ (1 - ϵ)n.

Consider such an instance G of MNC as generated by the above transformation.We first construct the following graph $G^+ = (V^+, E^+)$ from G. Let k = 1 + $\lfloor log_2^n \rfloor$ and let b(j) = $b_{k-1}(j)b_{k-2}(j)...b_1(j)b_0(j)$ be the binary representation of an integer $j \in \{1, 2, ..., n\}$ using exactly k bits (e.g., if n = 5 then b(3) = $\overset{b_2(3)}{0}\ \overset{b_1(3)}{1}\ \overset{b_0(3)}{1}$).Let $u_1, u_2, ..., u_n$ be an arbitrary ordering of the nodes in V. Then,

- $V^+ = V \bigcup V_1^+$ where $V_1^+ = \{v_1, v_2, ..., v_{k-1}, y\}$ is a set of k new nodes, and
- $E^+ = E \bigcup (\cup_{j=1}^{n} \left\{u_j, v_l\right\} | b_l(j) = 1) \bigcup (\cup_{j=1}^{k-1} \left\{y, v_j\right\})$

Thus $|V^+|$= n+k and $|E^+|$¡$|E|+\frac{nk}{2}$+k. Now, note that if $V' \subset V$ is a solution of MNC on G, then $V' \cup V'_+$ is a solution of MNC on G^+, implying MNC(G^+)$\leq$ MNC(G) +k, and, conversely, if $V' \subset V^+$ is a solution of MNC on G^+, then $V' \setminus V_1^+$ is a solution of MNC on G, implying MNC(G) $\leq$ MNC(G^+). Combining the above inequalities with that in ($\star$), we have

$$(\star\star)\quad \begin{array}{ll} \textbf{(YES case)} & \text{if } \phi \text{ is satisfiable then MNC}(G^+) \leq (\frac{1}{2} + \epsilon)\text{n} + log_2^n + 1, \text{ and} \\ \textbf{(NO case)} & \text{if } \phi \text{ is not satisfiable then MNC}(G^+) \geq (1 - \epsilon)\text{n}. \end{array}$$

We now build the graph $\widetilde{G^+} = (\widetilde{V^+}, \widetilde{E^+})$ from G using the construction in Theorem 2(b).

Claim 1. No two nodes in $\widetilde{G^+}$ have the same neighborhood.

Proof. The following careful case analysis proves the claim:

- For any $i \neq j$, since $b(i) \neq b(j)$, there exists an index t such that $b_t(i) \neq b_t(j)$, say $b_t(i)$=0 and $b_t(j)$ = 1. Thus,$\mathsf{Nbr}(u_i) \neq \mathsf{Nbr}(u_j)$ since $v_t \in \mathsf{Nbr}(u_i)$ but $v_t \notin \mathsf{Nbr}(u_j)$
- Since $b(i) \neq 0$ for any i and b(1),b(2),... ,b(n) are distinct binary numbers each of exactly k bits,for any $t \neq t'$ there is an index i such that $b_t(i) \neq b_{t'}(i)$, say $b_t(i)$=0 and $b_{t'}(i)$=1. Thus,$\mathsf{Nbr}(v_t) \neq \mathsf{Nbr}(v_{t'})$ since $u_i \in \mathsf{Nbr}(v_{t'})$ but $u_i \notin \mathsf{Nbr}(v_t)$
- For any i and j,$\mathsf{Nbr}(u_i) \neq \mathsf{Nbr}(v_j)$ since $y \in \mathsf{Nbr}(v_j)$ but $y \notin \mathsf{Nbr}(u_i)$
- For any i,$b(i \neq 0)$ and thus there exists an index j such that $b_j(i) = 1$.This implies $u_j \in \mathsf{Nbr}(v_i)$ but $u_j \notin \mathsf{Nbr}(y)$ and therefore $\mathsf{Nbr}(v_i) \neq \mathsf{Nbr}(y)$
- Since G is a connected graph, for every node u_i there exists a node u_j such that $\{u_i, u_j\} \in E^+$.Thus,$u_j \in \mathsf{Nbr}(u_i)$ but $u_j \notin \mathsf{Nbr}(y)$), implying $\mathsf{Nbr}(u_i) \neq \mathsf{Nbr}(y)$

Thus,no two nodes in G^+ have the same neighborhood, implying $\kappa = 0$ and sdim($\widetilde{G^+}$)= MNC(G^+). Thus,setting $\epsilon' = \epsilon + \frac{log_2^n + 1}{n} > \epsilon$ to be any arbitrarily small constant, it follows from ($\star\star$) that

$$(\star\star\star)\quad \begin{array}{ll} \textbf{(YES case)} & \text{if } \phi \text{ is satisfiable then MNC}(G^+) < (\frac{1}{2} + \epsilon')\text{n}, \text{ and} \\ \textbf{(NO case)} & \text{if } \phi \text{ is not satisfiable then MNC}(G^+) \geq (1 - \epsilon')\text{n}. \end{array}$$

This proves Theorem 1(b)(i) since diam($\widetilde{G^+}$) =2 by Theorem 2(b).

To prove Theorem 1(b)(ii), we modify the graph $\widetilde{G^+}$ to a new graph $G' = (V', E')$ by splitting every edge

into a sequence of two edges, i.e., for every edge $\{u, v\}$ in $\widetilde{G^+}$ we add a new node x_{uv} in G' and replace the edge u,v by the two edges $\{u, x_{uv}\}$ and $\{v, x_{uv}\}$. Clearly G' is bipartite since all its cycles are of even length and diam(G')$\leq$ 2 + diam($\widetilde{G^+}$)=4.

Claim 2. sdim($\widetilde{G^+}$) = MNC($\widetilde{\widetilde{G^+}}$) =MNC($\widehat{G'}$)= sdim($G'$).

Proof. No maximal shortest path in G' ends at a node x_{uv} for any distinct pair of nodes u and v. Indeed, if a maximal shortest path $\wp$ from some node z ends at x_{uv},it must use one of the two edges $\{u, x_{uv}\}$ or $\{v, x_{uv}\}$, say $\{u, x_{uv}\}$. Then adding the edge $\{v, x_{uv}\}$ to the path provide a shortest path between v and z and thus $\wp$ was not maximal. Using this and the construction in Theorem 2(a), we have $\widehat{\widetilde{G^+}} = \widehat{G'}$ and therefore sdim($\widetilde{G^+}$) = MNC($\widehat{\widetilde{G^+}}$) =MNC($\widehat{G'}$)= sdim($G'$) As a result, the inapproximability result for $\widetilde{G^+}$ directly translates to that for G' and concludes the proof.

4.4.2.4 Proof of Theorem 1(c)

The same proof as in (b) works provided, instead of the result in (77), our starting point is the following result shown by Dinur and Safra (78):

> Assuming $P \neq NP$, there exists a polynomial time algorithm that transforms a given instance ϕ of Sat to an input instance graph $G = (V, E)$ of MNC with n nodes such that, for any constant $0 < \epsilon < 168\sqrt{5}$ and for some $0 < \alpha < 2n$, the following holds:
>
> ($\star$) **(YES case)** if ϕ is satisfiable then MNC(G)$\leq$ ($\frac{\sqrt{5}-1}{2}$ + ϵ')α, and
>
> **(NO case)** if ϕ is not satisfiable then MNC(G)$\geq$ ($\frac{71-31\sqrt{5}}{2}$ - ϵ')α.

4.4.2.5 Proof of Theorem 1(d)

We first show how to prove Theorem 1(d)(i). Suppose, for the sake of contradiction, that there does exist a $O^*(2^{O(n)})$-time algorithm that exactly computes $sdim(G)$. We start with an instance ϕ of 3-Sat having n variables and m clauses. The sparsification lemma in (79) proves the following result:

> for every constant $\epsilon > 0$, there is a constant $c > 0$ such that there exists a $O^*(2^{\epsilon n})$-time algorithm that produces from ϕ a set of t instances $\phi_1, ..., \phi_t$ of 3-Sat on these n variables with the following properties:
>
> - $t \leq 2^{\epsilon n}$,
> - each ϕ_j is an instance of 3-Sat with $n_j \leq n$ variables and $m_j \leq cn$ clauses, and
> - ϕ is satisfiable if and only if at least one of $\phi_1, ..., \phi_t$ is satisfiable.

For each such above-produced 3-Sat instance ϕ_j, we now use the classical textbook reduction from 3-Sat to the node cover problem producing an instance $G = (V, E)$ of MNC of $|V| = 3n_j + 2m_j \leq (3 + 2c)n$ nodes and $|E| = n_j + m_j \leq (1 + c)n$ edges such that ϕ_j is satisfiable if and only if $MNC(G) = n_j + 2m_j$. Moreover, it is also easy to check that this classical reduction does not produce two nodes in V that have the same neighborhood. Thus, setting $\kappa = 0$ in Theorem 2(b) we get $sdim(\widetilde{G}) = MNC(G)$ where $\widetilde{G}$ is a graph with $\widetilde{n} = |\widetilde{V}| = |V| + 1 \leq (3 + 2c)n + 1$ nodes. By assumption, we can compute sdim($\widetilde{G}$) in $O^*(2^{O(\widetilde{n})})$ time, and and consequently $MNC(G)$ in $O^*(2^{O(n)})$ time, which leads us to decide in $O^*(2^{O(n)})$ time if ϕ_j is satisfiable. Since $t \leq 2^{\epsilon n}$ for any constant $\epsilon > 0$, this provides a $O^*(2^{O(n)})$-time algorithm for 3-Sat, contradicting ETH. To prove Theorem 1(d)(ii) suppose again, for the sake of contradiction, that there exists a $O^*(n^{O(k)})$-time algorithm for exactly computing $sdim(G)$ if $sdim(G) \leq k$. Our proof is very similar to the previous one, but this time we start with the following lower bound result on parameterized complexity (e.g., see [(80), Theorem 14.21]):

assuming ETH to be true, if $Mnc(G) \leq k$ then there is no $O^*(n^{O(k)})$-time algorithm for exactly computing Mnc(G).

Using the encoding as described in part (b) of this proof with the corresponding Claim 2, we can set $\kappa = 0$ in Theorem 2(b) to obtain the graph $\widetilde{G^+} = (\widetilde{V^+}, \widetilde{E^+})$ such that $\widetilde{n^+} = |\widetilde{V^+}| = |V| + (1 + \lfloor \log_2 n \rfloor + 1) = n + \lfloor \log_2 n \rfloor + 2$ and $sdim(\widetilde{G^+}) = MNC(G)$. By our assumption, we can compute $sdim(\widetilde{G^+})$ in $O^*(\widetilde{n^+}^{O(k)})$-time algorithm if $sdim(G) \leq k$. This then provides an algorithm running in $O^*(\widetilde{n^+}^{O(k)}) = O^*(n^{O(k)})$ time if $MNC(G) = sdim(G) \leq k$, contradicting ETH.

Although, the ability of strong metric set to uniquely identify a graph through a subset of nodes is quite useful in many applications as we discussed, it also raises concerns over privacy. In the next chapter, we look into another property in complex networks that has been inspired by and derived from strong metric dimension and measures the privacy in networks.

CHAPTER 5

PRIVACY IN SOCIAL NETWORKS AND (K, ℓ)-ANONYMITY

5.1 Introduction

Due to a significant growth of applications of graph-theoretic methods to the field of social sciences in recent days, it is by now a standard practice to use the concepts and terminologies of network science to those social networks that focus on interconnections between people. However, social networks in general may represent much more than just networks of interconnections between people. Rapid evolution of popular social networks such as *Facebook*, *Twitter* and *LinkedIn* have rendered modern society heavily dependent on such virtual platforms for their day-to-day operation. The powers and implications of social network analysis are indeed *indisputable*; for example, such analysis may uncover previously unknown knowledge on community-based involvements, media usages and individual engagements. However, all these benefits are *not* necessarily cost-free since a malicious individual could compromise privacy of users of these social networks for harmful purposes that may result in the disclosure of sensitive data (attributes) that may be linked to its users, such as node degrees, inter-node distances or network connectivity. A natural way to avoid this consists of an "anonymization process" of the relevant social network in question. However, since such anonymization processes may *not* always succeed, an important research goal is to be able to quantify and measure how much privacy a given social network can achieve. Towards this goal, the recent work in (81) aimed at evaluating the *resistance* of a social network against active privacy-violating attacks by introducing and studying theoretically a new and meaningful privacy measure for social networks. This privacy measure arises from the concept of the so-called k-metric antidimension of graphs.

Given a connected simple graph $G = (V, E)$, and an ordered sequence of nodes $S = (v_1, \ldots, v_t)$, the *metric representation* of a node u that is *not* in S with respect to S is the vector (of t components) $\mathbf{d}_{u,-S} =$

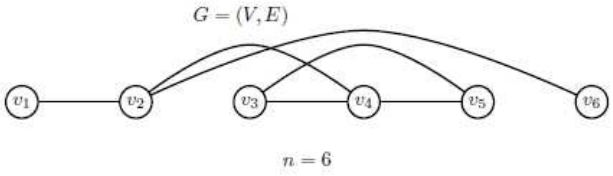
$(\text{dist}_{u,v_1}, \ldots, \text{dist}_{u,v_t})$, where $\text{dist}_{u,v}$ represents the length of a shortest path between nodes u and v. The set S is then a *k-antiresolving set* if k is the largest positive integer such that for every node v not in S there also exist *at least* other $k-1$ different nodes $v_{j_1}, \ldots, v_{j_{k-1}}$ not in S such that $v, v_{j_1}, \ldots, v_{j_{k-1}}$ have the *same* metric representation with respect to S (*i.e.*, $\mathbf{d}_{v,-S} = \mathbf{d}_{v_{j_1},-S} = \cdots = \mathbf{d}_{v_{j_{k-1}},-S}$). The *k-metric antidimension* of G is defined to be value of the minimum cardinality among all the k-antiresolving sets of G (81). If a set of attacker nodes S represents a k-antiresolving set in a graph G, then an adversary controlling the nodes in S cannot *uniquely* re-identify other nodes in the network (*based on the metric representation*) with probability higher than $1/k$. However, given that S is unknown, any privacy measure for a social network should quantify over *all* possible subsets S of nodes. In this sense, a social network G meets (k, ℓ)-*anonymity* with respect to active attacks to its privacy if k is the smallest positive integer such that the k-metric antidimension of G is no more than ℓ. In this definition of (k, ℓ)-anonymity the parameter k is used for a *privacy threshold*, while the parameter ℓ represents an *upper bound* on the expected number of attacker nodes in the network. Since attacker nodes are in general difficult to inject without being detected, the value ℓ could be estimated based on some statistical analysis of other known networks. A simple example that explains the role of k and ℓ is as follows: Consider a complete network K_n on n nodes in which every node is connected with every other node. It is readily seen that for any $0 < \ell < n$, this network meets $(n - \ell, \ell)$-anonymity. In other words, this means that a social network K_n guarantees that a user cannot be re-identified (based on the metric representation) with a probability higher than $1/(n - \ell)$ by an adversary controlling at most ℓ attacker nodes. Chatterjee *et al.* in (82) (see also (83)) formalized and analyzed the computational complexities of several optimization problemsmotivated by the **(k, ℓ)-anonymity** of a network as described in (81). In this chapter, we consider three of these optimization problems from (82), namely Problems 1–3 as defined in Section 5.2.

To avoid any possible misgivings or confusions regarding the technical content of this chapter as well as to help the reader towards understanding the remaining content, we believe the following comments and explanations may be relevant.

- The computational complexity investigations in this chapter has nothing to do with the model in the paper by Backstrom *et al.* (84). The notion of active attack is very different in that paper, and therefore the computational problems that arise in that paper are very different from those in the current work and in fact incomparable. Finally, the goal of this work is not to compare various network privacy models but to investigate, theoretically and empirically, the model in (81).
- This chapter does *not* introduce any new privacy model or measure, but simply investigates, both theoretically and empirically, computational problems for a model that is published in "*Information Sciences*, 328, 403–417, 2016" (reference (81)).
- The network privacy model we investigate was introduced in (81) and therefore the best option for clarification of any confusion regarding the model would be to look at that reference. However, we provide the following clarification just in case. In this model, nobody is trying to prevent adversaries. Informally, the privacy measure only gives a "measure" on how much secure a graph is against active attacks, *i.e.*, a probability with which we can assert that, if there are controlled nodes in a graph, then we can in some sense know which is the probability to be reidentified in such graph (for details please see the texts preceding and following the statements of Problems 1–3 in Section 5.2). No new nodes are added at all. This is not a problem that involves dynamic graphs.

5.2 Basic notations, relevant background and problem formulations

Let $G = (V, E)$ be the undirected input network over n nodes $v_1, \ldots, v_n$. The authors in (82) formalized and analyzed the computational complexities of several optimization problems motivated by the (k, ℓ)-



dist_{v_i,v_j} values

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

Figure 11: An example for illustration of some basic definitions and notations in Section 5.2.

anonymity of a network as described in (81). The notations and terminologies from (82) relevant for this chapter are as follows (*see Figure 11 for an illustration*)

- $\mathbf{d}_{v_i} = (\text{dist}_{v_i,v_1}, \text{dist}_{v_i,v_2}, \ldots, \text{dist}_{v_i,v_n})$ denotes the metric representation of a node v_i. For example, in Figure 11, $\mathbf{d}_{v_1} = (0, 1, 3, 2, 3, 2)$.

- $\mathsf{Nbr}(v_\ell) = \left\{ v_j \mid \left\{v_\ell, v_j\right\} \in E \right\}$ is the (open) *neighborhood* of node v_ℓ in $G = (V, E)$. For example, in Figure 11, $\mathsf{Nbr}(v_2) = \{ v_1, v_4, v_6 \}$.

- For a subset of nodes $V' = \left\{v_{j_1}, v_{j_2}, \ldots, v_{j_t}\right\} \subset V$ with $j_1 < j_2 < \cdots < j_t$ and any other node $v_i \in V \setminus V'$, $\mathbf{d}_{v_i,-V'} = (\text{dist}_{v_i,v_{j_1}}, \text{dist}_{v_i,v_{j_2}}, \ldots, \text{dist}_{v_i,v_{j_t}})$ denotes the metric representation of v_i with respect to V'. The notation is further generalized by defining $\mathcal{D}_{V'',-V'} = \{ \mathbf{d}_{v_i,-V'} \mid v_i \in V'' \}$ for any $V'' \subseteq V \setminus V'$. For example, in Figure 11, $\mathbf{d}_{v_3,-\{v_1,v_5,v_6\}} = (\underset{v_1}{3}, \underset{v_5}{1}, \underset{v_6}{3})$ and $\mathcal{D}_{\{v_2,v_3\},-\{v_1,v_5,v_6\}} = \{(\overbrace{\underset{v_1}{1}, \underset{v_5}{2}, \underset{v_6}{1}}^{\text{from } v_2}), (\overbrace{\underset{v_1}{3}, \underset{v_5}{1}, \underset{v_6}{3}}^{\text{from } v_3})\}$.

- A partition $\Pi' = \left\{V'_1, V'_2, \ldots, V'_\ell\right\}$ of $S' \subseteq V$ is called a *refinement* of a partition $\Pi = \{V_1, V_2, \ldots, V_k\}$ of $S \supseteq S'$, denoted by $\Pi' \prec_r \Pi$, provided Π' can be obtained from Π in the following manner:

  - For every node $v_i \in \left(\cup_{t=1}^{k} V_t\right) \setminus \left(\cup_{t=1}^{\ell} V'_t\right)$, remove v_i from the set in Π that contains it.

  - *Optionally*, for every set V_ℓ in Π, replace V_ℓ by a partition of V_ℓ.

  - Remove empty sets, if any.

  For example, for Figure 11, $\{\{v_2\}, \{v_3\}, \{v_4, v_5\}\} \prec_r \{\{v_1, v_2, v_3\}, \{v_4, v_5\}\}$.

- The following notations pertain to the equality relation (an equivalence relation) over the set of (same length) vectors $\mathcal{D}_{V\setminus V',-V'}$ for some $\emptyset \subset V' \subset V$:

  - ▹ The set of equivalence classes, which forms a partition of $\mathcal{D}_{V\setminus V',-V'}$, is denoted by $\Pi^{=}_{V\setminus V',-V'}$. For example, in Figure 11, $\mathcal{D}_{\{v_2,v_3,v_4,v_5\},-\{v_1,v_6\}} = \{(\overbrace{1,1}^{\text{from } v_2}_{v_1\ v_6}), (\overbrace{3,3}^{\text{from } v_3}_{v_1\ v_6}), (\overbrace{2,2}^{\text{from } v_4}_{v_1\ v_6}), (\overbrace{3,3}^{\text{from } v_5}_{v_1\ v_6})\}$ and $\Pi^{=}_{\{v_2,v_3,v_4,v_5\},-\{v_1,v_6\}} = \Big\{\{(\overbrace{1,1}^{\text{from } v_2}_{v_1\ v_6})\}, \{(\overbrace{2,2}^{\text{from } v_4}_{v_1\ v_6})\}, \{(\overbrace{3,3}^{\text{from } v_3}_{v_1\ v_6}), (\overbrace{3,3}^{\text{from } v_5}_{v_1\ v_6})\}\Big\}$.
  - ▹ Abusing terminologies slightly, two nodes $v_i, v_j \in V \setminus V'$ will be said to belong to the *same* equivalence class if $\mathbf{d}_{v_i,-V'}$ and $\mathbf{d}_{v_j,-V'}$ belong to the same equivalence class in $\Pi^{=}_{V\setminus V',-V'}$, and thus $\Pi^{=}_{V\setminus V',-V'}$ also defines a partition into equivalence classes of $V \setminus V'$. For example, in Figure 11, v_3 and v_5 belong to the same equivalence class in $\Pi^{=}_{\{v_2,v_3,v_4,v_5\},-\{v_1,v_6\}}$ and $\Pi^{=}_{\{v_2,v_3,v_4,v_5\},-\{v_1,v_6\}}$ also defines the partition $\{\{v_2\}, \{v_4\}, \{v_3, v_5\}\}$.
  - ▹ The *measure* of the equivalence relation is defined as $\mu\left(\mathcal{D}_{V\setminus V',-V'}\right) \stackrel{\text{def}}{=} \min_{\mathcal{Y}\in\Pi^{=}_{V\setminus V',-V'}} \{\,|\mathcal{Y}|\,\}$. Thus, if a set S is a k-antiresolving set, then $\mathcal{D}_{V\setminus S,-S}$ defines a partition into equivalence classes whose measure is k. For example, in Figure 11, $\mu\left(\Pi^{=}_{\{v_2,v_3,v_4,v_5\},-\{v_1,v_6\}}\right) = 1$.

By using the terminologies mentioned above, the following three optimization problems were formalized and studied in (82). We need to stress that one really needs to study the three different problems and consequently the three objectives (namely, k_{opt}, $\mathcal{L}^{\geq k}_{\text{opt}}$ and $\mathcal{L}^{=k}_{\text{opt}}$) separately because they are motivated by different considerations as explained before and after the problem definitions and as stated in (★), (⋈) and (♠). Informally and briefly, Problem 1 and k_{opt} are used to provide an absolute privacy violation bound assuming the attacker can control as many nodes as it needs, restricting the number of attacker nodes employed by the adversary leads to Problem 2, and Problem 3 is motivated by a type of trade-off question between (k, ℓ)-anonymity vs. (k', ℓ')-anonymity. Thus, it is simply not possible to combine them into fewer than three problems.

Problem 1 (metric anti-dimension or ADIM)) *Find a subset of nodes V' such that* $k_{\text{opt}} = \mu\left(\mathcal{D}_{V\setminus V',-V'}\right) = \max_{\emptyset\subset S\subset V}\{\mu\left(\mathcal{D}_{V\setminus S,-S}\right)\}$.

A solution of Problem 1 asserts the following:

(⋆) Assuming that there is *no* restriction on the number of nodes that can be controlled by an adversary, the following statements hold:

(*a*) The network administrator *cannot* guarantee that an adversary will not be able to uniquely re-identify any node in the network (based on the metric representation) with probability $1/k_{\text{opt}}$ or less.

(*b*) It *is* possible for an adversary to uniquely re-identify k_{opt} nodes in the network (based on the metric representation) with probability $1/k_{\text{opt}}$.

Thus, informally, Problem 1 and k_{opt} give an absolute privacy violation bound assuming the attacker can control as many nodes as it needs. In practice, however, the number of attacker nodes employed by the adversary *may* be restricted. This leads us to Problem 2.

Problem 2 ($k_{\geq}$-metric anti-dimension or ADIM$_{\geq k}$) *Given a positive integer k, find a subset* $V_{\text{opt}}^{\geq k}$ *of nodes of* minimum *cardinality* $\mathcal{L}_{\text{opt}}^{\geq k} = \left|V_{\text{opt}}^{\geq k}\right|$*, if one such subset at all exists, such that* $\mu(\mathcal{D}_{V\setminus V_{\text{opt}}^{\geq k},-V_{\text{opt}}^{\geq k}}) \geq k$.

Similar to (⋆), a solution of Problem 2 (if it exists) asserts the following:

(⋈) Assuming that an adversary may control up to α nodes, the following statements hold:

(*a*) If $\alpha < \mathcal{L}_{\text{opt}}^{\geq k}$ then the network administrator *can* guarantee that an adversary will not be able to uniquely re-identify any node in the network (based on the metric representation) with probability $1/k$ or less.

(*b*) If $\alpha \geq \mathcal{L}_{\text{opt}}^{\geq k}$ then the network administrator *cannot* guarantee that an adversary will not be able to uniquely re-identify any node in the network (based on the metric representation) with probability $1/k$ or less.

(*c*) If $\alpha \geq \mathcal{L}_{\text{opt}}^{\geq k}$ then it *is* possible for an adversary to uniquely re-identify a subset of β nodes in the network (based on the metric representation) with probability $1/\beta$ for some $\beta \geq k$ (note that β may be much larger compared to k).

The remaining third problem is motivated by the following trade-off question between (k, ℓ)-anonymity vs. (k', ℓ')-anonymity: if $k' > k$ but $\ell' < \ell$ then (k', ℓ')-anonymity has *smaller* privacy violation probability $1/k' < 1/k$ compared to (k, ℓ)-anonymity but can only tolerate attack on *fewer* $\ell' < \ell$ number of nodes.

Problem 3 ($k_=$-metric antidimension or $\textsc{Adim}_{=k}$) *Given a positive integer k, find a subset* $V_{\text{opt}}^{=k}$ *of nodes of* minimum *cardinality* $\mathcal{L}_{\text{opt}}^{=k} = \left|V_{\text{opt}}^{=k}\right|$*, if one such subset at all exists, such that* $\mu(\mathcal{D}_{V \setminus V_{\text{opt}}^{=k}, -V_{\text{opt}}^{=k}}) = k$.

One can describe assertions to a solution of Problem 2 (if it exists) in a manner similar to that in (⋆) and (⋈). Chatterjee *et al.* in (82) studied the computational complexity aspects of Problems 1–3. They provided efficient (polynomial-time) algorithms to solve Problems 1 and 2 and showed that Problem 3 is *provably* computationally hard for exact solution but admits an efficient approximation for the particular case of $k = 1$ (see Algorithm II). Since we use this approximation algorithm for $k = 1$, we explicitly state below the implication of a solution of $\textsc{Adim}_{=1}$ (note that a solution of $\textsc{Adim}_{=1}$ always exists and $\mathcal{L}_{\text{opt}}^{=1}$ is trivially at most $n - 1$):

(♠) It suffices for an adversary to control a *suitable* subset of $\mathcal{L}_{\text{opt}}^{=1}$ nodes in the network to *uniquely* re-identify at least one node in the network (based on the metric representation) with *absolute certainty* (*i.e.*, with a probability of one).

5.3 Theoretical and Empirical Results

5.3.1 Theoretical Result

Suppose that a given graph G is a "k'-metric antidimensional" graph, *i.e.*, k' is the largest positive integer such that G has *at least* one k'-antiresolving set. Then obviously G does *not* contain any k''-antiresolving set

for every $k'' > k'$. In contrast, it is not *a priori* clear if G contains k-metric antiresolving sets for any $k < k'$. For instance, a complete graph K_n on n nodes is $(n-1)$-metric antidimensional and moreover, for every $1 \leq k \leq n-1$, there exists a set of nodes in K_n which is a k-antiresolving set. *Au contraire*, if we consider the wheel graph $W_{1,n}$ (see Figure 12 for an illustration for $n = 16$), it is easy to see that the central node v_n is the *unique* n-antiresolving set, 1-antiresolving and 2-antiresolving sets exist, 3-antiresolving sets also exist (if n is larger than 5), but *no* k-antiresolving set exists for $4 \leq k \leq n-1$. This motivates the following research question:

> *For a given class of k'-metric antidimensional networks, can we decide if they also have k-antiresolving sets for all $1 \leq k \leq k' - 1$?*

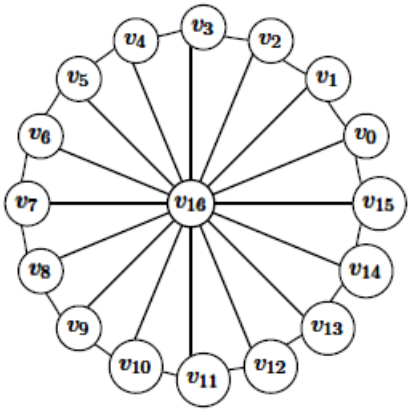


Figure 12: The wheel graph $W_{1,n}$ for $n = 16$.

The following theorem answers the question affirmatively for all networks without a cycle.

Theorem 1 *If T is a k'-metric antidimensional tree, then for every $1 \leq k \leq k'$ there exists a k-antiresolving set for T.*

Some consequences of Theorem 1

Some consequences of the above result in relation to the (k, ℓ)-anonymity measure are as follows. Clearly, since trees have nodes of degree one (called leaves), it is always possible to identify at least one node of the tree (85). However, if the network manager introduces some "fake" nodes as leaves, then this advantage for the adversary is avoided. In this sense, the result above asserts that an adversary will never be sure that the set of nodes which it could control will always identify at least one node of the given tree. Another related interesting observation is that for this to happen, the tree must be k-metric antidimensional for some $k \geq 2$, otherwise the tree is *completely insecure*. A characterization of that trees which are 1-metric antidimensional (graphs that contain only 1-antiresolving sets) was given in (86). The topology need not be "fully" controlled by a network manager, but can be influenced by adding extra nodes.

Proof of Theorem 1

We will use the following result from (86) in our proof.

Lemma 2(86) *Any k-antiresolving set S in a tree T with $k \geq 2$ induces a connected subgraph of T.*

Since Problem 1 was shown to be solvable in polynomial time in (82), we may assume that we know the value k' for which the tree T is k'-metric antidimensional. If $k = 1$ or $k = k'$ then a k-antiresolving set for T clearly exists. We may also assume $k > 1$, since otherwise our result follows trivially. Suppose that $k = k' - 1$ and let S be a k'-antiresolving set of minimum cardinality for T. By Lemma 2, S induces a connected subgraph of T. Moreover, according to the definition of a k-antiresolving set, there exists an equivalence class $Q \in \Pi^{=}_{V \setminus S, -S}$ such that $|Q| = k'$. Select $v \in S$ such that $\mathsf{Nbr}(v) \setminus S \neq \emptyset$ and let $v_1, v_2, \ldots, v_r \in \mathsf{Nbr}(v) \setminus S$ for some $r \geq 1$. Clearly, the set $A_1 = \{v_1, v_2, \ldots, v_r\}$ forms an equivalence class of $\Pi^{=}_{V \setminus S, -S}$. Moreover, the set $A_2 = \bigcup_{i=1}^{r} \mathsf{Nbr}(v_i) \setminus \{v\}$, if not empty, also forms an equivalence class of $\Pi^{=}_{V \setminus S, -S}$. Figure 13 shows two examples which are useful to clarify all the notations of this proof (recall that the *eccentricity* of a node v is the maximum over the set of distances between v to all other nodes in the graph).

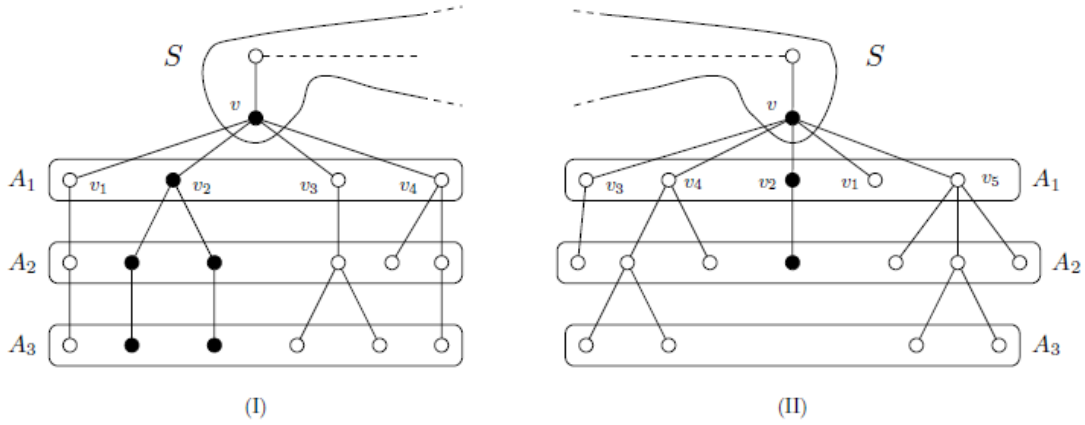


Figure 13: Two auxiliary trees. Notice that eccentricity of v in the subtrees is three in both cases. The set S is a 4-antiresolving set. The nodes of the subtree T_2 are shown in bold in both trees.

Assume that T is rooted at node v and, for every $v_i \in A_1$, let T_i be the subtree of T with node set $V(T_i)$ formed by v, v_i, and the set of descendants of v_i. Let e_i be the eccentricity of v in T_i for $1 \leq i \leq r$. Moreover, let A_j be the subset of nodes x in $\bigcup_{i=1}^{r} V(T_i)$ such that $\text{dist}_{v,x} = j$ for every $1 \leq j \leq \max\{e_i : 1 \leq i \leq r\}$. Observe that each A_j, with $1 \leq j \leq \max\{e_i : 1 \leq i \leq r\}$, is an equivalence class of $\Pi^{=}_{V\setminus S,-S}$ and thus, $|A_j| \geq k'$ since otherwise S is *not* a k'-antiresolving set. Moreover, without loss of generality, we can assume there exists a set A_q such that $|A_q| = k'$ (e.g., in Figure 13 the sets A_1 and A_4). If there is no such set, then we choose another node v' of T for which this situation happen. If there is no such node v' at all, then the cardinality of every equivalence class of $\Pi^{=}_{V\setminus S,-S}$ is *strictly* larger than k', which contradicts the definition of a k'-antiresolving set. We now consider the following situations.

Case 1: $e_1 = e_2 = \cdots = e_r$ (e.g., in Figure 13(I) all the eccentricities are equal to 3). Notice that in this case $A_j \cap V(T_i) \neq \emptyset$ for every $1 \leq j \leq \max\{e_i : 1 \leq i \leq r\}$ and every $1 \leq i \leq r$. Moreover, there exist α, β such that $|A_\alpha \cap V(T_\beta)| = 1$ (e.g., in Figure 13(I) $\alpha = 1$ and β can take any value between 1 and 4). Thus, for the set $S' = S \cup V(T_\beta)$ it follows that $A_\alpha \setminus V(T_\beta)$ is an equivalence class of the equivalence relation $\Pi^{=}_{V\setminus S',-S'}$ and

$|A_\alpha - V(T_\beta)| = k' - 1$. Moreover, for every other equivalence class X of $\Pi^{=}_{V\setminus S', -S'}$ it follows $|X| \geq k' - 1 = k$. Thus, X is a $(k' - 1)$-antiresolving set. Clearly, X could not be of minimum cardinality.

Case 2: There are at least two subtrees T_i and T_j such that $e_i \neq e_j$. Without loss of generality, assume that $e_1 \leq e_2 \leq \cdots \leq e_r$. As in Case 1, there exist γ such that $|A_\gamma| = k'$ (e.g., in Figure 13(II) $\alpha = 3$). Let $S_1 = S \cup V(T_1)$ (note that T_1 is the subtree in which v has the minimum eccentricity). If $|A_j^{(1)}| \geq k'$ for every $A_j^{(1)} = A_j \setminus V(T_1)$ with $1 \leq j \leq e_1$, then $\gamma > e_1$ and thus S_1 is *also* a k'-antiresolving set. Hence, we consider $S_2 = S_1 \cup V(T_2)$ (note that T_2 is the subtree in which v has the second minimum eccentricity). If $|A_j^{(2)}| \geq k'$ for every $A_j^{(2)} = A_j^{(1)} \setminus V(T_2)$ with $1 \leq j \leq e_2$, then $\gamma > e_2$. Repeating this procedure, we shall find a set $S_q = S_{q-1} \cup V(T_q)$ such that $\gamma \leq e_q$ and moreover, $|A_{\alpha'} \cap V(T_{\beta'})| = 1$ for some $1 \leq \alpha' \leq e_r$ and $q \leq \beta' \leq r$. Thus, the set $A_j^{(q+1)} = A_j^{(q)} \setminus V(T_{q+1})$ satisfies $|A_j^{(q)}| = k' - 1$ and consequently $S_{q+1} = S_q \cup V(T_{q+1})$ is a $(k' - 1)$-antiresolving set (e.g., in Figure 13(II) the process must be done two times, first we remove the nodes in the set $V(T_1) \setminus \{v\}$ and next we remove the nodes in the set $V(T_2) \setminus \{v\}$, thereby getting the required $(k' - 1)$-antiresolving set).

Thus, in both cases we obtain a $(k' - 1)$-antiresolving set. By using the same procedure and a $(k' - 1)$-antiresolving set of minimum cardinality, we can find a $(k' - 2)$-antiresolving set and in general a k-antiresolving set for every $2 \leq k \leq k' - 1$, which completes the proof.

5.3.2 Empirical Results

We remind the readers about the assertions in ($\star$), ($\bowtie$) and ($\spadesuit$) while we report our empirical results and related conclusions.

5.3.2.1 Algorithms for Problems 1–3 (Algorithms I and II)

We obtain an exact solution for Problem 2 by implementing the following algorithm (Algorithm I) devised in (82) by Chatterjee *et al.*. In this algorithm, an absence of a valid solution is indicated by $\mathcal{L}_{\text{opt}}^{\geq k} \leftarrow \infty$ and $V_{\text{opt}}^{\geq k} \leftarrow \emptyset$.

(* Algorithm I *)

1. Compute $\mathbf{d}_{v_i}$ for all $i = 1, \ldots, n$ using any algorithm that solves *all-pairs-shortest-path* problem (87).

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

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

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

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

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

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

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

3.2.4 **else** let $V_1, V_2, \ldots, V_\ell$ be the *only* $\ell > 0$ equivalence classes

in $\Pi^{=}_{V\setminus V', -V'}$ such that $|V_1| = \cdots = |V_\ell| = \mu\left(\mathcal{D}_{V\setminus V', -V'}\right)$

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

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

We obtain exact solutions for Problem 1 and find k_{opt} by using Algorithm I in the following straightforward manner:

1. $k \leftarrow n - 1$; done $\leftarrow$ FALSE

2. **while** ($(k \geq 1)$ AND (NOT done)) **do**

2.1 **execute** Algorithm I

2.2 **if** $(V_{\text{opt}}^{\geq k} \neq \emptyset)$ **then** $k_{\text{opt}} \leftarrow k$; done $\leftarrow$ TRUE

Although $\textsc{Adim}_{=k}$ is NP-hard for almost all k, for $k = 1$ we implement the following logarithmic-approximation algorithm devised in (82) by Chatterjee *et al.* for $\textsc{Adim}_{=1}$ computing $\mathcal{L}_{\text{opt}}^{=1}$ and $V_{\text{opt}}^{=1}$.

(* Algorithm II *)

1. Compute $\mathbf{d}_{v_i}$ for all $i = 1, \ldots, n$ using any algorithm that solves *all-pairs-shortest-path* problem (87).

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

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

3.1 create the following instance of the set-cover problem (88)

containing $n-1$ elements and $n-1$ sets:

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

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

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

3.2.1 run the algorithm of Johnson in (88) for this instance of set-cover

giving a solution $\mathcal{I} \subseteq \{1, 2, \ldots, n\} \setminus \{i\}$

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

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

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

Remarks on the implementations of algorithms

Both Algorithm I and Algorithm II use the all-pairs-shortest-path computation. Just like the measures in this chapter, the all-pairs-shortest-path computation is *unavoidable* for a large variety of other geodesic-based network properties that are often used for real networks such as the betweenness centrality, closeness centrality or Gromov-hyperbolic curvature measure (89; 90; 8; 91). In practice, for larger networks the running time of the Floyd-Warshall algorithm of all-pairs-shortest-path (87) can often be improved by using algorithmic engineering tricks such as early termination criteria that are known in the algorithms community. Moreover, for specific networks under consideration, practitioners also consider using other algorithmic approaches, such as repeated use of Dijkstra's single-source shortest path or Johnson's algorithm (87), if they are run faster. For our networks, we found the Floyd-Warshall algorithm with appropriate data structures and

algorithmic engineering techniques to be sufficient. For increasing the efficiency and speed of the algorithms we used various data structures such as *STL nested maps* and *vectors* to improve comparisons and lookup operations. Furthermore, for Algorithm I, we prematurely terminate the algorithm if $|V_{\text{opt}}|$ reaches 1 as 1 is the smallest value of the size of attacker nodes.

5.3.3 Synthetic networks: models and algorithmic generations

We use two major types of synthetic networks, namely the *Erdös-Rényi random networks* and the *scale-free random networks* generated by the Barábasi-Albert *preferential-attachment* model (2). Although the Erdös-Rényi network model has been used by prior network researchers as a real-network model in several application domains (e.g., see (92; 93; 94; 95)) it is also known that this particular model is probably not very good a model for real networks in many other application domains. Thus, we also consider networks generated by the scale-free random network model which is more widely considered to be a real-network model in many network applications (e.g, see (2; 96; 97; 98; 7)).

Erdös-Rényi model This is the classical undirected Erdös-Rényi model $G(n, p)$, where n is the number of nodes and every possible edge in the network is selected independently with a probability of p. The average degree of any node in $G(n, p)$ is $(n-1)p \approx np$, leading to $\frac{n(n-1)p}{2} \approx \frac{n^2p}{2}$ as the average number of edges in the network. Our privacy measures assume that the given graph is connected since one connected component has no influence on the privacy of another connected component. Thus, it is imperative to select only those combinations of n and p that keeps the graph connected by keeping the average degree of every node to be at least 1. However, we actually need to make sure that the average degree is *at least* 2 since, for example, $\mathcal{L}_{\text{opt}}^{=1}$ is trivially equal to 1 otherwise. This implies that at the very least we must ensure that $(n-1)p \geq 2$, or *roughly* $np \geq 2$. However, in practice, while generating the actual random networks one may need to select a p that is slightly higher (in our case, $np \geq 2.5$). Note that the giant-component formation in ER networks happens around $np \approx 1$, so we are indeed further away from this phenomenon where slight variations in p cause abrupt changes in topological behavior of the network. We used the following four combinations of n

and p to generate our synthetic networks to capture the smaller average degree of 2.5 and the larger average degree of 5:

$n = 500$	$n = 500$	$n = 1000$	$n = 1000$
$p = 0.005$	$p = 0.01$	$p = 0.005$	$p = 0.01$
$np = 2.5$	$np = 5$	$np = 2.5$	$np = 5$

For $n = 500$ (respectively, for $n = 1000$) we generated 1000 random networks (respectively, 100 random networks) for each corresponding value of p, and then calculated relevant statistics using Algorithms I and II.

Scale-free model We use the Barábasi-Albert *preferential-attachment* model (2) to generate random scale-free networks. The algorithm for generating a random scale-free $G(n, q)$,where n is number of nodes and $q \ll n$ is the number of connections each new node makes, is as follows:

- Initialize G to have q nodes and *no* edges. Add these nodes to a "*list of repeated nodes*".
- Repeat the following steps till G has n nodes:
  - Randomly select q distinct nodes, say $u_1, \ldots, u_q$, from the *list of repeated nodes*.
  - Add a new node w and undirected edges $\{w, u_1\}, \ldots, \{w, u_q\}$ in G.
  - Add w and $u_1, \ldots, u_q$ to the current *list of repeated nodes*.

The larger the q is, the more dense is the network $G(n, q)$. We used the following four combinations of n and q to generate our synthetic scale-free networks:

$n = 500$	$n = 500$	$n = 1000$	$n = 1000$
$q = 5$	$q = 10$	$q = 5$	$q = 10$

For $n = 500$ (respectively, for $n = 1000$) we generated 1000 random networks (respectively, 100 random networks) for each corresponding value of q, and then calculated relevant statistics using Algorithms I and II.

TABLE XIV: List of real social networks studied in this study.

Name	# of nodes	# of edges	Description
(A) Zachary Karate Club (48)	34	78	Network of friendships between 34 members of a karate club at a US university in the 1970s
(B) San Juan Community (51)	75	144	Network for visiting relations between families living in farms in the neighborhood San Juan Sur, Costa Rica, 1948
(C) Jazz Musician Network (50)	198	2842	A social network of Jazz musicians
(D) University Rovira i Virgili emails (99)	1133	10903	the network of e-mail interchanges between members of the University Rovira i Virgili
(E) Enron Email Data set (100)	1088	1767	Enron email network
(F) Email Eu core (101)	986	24989	Emails from a large European research institution
(G) UC Irvine College Message platform (102)	1896	59835	Messages on a Facebook-like platform at UC-Irvine
(H) Hamsterster friendships (103)	1788	12476	This Network contains friendships between users of the website `hamsterster.com`

5.3.3.1 Real networks

Table XXI shows the list of eight well-known unweighted social networks that we investigated. All the networks except one were undirected; for the only directed *UC Irvine College Message platform* network, we ignored the direction of edges. For each network the *largest* connected component was selected and tested.

5.3.3.2 Results for real networks in Table XXI

Results for $\textsc{Adim}$ and $\textsc{Adim}_{\geq k}$ Table XXII shows the results for $\textsc{Adim}$ via applying Algorithm I to these networks. From these results we may conclude:

① For all networks *except* the "Enron Email Data" network, an attacker needs to control *only* *one* suitable node of the network to uniquely re-identify (based on the metric representa-

tion) a significant percentage of nodes in the network (ranging from 2.6% of nodes for the "University Rovira i Virgili emails" network to 26.5% of nodes for the "Zachary Karate Club" network).

② For all networks *except* the "Enron Email Data" network, the minimum privacy violation probability guarantee is significantly further from zero (ranging from 0.019 for the "UC Irvine College Message platform" network to 0.25 for the "Hamsterster friendships" network). The minimum privacy violation probability guarantee for the "Hamsterster friendships" network is significantly higher than all other networks.

③ The "Zachary Karate Club" and the "San Juan Community" networks are *more* vulnerable to privacy attacks in terms of the percentage of nodes in the networks whose privacy can be violated by the adversary.

TABLE XV: Results for ADIM using Algorithm I. n is the number of nodes and k_{opt} is the largest value of k such that $V_{\text{opt}}^{\geq k} \neq \emptyset$ (*cf.* Problem 1).

Name	n	k_{opt}	$p_{\text{opt}} = 1/k_{\text{opt}}$	$\mathcal{L}_{\text{opt}}^{\geq k_{\text{opt}}} = \mathcal{L}_{\text{opt}}^{= k_{\text{opt}}}$	$\frac{k_{\text{opt}}}{n}$
(A) Zachary Karate Club	34	9	0.111	1	26.5%
(B) San Juan Community	75	7	0.143	1	9.3%
(C) Jazz Musician Network	198	12	0.084	1	6.0%
(D) University Rovira i Virgili emails	1133	29	0.035	1	2.6%
(E) Enron Email Data set	1088	153	0.007	935	14.1%
(F) Email Eu core	986	39	0.026	1	3.4%
(G) UC Irvine College Message platform	1896	55	0.019	1	2.9%
(H) Hamsterster friendships	1788	4	0.25	1	0.22%

For the "Enron Email Data" network, $\mathcal{L}_{\text{opt}}^{\geq k_{\text{opt}}} = 935$ implies that even to achieve a modest value of $p_{\text{opt}} = 0.007$ an adversary needs to control a large percentage (at least $\frac{935\times100}{1088}\% \approx 86\%$) of its nodes, a possibility unlikely to happen in practice. Thus, we continue further investigation about this network to check if a value of k *somewhat* smaller than k_{opt} may allow a *sufficiently steep* decline in the number of nodes that the attacker need to control, and report the values of $\mathcal{L}_{\text{opt}}^{\geq k}$ corresponding to relevant values of $k > 1$ in Table XVI. As can be seen, the values of $\mathcal{L}_{\text{opt}}^{\geq k}$ does not decline unless k is really further away from k_{opt}, leading us to conclude the following:

④ For the "Enron Email Data" network, privacy violation of a large number of nodes of the network by an attacker cannot be guaranteed in a *practical* sense (*i.e.*, without gaining control of a large number of nodes).

TABLE XVI: Values of $\mathcal{L}_{\text{opt}}^{\geq k}$ corresponding to values for $k > 1$ for "Enron Email Data" network. Only those values of $k > 1$ for which $\mathcal{L}_{\text{opt}}^{\geq k} \neq \mathcal{L}_{\text{opt}}^{\geq k-1}$ are shown.

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

Results for $\textsc{Adim}_{=1}$ Algorithm II returns $\mathcal{L}_{\text{opt}}^{=1} = 1$ for all of our networks except the "Hamsterster friendships" network. For the "Hamsterster friendships" network, Algorithm II returns $\mathcal{L}_{\text{opt}}^{=1} = 2$. Thus, we conclude:

⑤ For all the real networks except the "Hamsterster friendships" network, an adversary controlling *just one* suitable node may uniquely re-identify (based on the metric represen-

tation) one other node in the network with certainty (*i.e.*, with a probability of 1). For the "Hamsterster friendships" network, the same conclusion holds provided the adversary controls two suitable nodes.

5.3.3.3 Results for Erdös-Rényi synthetic networks

Results for $\textsc{Adim}_{\geq k}$ Table XVII shows the results for $\textsc{Adim}_{\geq k}$ via applying Algorithm I to these networks. From these results we may conclude:

⑥ For *most* synthetic Erdös-Rényi networks, k_{opt} is a value that is *much smaller* compared to the number of nodes n. Thus, for our synthetic Erdös-Rényi networks, with high probability privacy violation of a large number of nodes of the network by an attacker *cannot* be achieved.

⑦ The values of $\frac{k_{\text{opt}}}{n}$ for denser Erdös-Rényi networks (corresponding to $p = 0.01$) is about 75% higher that those for sparser Erdös-Rényi networks (corresponding to $p = 0.005$) irrespective of the number of nodes. Thus, we conclude that our sparser synthetic Erdös-Rényi networks are more privacy-secure compared to their denser counter-parts.

Results for $\textsc{Adim}_{=1}$ Table XVIII shows the result of our experiments of computation of $\mathcal{L}_{\text{opt}}^{=1}$ using Algorithm II. From these results, we conclude:

⑧ For our synthetic Erdös-Rényi networks, with high probability an adversary controlling *at most two* nodes may uniquely re-identify (based on the metric representation) *at least* one other node in the network.

5.3.3.4 Results for scale-free synthetic networks

Results for $\textsc{Adim}_{\geq k}$ Table XIX shows the results for $\textsc{Adim}_{\geq k}$ via applying Algorithm I to these networks. From these results we may conclude:

TABLE XVII: Results for $\textsc{Adim}_{\geq k}$ using Algorithm I for classical Erdös-Rényi model $G(n, p)$. k_{opt} is the largest value of k such that $V_{\text{opt}}^{\geq k} \neq \emptyset$ (cf. Problem 1). The %-values indicate the percentage of the generated networks for those particular values of k_{opt} (e.g., for $n = 500$ and $p = 0.005$, 980 out of the 1000 networks have $k_{\text{opt}} \geq 5$).

Network parameters									
n	p								
500	0.005	k_{opt}	≥ 4	≥ 5	≥ 6	≥ 7	≥ 8	≥ 9	≥ 10	> 10
		$p_{\text{opt}} = 1/k_{\text{opt}}$	≤ 0.25	≤ 0.2	≤ 0.166	≤ 0.142	≤ 0.125	≤ 0.111	≤ 0.1	< 0.1
		% of networks	100%	98%	81.8%	54.6%	21.5%	8%	3%	1%
		At least 90% of networks have $k_{\text{opt}} \leq 8$ and $\frac{k_{\text{opt}}}{n} \leq 0.016$								
500	0.010	k_{opt}	≥ 9	≥ 10	≥ 11	≥ 12	≥ 13	≥ 14	≥ 15	> 15
		$p_{\text{opt}} = 1/k_{\text{opt}}$	≤ 0.11	≤ 0.1	≤ 0.09	≤ 0.083	≤ 0.077	≤ 0.071	≤ 0.066	< 0.066
		% of networks	100%	98%	94%	81.4%	49.4%	21.4%	6.8%	0.6%
		At least 90% of networks have $k_{\text{opt}} \leq 14$ and $\frac{k_{\text{opt}}}{n} \leq 0.028$								
1000	0.005	k_{opt}	≥ 10	≥ 11	≥ 12	≥ 13	≥ 14	> 14		
		$p_{\text{opt}} = 1/k_{\text{opt}}$	≤ 0.1	≤ 0.09	≤ 0.083	≤ 0.077	≤ 0.071	< 0.066		
		% of networks	100%	99%	65%	16%	7%	1%		
		At least 90% of networks have $k_{\text{opt}} \leq 13$ and $\frac{k_{\text{opt}}}{n} \leq 0.013$								
1000	0.010	k_{opt}	≥ 18	≥ 19	≥ 20	≥ 21	≥ 22	≥ 23	≥ 24	> 24
		$p_{\text{opt}} = 1/k_{\text{opt}}$	≤ 0.055	≤ 0.052	≤ 0.05	≤ 0.047	≤ 0.045	≤ 0.043	≤ 0.041	< 0.041
		% of networks	100%	99%	90%	75%	47%	26%	9%	1%
		At least 90% of networks have $k_{\text{opt}} \leq 23$ and $\frac{k_{\text{opt}}}{n} \leq 0.023$								

⑨ The value of k_{opt} relative to the size n of the network is much larger for synthetic scale-free networks compared to those for the synthetic Erdös-Rényi networks. Thus, compared to synthetic Erdös-Rényi networks, synthetic scale-free networks may allow privacy violation of a largeer number of nodes of the network by an attacker.

TABLE XVIII: Results for $\textsc{Adim}_{=1}$ using Algorithm II for classical Erdös-Rényi model $G(n, p)$. The %-values indicate the percentage of the generated networks that have the corresponding value of $\mathcal{L}_{\text{opt}}^{=1}$ (e.g., for $n = 500$ and $p = 0.01$, 920 out of the 1000 networks have $\mathcal{L}_{\text{opt}}^{=1} = 1$).

Network parameters		$\mathcal{L}_{\text{opt}}^{=1}$		
n	p	1	2	> 2
500	0.01	92%	7%	1%
500	0.005	5.9%	89.3%	4.8%
1000	0.01	8%	90%	2%
1000	0.005	5%	93%	1%

⑩ Unlike the synthetic Erdös-Rényi networks, the values of $\frac{k_{\text{opt}}}{n}$ for denser scale-free networks (corresponding to $q = 10$) may be smaller or larger than those for sparser scale-free networks (corresponding to $q = 5$). Thus, density of scale-free networks does not seem to be well-correlated to privacy-security of these networks.

Results for $\textsc{Adim}_{=1}$ Table XX shows the result of our experiments of computation of $\mathcal{L}_{\text{opt}}^{=1}$ using Algorithm II. From these results, we conclude:

⑪ Similar to synthetic synthetic Erdös-Rényi networks, for synthetic scale-free networks also with high probability an adversary controlling *at most two* nodes may uniquely re-identify (based on the metric representation) *at least* one other node in the network.

TABLE XIX: Results for $\textsc{Adim}_{\geq k}$ using Algorithm I for the Barábasi-Albert preferential-attachment scale-free model $G(n, q)$. k_{opt} is the largest value of k such that $V_{\text{opt}}^{\geq k} \neq \emptyset$ (cf. Problem 1). The %-values indicate the percentage of the generated networks for those particular values of k_{opt} (e.g., for $n = 500$ and $q = 5$, 990 out of the 1000 networks have $k_{\text{opt}} \geq 50$).

Network parameters									
n	q								
500	5	k_{opt}	≥ 49	≥ 50	≥ 55	≥ 60	≥ 65	≥ 70	> 70
		$p_{\text{opt}} = 1/k_{\text{opt}}$	≤ 0.0204	≤ 0.02	≤ 0.018	≤ 0.016	≤ 0.015	≤ 0.014	< 0.014
		% of networks	100%	99%	97%	89%	42%	10%	6%
		At least 90% of networks have $k_{\text{opt}} \leq 65$ and $\frac{k_{\text{opt}}}{n} \leq 0.13$							
500	10	k_{opt}	≥ 45	≥ 60	≥ 80	≥ 100	≥ 120	≥ 140	> 140
		$p_{\text{opt}} = 1/k_{\text{opt}}$	≤ 0.022	≤ 0.016	≤ 0.0125	≤ 0.001	≤ 0.008	≤ 0.007	< 0.007
		% of networks	100%	50%	48%	47%	27%	5%	4%
		At least 95% of networks have $k_{\text{opt}} \leq 120$ and $\frac{k_{\text{opt}}}{n} \leq 0.24$							
1000	5	k_{opt}	≥ 88	≥ 90	≥ 100	≥ 110	≥ 120	≥ 130	≥ 135
		$p_{\text{opt}} = 1/k_{\text{opt}}$	≤ 0.011	≤ 0.010	≤ 0.001	≤ 0.009	≤ 0.008	≤ 0.007	≤ 0.0074
		% of networks	100%	98%	94%	66%	32%	11%	1%
		At least 89% of networks have $k_{\text{opt}} \leq 120$ and $\frac{k_{\text{opt}}}{n} \leq 0.12$							
1000	10	k_{opt}	≥ 86	≥ 88	≥ 90	≥ 92	≥ 94	≥ 96	≥ 98	> 100
		$p_{\text{opt}} = 1/k_{\text{opt}}$	≤ 0.0116	≤ 0.0113	≤ 0.0111	≤ 0.0108	≤ 0.0106	≤ 0.0104	≤ 0.0102	< 0.001
		% of networks	100%	77%	67%	56%	43%	30%	13%	3%
		At least 87% of networks have $k_{\text{opt}} \leq 96$ and $\frac{k_{\text{opt}}}{n} \leq 0.096$								

TABLE XX: Results for $\textsc{Adim}_{=1}$ using Algorithm II for the Barábasi-Albert preferential-attachment scale-free model $G(n, q)$. The %-values indicate the percentage of the generated networks that have the corresponding value of $\mathcal{L}_{\text{opt}}^{=1}$ (e.g., for $n = 500$ and $q = 5$, 990 out of the 1000 networks have $\mathcal{L}_{\text{opt}}^{=1} = 2$).

Network parameters		$\mathcal{L}_{\text{opt}}^{=1}$	
n	q	2	> 2
500	5	99%	1%
500	10	99.5%	0.5%
1000	5	99%	1%
1000	10	99%	1%

CHAPTER 6

CONCLUSIONS

In this thesis, we examined various geodesic-based metrics in complex networks and investigated the algorithmic perspective related to these measures as well as their implications on network behavior.
In first part, Our empirical results showed that many biological and social networks are hyperbolic. We showed that hyperbolicity in real-world networks has some interesting implications on the shortest and approximately shortest paths. Our theoretical results led to methodologies for determining relevant paths between a source and a target in a signal transduction network, and identifying the most important nodes that mediate these paths. We also described the interesting impact of hyperbolicity on existence of cross-talks in regulatory networks. This raise a question about the interplay between hyperbolicity and cross-talk in these networks, that can be studied in order to obtain a better understanding of this phenomenon:

- *Are these cross-talks the reason that biological networks are hyperbolic? or does the hyperbolicity results in cross-talks? For answering this question one can consider a time-varying nature of the networks and study the changes in hyperbolicity over time.*

In second part, we discussed another geodesic-based property known as strong metric dimension and showed that the problem of computing the strong metric dimension of a graph can be reduced to a simpler problem known as Minimum Vertex Cover in a transformed graph. Our results led to both a 2-approximation algorithm and a (2-ϵ)-inapproximability for the problem of computing the strong metric dimension of a graph.
There are still several interesting computational complexity questions still remain open:

- Does the (2 ϵ)-inapproximability result for computing $sdim(G)$ hold even when G is bipartite and $diam(G) \leq 3$?

- Are there interesting non-trivial classes of graphs for which $sdim(G)$ can be computed in polynomial time?

In the last chapter, we examined (k, ℓ)-*anonymity*, a privacy measure for complex networks which has a similar nature to strong metric dimension. We investigated, both theoretically and empirically, quantifications of privacy violation measures of large networks under active attacks. Our empirical results shed light on privacy violation properties of eight real social networks as well as synthetic networks generated by the classical Erdös Rènyi model.

APPENDIX

PROOF OF THEOREMS IN CHAPTER 3

A.1 Theorem 3

Theorem 3 *Suppose that G has a cycle of $k \geq 4$ nodes which has no path-chord. Then, $\delta^{+}_{\text{worst}}(G) \geq \lceil k/4 \rceil$.*

Proof. In our proofs we will use the consequences of the 4-node condition when the 4 nodes are chosen in a specific manner as stated below in Lemma 4.

Lemma 4 *Let u_0, u_1, u_2, u_3 be four nodes such that u_3 is on a shortest path between u_1 and u_2. Suppose also that all the inter-node distances are* strictly *positive except for d_{u_1,u_3} and $d_{u_1,u_3} = \left\lceil \frac{d_{u_1,u_2}+d_{u_0,u_1}-d_{u_0,u_2}}{2} \right\rceil$. Then,*

$$\left\lceil \frac{d_{u_0,u_1} + d_{u_0,u_2} + d_{u_1,u_2}}{2} \right\rceil \quad \leq \quad d_{u_0,u_3} \quad + \quad d_{u_1,u_2} \quad \leq \quad \left\lceil \frac{d_{u_0,u_1} + d_{u_0,u_2} + d_{u_1,u_2}}{2} \right\rceil \quad + \quad 2\,\delta^{+}_{u_0,u_1,u_2,u_3}$$

Proof. Note that due to triangle inequality $0 \leq \left\lceil \frac{d_{u_1,u_2}+d_{u_0,u_1}-d_{u_0,u_2}}{2} \right\rceil \leq d_{u_1,u_2}$ and thus node u_3 always exists.

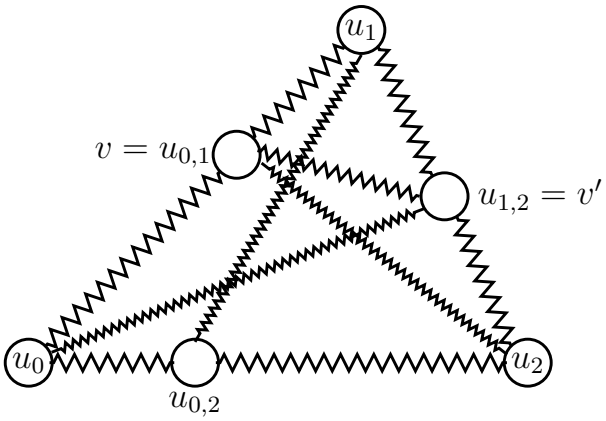


Figure 14: Case 1 of Theorem 5: $v = u_{0,1}$, $v' = u_{1,2}$.

APPENDIX (Continued)

First, consider the case when $0 < d_{u_1,u_3} < d_{u_1,u_2}$. Consider the three quantities involved in the 4-node condition for the nodes u_0, u_1, u_2, u_3, namely the quantities $d_{u_0,u_3} + d_{u_1,u_2}$, $d_{u_0,u_2} + d_{u_1,u_3}$ and $d_{u_0,u_1} + d_{u_2,u_3}$. Note that

$$2\,(d_{u_0,u_3} + d_{u_1,u_2}) = (d_{u_0,u_3} + d_{u_1,u_3}) + (d_{u_0,u_3} + d_{u_2,u_3}) + d_{u_1,u_2} \geq d_{u_0,u_1} + d_{u_0,u_2} + d_{u_1,u_2}$$

$$\Rightarrow\ d_{u_0,u_3} + d_{u_1,u_2} \geq \left\lceil \frac{d_{u_0,u_1} + d_{u_0,u_2} + d_{u_1,u_2}}{2} \right\rceil$$

$$d_{u_0,u_2} \;+\; d_{u_1,u_3} \quad = \quad d_{u_0,u_2} \;+\; \left\lfloor \frac{d_{u_1,u_2} + d_{u_0,u_1} - d_{u_0,u_2}}{2} \right\rfloor \quad = \quad \left\lfloor \frac{d_{u_0,u_1} + d_{u_0,u_2} + d_{u_1,u_2}}{2} \right\rfloor$$

$$d_{u_0,u_1} \;+\; d_{u_2,u_3} \quad = \quad d_{u_0,u_1} \;+\; \left\lceil \frac{d_{u_1,u_2} + d_{u_0,u_2} - d_{u_0,u_1}}{2} \right\rceil \quad = \quad \left\lceil \frac{d_{u_0,u_1} + d_{u_0,u_2} + d_{u_1,u_2}}{2} \right\rceil$$

Thus, $d_{u_0,u_3} + d_{u_1,u_2} \geq \max\{\, d_{u_0,u_2} + d_{u_1,u_3},\ d_{u_0,u_1} + d_{u_2,u_3} \,\}$ and using the definition of $\delta^+_{u_0,u_1,u_2,u_3}$ we have

$$\left\lceil \frac{d_{u_0,u_1} + d_{u_0,u_2} + d_{u_1,u_2}}{2} \right\rceil \quad \leq \quad d_{u_0,u_3} \;+\; d_{u_1,u_2} \quad \leq \quad \left\lceil \frac{d_{u_0,u_1} + d_{u_0,u_2} + d_{u_1,u_2}}{2} \right\rceil \;+\; 2\,\delta^+_{u_0,u_1,u_2,u_3}$$

Next, consider the case when $d_{u_1,u_3} = 0$. This implies

$$d_{u_0,u_1} \;+\; d_{u_1,u_3} \;=\; d_{u_0,u_1} \;+\; d_{u_1,u_2} \;=\; d_{u_0,u_2} \;=\; \frac{d_{u_0,u_1} + d_{u_0,u_2} + d_{u_1,u_2}}{2} \;\leq\; \left\lceil \frac{d_{u_0,u_1} + d_{u_0,u_2} + d_{u_1,u_2}}{2} \right\rceil$$

Finally, consider the case when $d_{u_1,u_3} = d_{u_1,u_2}$. This implies

$$d_{u_1,u_2} \;-\; \frac{d_{u_1,u_2} + d_{u_0,u_1} - d_{u_0,u_2}}{2} \;<\; 1 \quad \equiv \quad d_{u_0,u_2} \;+\; d_{u_1,u_2} \;=\; d_{u_0,u_1} \;+\; 2 \;-\; 2\,\varepsilon \quad \text{for some } 0 < \varepsilon \leq 1$$

APPENDIX (Continued)

Thus, it easily follows that

$$d_{u_0,u_3} + d_{u_1,u_2} = d_{u_0,u_2} + d_{u_1,u_2} = \frac{d_{u_0,u_2} + d_{u_1,u_2} + d_{u_0,u_1} + 2 - 2\varepsilon}{2} = \frac{d_{u_0,u_2} + d_{u_1,u_2} + d_{u_0,u_1}}{2} + 1 - \varepsilon$$

$$\Rightarrow\ d_{u_0,u_3} + d_{u_1,u_2} \leq \left\lceil \frac{d_{u_0,u_1} + d_{u_0,u_2} + d_{u_1,u_2}}{2} \right\rceil$$

❑

We can now prove Theorem 3 as follows. Let $C = (u_0, u_1, \ldots, u_{k-1}, u_0)$ be the cycle of $k = 4r + r'$ nodes for some integers r and $0 \leq r' < 4$. Consider the four nodes $u_0, u_{r+\lceil r'/2 \rceil}, u_{2r+\lfloor (r'+\lceil r'/2 \rceil)/2 \rfloor}$ and $u_{3r+r'}$. Since C has no path-chord, we have $d_{u_0,u_{r+\lceil r'/2 \rceil}} = r + \lceil r'/2 \rceil$, $d_{u_0,u_{2r+\lfloor (r'+\lceil r'/2 \rceil)/2 \rfloor}} = 2r + \left\lfloor \frac{r'+\lceil r'/2 \rceil}{2} \right\rfloor$ $d_{u_{r+\lceil r'/2 \rceil},u_{3r+r'}} = 2r + r' - \lceil r'/2 \rceil \leq 2r + \lceil r'/2 \rceil$, $d_{u_0,u_{3r+r'}} = r$ and $u_{2r+\lfloor (r'+\lceil r'/2 \rceil)/2 \rfloor}$ is on a shortest path between u_r and $u_{3r+r'}$. Thus, applying the bound of Lemma 4, we get

$$\delta^+_{\text{worst}}(G) \geq \delta^+_{u_0,\, u_{r+\lceil r'/2 \rceil}, u_{2r+\lfloor (r'+\lceil r'/2 \rceil)/2 \rfloor}, u_{3r+r'}} \geq \frac{d_{u_0,u_{2r+\left\lfloor \frac{r'+\lceil r'/2 \rceil}{2} \right\rfloor}} + d_{u_{r+\left\lceil \frac{r'}{2} \right\rceil},u_{3r+r}} - \left\lceil \frac{d_{u_0,u_{r+\left\lceil \frac{r'}{2} \right\rceil}} + d_{u_{r+\left\lceil \frac{r'}{2} \right\rceil},u_{3r+r'}} + d_{u_{3r+r'},u_0}}{2} \right\rceil}{2}$$

$$= \frac{4r + \left\lfloor \frac{r'+\lceil r'/2 \rceil}{2} \right\rfloor - r' + \lceil r'/2 \rceil - \left\lceil \frac{4r + r'}{2} \right\rceil}{2} = r + \frac{\left\lfloor \frac{r'+\lceil r'/2 \rceil}{2} \right\rfloor - r'}{2} \geq r - 1/4 \ \Rightarrow\ \delta^+_{\text{worst}}(G) \geq r = \lceil k/4 \rceil$$

❑

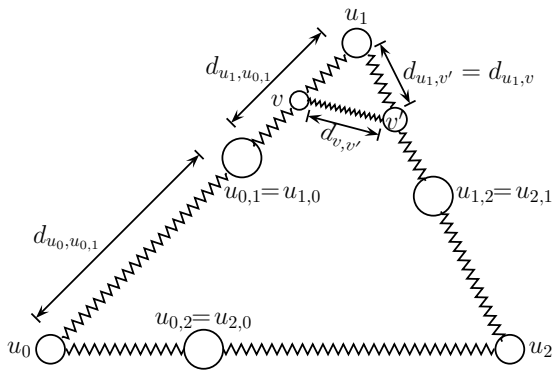
A.2 Theorem 5 and Corollary 6

The Gromov product nodes $u_{0,1}, u_{0,2}, u_{1,2}$ of a shortest-path triangle $\Delta_{\{u_0,u_1,u_2\}}$ are three nodes satisfying the following[1]:

---

[1]To simplify exposition, we assume that $d_{u_0,u_1} + d_{u_1,u_2} + d_{u_0,u_2}$ is an even number. Otherwise, the definition will require minor changes.

APPENDIX (Continued)



$$d_{u_0,u_{0,1}} = \left\lfloor \frac{d_{u_0,u_1} + d_{u_0,u_2} - d_{u_1,u_2}}{2} \right\rfloor \quad d_{u_1,u_{0,1}} = \left\lceil \frac{d_{u_1,u_2} + d_{u_1,u_0} - d_{u_2,u_0}}{2} \right\rceil$$

$$d_{u_1,u_{0,1}} = d_{u_1,u_{1,2}} \quad d_{u_0,u_{0,1}} = d_{u_0,u_{0,2}} \quad d_{u_2,u_{0,2}} = d_{u_2,u_{1,2}}$$

$$\delta^+_{\Delta_{\{u_0,u_1,u_2\}}} \leq \delta^+_{\text{worst}}(G) \quad d_{v,v'} \leq \max\left\{6\,\delta^+_{\Delta_{\{u_0,u_1,u_2\}}},\, 2\right\}$$

Figure 15: A pictorial illustration of the claim in Theorem 5.

- $u_{0,1}$, $u_{0,2}$ and $u_{1,2}$ are located on the paths $\mathcal{P}_\Delta(u_0, u_1)$, $\mathcal{P}_\Delta(u_0, u_2)$ and $\mathcal{P}_\Delta(u_1, u_2)$, respectively, and
- the distances of these three nodes from u_0, u_1 and u_2 satisfy the following constraints:

$$d_{u_0,u_{0,1}} + d_{u_1,u_{0,1}} = d_{u_0,u_1}, \quad d_{u_0,u_{0,2}} + d_{u_2,u_{0,2}} = d_{u_0,u_2} d_{u_1,u_{1,2}} + d_{u_2,u_{1,2}} = d_{u_1,u_2}, \quad d_{u_1,u_{0,1}} = d_{u_1,u_{1,2}}$$

$$d_{u_0,u_{0,1}} = d_{u_0,u_{0,2}} = \left\lfloor \frac{d_{u_0,u_1} + d_{u_0,u_2} - d_{u_1,u_2}}{2} \right\rfloor$$

It is not difficult to see that a set of such three nodes always exists. For convenience, the nodes $u_{1,0}$, $u_{2,0}$ and $u_{2,1}$ are assumed to be the same as the nodes $u_{0,1}$, $u_{0,2}$ and $u_{1,2}$, respectively.

Theorem 5 (see Fig. Figure 15 for a visual illustration) *For a shortest-path triangle $\Delta_{\{u_0,u_1,u_2\}}$ and for $0 \leq i \leq 2$, let v and v' be two nodes on the paths $u_i \overset{\mathcal{P}_\Delta(u_i,u_{i+2 \pmod 3})}{\longleftrightarrow\!\!\sim} u_{i,\,i+2 \pmod 3}$ and $u_i \overset{\mathcal{P}_\Delta(u_i,u_{i+1 \pmod 3})}{\longleftrightarrow\!\!\sim} u_{i,\,i+1 \pmod 3}$, respectively, such that $d_{u_i,v} = d_{u_i,v'}$. Then,*

$$d_{v,v'} \leq 6\,\delta^+_{\Delta_{\{u_0,u_1,u_2\}}} + 2$$

APPENDIX (Continued)

where $\delta^+_{\Delta_{\{u_0,u_1,u_2\}}} \leq \delta^+_{\text{worst}}(G)$ *is the largest worst-case hyperbolicity among all combinations of four nodes in the three shortest paths defining the triangle.*

Corollary 6 (Hausdorff distance between shortest paths) *Suppose that* $\mathcal{P}_1$ *and* $\mathcal{P}_2$ *are two shortest paths between two nodes* u_0 *and* u_1*. Then, the Hausdorff distance* $d_H(\mathcal{P}_1, \mathcal{P}_2)$ *between these two paths can be bounded as:*

$$d_H(\mathcal{P}_1,\mathcal{P}_2) \quad \stackrel{\text{def}}{=} \quad \max\left\{ \max_{v_1\in\mathcal{P}_1} \min_{v_2\in\mathcal{P}_2} \left\{ d_{v_1,v_2} \right\}, \max_{v_2\in\mathcal{P}_2} \min_{v_1\in\mathcal{P}_1} \left\{ d_{v_1,v_2} \right\} \right\} \quad \leq \quad 6\,\delta^+_{\Delta_{\{u_0,u_1,u_2\}}} + 2$$

where u_2 *is any node on the path* $\mathcal{P}_2$*.*

Proof of Theorem 5. To simplify exposition, we assume that $d_{u_0,u_1}+d_{u_1,u_2}+d_{u_0,u_2}$ is even and prove a slightly improve bound of $d_{v,v'} \leq 6\,\delta^+_{\Delta_{\{u_0,u_1,u_2\}}} + 1$. It is easy to modify the proof to show that $d_{v,v'} \leq 6\,\delta^+_{\Delta_{\{u_0,u_1,u_2\}}} + 2$ if $d_{u_0,u_1} + d_{u_1,u_2} + d_{u_0,u_2}$ is odd.

We will prove the result for $i = 1$ only; similar arguments will hold for $i = 0$ and $i = 2$. If $d_{u_1,u_{0,1}} = 0$ then $v = v' = u_1$ and the claim holds trivially, Thus, we assume that $d_{u_1,u_{0,1}} > 0$.

Case 1: $v = u_{0,1}$ **and** $v' = u_{1,2}$. In this case we need to prove that $d_{u_{0,1},u_{1,2}} \leq 6\,\delta^+_{\Delta_{\{u_0,u_1,u_2\}}} + 1$ (see Fig. Figure 14). Assume that $d_{u_{0,1},u_{1,2}} > 0$ since otherwise the claim is trivially true. Using Lemma 4 for the four nodes $u_0, u_1, u_2, u_{1,2}$, we get

$$d_{u_0,u_{1,2}} + d_{u_1,u_2} \leq \left\lceil \frac{d_{u_0,u_1} + d_{u_1,u_2} + d_{u_0,u_2}}{2} \right\rceil + 2\,\delta^+_{u_0,u_1,u_2,u_{1,2}} \tag{A.1}$$

Now, we note that

$$d_{u_1,u_2} \quad + \quad d_{u_0,u_{0,2}} \quad = \quad d_{u_1,u_2} \quad + \quad \left\lfloor \frac{d_{u_0,u_1} + d_{u_0,u_2} - d_{u_1,u_2}}{2} \right\rfloor \quad = \quad \left\lfloor \frac{d_{u_0,u_1} + d_{u_0,u_2} + d_{u_1,u_2}}{2} \right\rfloor \tag{A.2}$$

APPENDIX (Continued)

which in turn implies

$$\left| d_{u_0,u_{1,2}} - d_{u_0,u_{0,2}} \right| = \left| \left(d_{u_0,u_{1,2}} + d_{u_1,u_2} \right) - \left(d_{u_1,u_2} + d_{u_0,u_{0,2}} \right) \right| \leq \left| \underbrace{\left\lceil \frac{d_{u_0,u_1} + d_{u_1,u_2} + d_{u_0,u_2}}{2} \right\rceil}_{\text{(by inequality (Equation A.1))}} + 2\,\delta^{+}_{u_0,u_1,u_2,u_{1,2}} \right.$$

$$\left. - \underbrace{\left\lfloor \frac{d_{u_0,u_1} + d_{u_0,u_2} + d_{u_1,u_2}}{2} \right\rfloor}_{\text{(by equality (Equation A.2))}} \right| \leq 2\,\delta^{+}_{u_0,u_1,u_2,u_{1,2}} + 1 \quad \text{(A.3)}$$

In a similar manner, we can prove the following analog of inequality (Equation A.3):

$$\left| d_{u_2,u_{0,1}} - d_{u_2,u_{0,2}} \right| \leq 2\,\delta^{+}_{u_0,u_1,u_2,u_{0,1}} \quad \text{(A.4)}$$

Using inequalities (Equation A.3) and (Equation A.4), it follows that

$$\left| \left(d_{u_0,u_{1,2}} + d_{u_2,u_{0,1}} \right) - d_{u_0,u_2} \right| = \left| \left(d_{u_0,u_{1,2}} + d_{u_2,u_{0,1}} \right) - \left(d_{u_0,u_{0,2}} + d_{u_2,u_{0,2}} \right) \right| = \left| \left(d_{u_0,u_{1,2}} - d_{u_0,u_{0,2}} \right) + \left(d_{u_2,u_{0,1}} - d_{u_2,u_{0,2}} \right) \right|$$

$$\leq \left| d_{u_0,u_{1,2}} - d_{u_0,u_{0,2}} \right| + \left| d_{u_2,u_{0,1}} - d_{u_2,u_{0,2}} \right| \leq 2\,\delta^{+}_{u_0,u_1,u_2,u_{1,2}} + 2\,\delta^{+}_{u_0,u_1,u_2,u_{0,1}} + 1 \quad \text{(A.5)}$$

Now, consider the three quantities involved in the 4-node condition for the nodes $u_0, u_2, u_{0,1}, u_{1,2}$, namely the quantities: $d_{u_0,u_2} + d_{u_{0,1},u_{1,2}}$, $d_{u_0,u_{1,2}} + d_{u_{0,1},u_2}$ and $d_{u_0,u_{0,1}} + d_{u_2,u_{1,2}}$. Note that

$$d_{u_0,u_{0,1}} + d_{u_2,u_{1,2}} = d_{u_0,u_{0,2}} + d_{u_2,u_{0,2}} = d_{u_0,u_2} < d_{u_0,u_2} + d_{u_{0,1},u_{1,2}} \quad \text{(A.6)}$$

If $d_{u_0,u_{1,2}} + d_{u_{0,1},u_2} \leq d_{u_0,u_{0,1}} + d_{u_2,u_{1,2}}$ then by the definition of $\delta^{+}_{u_0,u_2,u_{0,1},u_{1,2}}$ we have

$$d_{u_{0,1},u_{1,2}} = \left(d_{u_0,u_2} + d_{u_{0,1},u_{1,2}} \right) - d_{u_0,u_2} = \left(d_{u_0,u_2} + d_{u_{0,1},u_{1,2}} \right) - \left(d_{u_0,u_{0,1}} + d_{u_2,u_{1,2}} \right) \leq 2\,\delta^{+}_{u_0,u_2,u_{0,1},u_{1,2}}$$

APPENDIX (Continued)

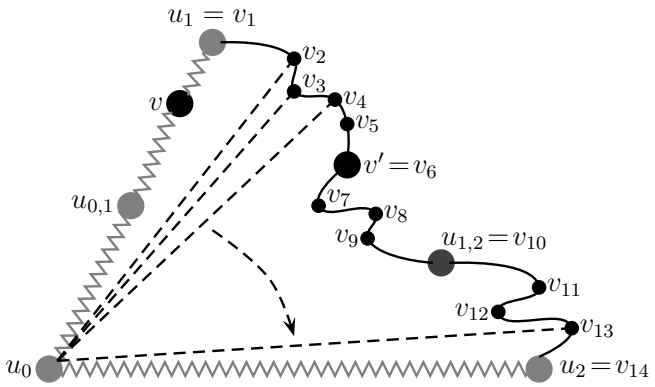


Figure 16: Case 2 of Theorem 5: $v \neq u_{0,1}$, $v' \neq u_{1,2}$.

Otherwise, $d_{u_0,u_{1,2}} + d_{u_{0,1},u_2} > d_{u_0,u_{0,1}} + d_{u_2,u_{1,2}}$ and then again by the definition of $2\,\delta^+_{u_0,u_2,u_{0,1},u_{1,2}}$ we have

$$\left| d_{u_0,u_{1,2}} + d_{u_{0,1},u_2} - d_{u_0,u_2} - d_{u_{0,1},u_{1,2}} \right| \leq 2\,\delta^+_{u_0,u_2,u_{0,1},u_{1,2}}$$

and now using inequality (Equation A.5) gives

$$\begin{aligned} d_{u_{0,1},u_{1,2}} &= \left(d_{u_0,u_{1,2}} + d_{u_2,u_{0,1}} - d_{u_0,u_2}\right) - \left(d_{u_0,u_{1,2}} + d_{u_{0,1},u_2} - d_{u_0,u_2} - d_{u_{0,1},u_{1,2}}\right) \\ &\leq \left|d_{u_0,u_{1,2}} + d_{u_2,u_{0,1}} - d_{u_0,u_2}\right| + \left|d_{u_0,u_{1,2}} + d_{u_{0,1},u_2} - d_{u_0,u_2} - d_{u_{0,1},u_{1,2}}\right| \\ &\leq 2\,\delta^+_{u_0,u_1,u_2,u_{1,2}} + 2\,\delta^+_{u_0,u_1,u_2,u_{0,1}} + 2\,\delta^+_{u_0,u_2,u_{0,1},u_{1,2}} + 1 \leq 6\,\delta^+_{\Delta_{\{u_0,u_1,u_2\}}} + 1 \end{aligned}$$

Case 2: $v \neq u_{0,1}$ **and** $v' \neq u_{1,2}$. The claim trivially holds if $d_{v,v'} \leq 1$, thus we assume that $d_{v,v'} > 1$. Let $(v_1 = u_1, v_2 = u_3, v_3, \ldots, v_h = v', \ldots, v_s = u_{1,2}, \ldots, v_r = u_2)$ be the *ordered* sequence of nodes in the given shortest path from u_1 to u_2 (see Fig. Figure 16). Consider the sequence of shortest-path triangles $\Delta_{\{u_0,u_1,v_2\}}, \Delta_{\{u_0,u_1,v_3\}}, \ldots, \Delta_{\{u_0,u_1,v_r\}}$, where each such triangle $\Delta_{\{u_0,u_1,v_j\}}$ is obtained by taking the shortest path $\mathcal{P}_\Delta(u_0,u_1)$, the sub-path $\mathcal{P}_\Delta(u_1,v_j)$ of the shortest path $\mathcal{P}_\Delta(u_1,u_2)$, from u_1 to v_j, and a shortest path $u_0 \overset{s}{\rightsquigarrow} v_j$

APPENDIX (Continued)

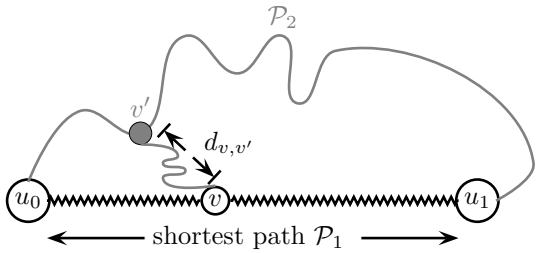


Figure 17: Illustration of the bound in Theorem 7.

from u_0 to v_j. Let $v_{1,j}$ be the Gromov product node on the side (shortest path) $\mathcal{P}_\Delta(u_1, v_j)$ for the shortest-path triangle $\Delta_{\{u_0,u_1,v_j\}}$.

We claim that if $v_{1,j} = v_p$ and $v_{1,j+1} = v_q$ then q is either p or $p+1$. Indeed, if $d_{u_1,v_p} = \left\lfloor \frac{d_{u_0,u_1}+d_{u_1,u_j}-d_{u_0,v_j}}{2} \right\rfloor$ and $d_{u_1,v_q} = \left\lfloor \frac{d_{u_0,u_1}+d_{u_1,u_{j+1}}-d_{u_0,v_{j+1}}}{2} \right\rfloor$ then

$$
\begin{aligned}
d_{u_1,v_q} - d_{u_1,v_p} &= \left\lfloor \frac{d_{u_0,u_1} + d_{u_1,v_{j+1}} - d_{u_0,v_{j+1}}}{2} \right\rfloor - \left\lfloor \frac{d_{u_0,u_1} + d_{u_1,v_j} - d_{u_0,v_j}}{2} \right\rfloor \\
&\leq \left\lfloor \frac{d_{u_0,u_1} + \left(1 + d_{u_1,v_j}\right) - \left(d_{u_0,v_{j+1}} - 1\right)}{2} \right\rfloor - \left\lfloor \frac{d_{u_0,u_1} + d_{u_1,v_j} - d_{u_0,v_j}}{2} \right\rfloor \\
&= \left\lfloor \frac{d_{u_0,u_1} + d_{u_1,v_j} - d_{u_0,v_j}}{2} + 1 \right\rfloor - \left\lfloor \frac{d_{u_0,u_1} + d_{u_1,v_j} - d_{u_0,v_j}}{2} \right\rfloor \leq 1
\end{aligned}
$$

and a similar proof of $d_{u_1,v_q} - d_{u_1,v_p} \leq 1$ can be obtained if $d_{u_1,v_p} = \left\lceil \frac{d_{u_0,u_1}+d_{u_1,u_j}-d_{u_0,v_j}}{2} \right\rceil$ and $d_{u_1,v_q} = \left\lceil \frac{d_{u_0,u_1}+d_{u_1,u_{j+1}}-d_{u_0,v_{j+1}}}{2} \right\rceil$. Thus, the ordered sequence of nodes $v_{1,1}, v_{1,2}, \ldots, v_{1,r}$ cover the ordered sequence of nodes $v_2, v_3, \ldots, v_s$ in a *consecutive manner* without skipping over any node. Since $v_{1,1}$ is either v_1 or v_2, and $v_{1,r} = v_s = u_{1,2}$, there must be an index t such that $v_{1,t} = v' = v_h$. Since $d_{u_1,v} = d_{u_1,v'}$, v and v' are the two Gromov product nodes for the shortest-path triangle $\Delta_{\{u_0,u_1,v_t\}}$ and thus applying Case **1.1** on $\Delta_{\{u_0,u_1,v_t\}}$ we have $d_{v,v'} \leq 6\,\delta^{+}_{\Delta_{\{u_0,u_1,u_2\}}} + 1$. ❑

APPENDIX (Continued)

A.2.1 Theorem 7 and Corollary 8

Theorem 7 (see Fig. Figure 17 for a visual illustration) *Let $\mathcal{P}_1 \equiv u_0 \overset{s}{\rightsquigarrow} u_1$ and $\mathcal{P}_2$ be a shortest path and an arbitrary path, respectively, between two nodes u_0 and u_1. Then, for every node v on $\mathcal{P}_1$, there exists a node v' on $\mathcal{P}_2$ such that*

$$\begin{aligned} d_{v,v'} &\leq \min\left\{ \Big(6\,\delta^{+}_{\text{worst}}(G)+2\Big)\Big(\lfloor \log_2 \ell\,(\mathcal{P}_2)\,\rfloor - 1\Big),\ \left\lfloor \tfrac{d_{u_0,u_1}}{2} \right\rfloor \right\} \\ &= \mathrm{O}\Big(\,\delta^{+}_{\text{worst}}(G)\,\log \ell\,(\mathcal{P}_2)\,\Big) \end{aligned}$$

Since $\ell\,(\mathcal{P}_2) \leq n$, the above bound also implies that

$$d_{v,v'} \leq \Big(6\,\delta^{+}_{\text{worst}}(G)+2\Big)\Big(\lfloor \log_2 n\,\rfloor - 1\Big) = \mathrm{O}\Big(\,\delta^{+}_{\text{worst}}(G)\,\log n\Big)$$

Corollary 8 *Suppose that there exists a node v on the shortest path between u_0 and u_1 such that $\min_{v' \in \mathcal{P}_2}\{d_{v,v'}\} \geq \gamma$. Then, $\ell\,(\mathcal{P}_2) \geq 2^{\frac{\gamma}{6\,\delta^{+}_{\text{worst}}(G)+2}+1} - 1 = \Omega\Big(2^{\gamma\,/\,\delta^{+}_{\text{worst}}(G)}\Big)$.*

Proof of Theorem 7. First, note that by selecting v' to be one of u_0 or u_1 appropriately we have $d_{v,v'} \leq \lfloor d_{u_0,u_1}/2 \rfloor$. Now, assume that $\ell\,(\mathcal{P}_2) > 2$. Let u_2 be the node on the path $\mathcal{P}_2$ such that $\ell(u_0 \overset{\mathcal{P}_2}{\rightsquigarrow} u_2) = \lceil\, \ell(\mathcal{P}_2)/2\,\rceil$. and consider the shortest-path triangle $\Delta_{\{u_0,u_1,u_2\}}$. By Theorem 5 there exists a node v' either on a shortest path between u_0 and u_2 or on a shortest path between u_1 and u_2 such that $d_{v,v'} \leq 6\,\delta^{+}_{\text{worst}}(G) + 2$. We move from v to v' and recursively solve the problem of finding a shortest path from v' to a node on a part of the path $\mathcal{P}_2$ containing at most $\lceil (\mathcal{P}_2)/2 \rceil$ edges. Let $D(y)$ denote the minimum distance from v to a node in a path of length y between u_0 and u_1. Thus, the worst-case recurrence for $D(y)$ is given by

$$D(y) \leq D\left(\left\lceil \tfrac{y}{2} \right\rceil\right) + 6\,\delta^{+}_{\text{worst}}(G) + 2, \quad \text{if } y > 2$$

$$D(2) = 1$$

APPENDIX (Continued)

A solution to the above recurrence satisfies $D(\ell(\mathcal{P}_2)) \leq \left(6\,\delta^+_{\text{worst}}(G)+2\right)\left(\lceil \log_2 \ell(\mathcal{P}_2)\rceil - 1\right)$. ❑

A.2.2 Theorem 9 and Corollary 10

For easy of display of long mathematical equations, we will denote $\delta^+_{\text{worst}}(G)$ simply as δ^+.

Theorem 9 *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be a shortest path and another path, respectively, between two nodes. Define $\eta_{\mathcal{P}_1,\mathcal{P}_2}$ as*

$$\begin{aligned}
&\eta_{\mathcal{P}_1,\mathcal{P}_2}\\
&= \left(6\,\delta^+ + 2\right)\log_2\left(\left(6\mu+2\right)\left(6\,\delta^+ + 2\right)\log_2\left[\,(6\,\delta^+ + 2)\;(3\mu+1)\,\mu\right] + \mu\right)\\
&= \mathrm{O}\left(\delta^+\,\log\left(\mu\,\delta^+\right)\right), \quad \textit{if } \mathcal{P}_2 \textit{ is } \mu\textit{-approximate short}
\end{aligned}$$

$$\begin{aligned}
&\eta_{\mathcal{P}_1,\mathcal{P}_2}\\
&= \left(6\,\delta^+ + 2\right)\log_2\left(8\left(6\,\delta^+ + 2\right)\log_2\left[\,(6\,\delta^+ + 2)\;(4+2\varepsilon)\right] + 1 + \frac{\varepsilon}{2}\right)\\
&= \mathrm{O}\left(\delta^+\log\left(\varepsilon + \delta^+\,\log\varepsilon\right)\right), \quad \textit{if } \mathcal{P}_2 \textit{ is } \varepsilon\textit{-additive-approximate short}
\end{aligned}$$

Then, the following statements are true.

(a) *For every node v on $\mathcal{P}_1$, there exists a node v' on $\mathcal{P}_2$ such that $d_{v,v'} \leq \lfloor \eta_{\mathcal{P}_1,\mathcal{P}_2} \rfloor$.*

(b) *For every node v' on $\mathcal{P}_2$, there exists a node v on $\mathcal{P}_1$ such that $d_{v,v'} \leq \zeta_{\mathcal{P}_1,\mathcal{P}_2}$ where*

$$\zeta_{\mathcal{P}_1,\mathcal{P}_2} = \begin{cases}
\min\left\{\left\lfloor (\mu+1)\,\eta_{\mathcal{P}_1,\mathcal{P}_2} + \frac{\mu}{2}\right\rfloor, \left\lfloor \frac{\mu\, d_{u_0,u_1}}{2}\right\rfloor\right\}\\
= \mathrm{O}\left(\mu\,\delta^+\,\log\left(\mu\,\delta^+\right)\right), \quad \textit{if } \mathcal{P}_2 \textit{ is } \mu\textit{-approximate short}\\
\min\left\{\left\lfloor 2\,\eta_{\mathcal{P}_1,\mathcal{P}_2} + \frac{1+\varepsilon}{2}\right\rfloor, \left\lfloor \frac{d_{u_0,u_1}+\varepsilon}{2}\right\rfloor\right\}\\
= \mathrm{O}\left(\varepsilon + \delta^+\log\left(\varepsilon + \delta^+\,\log\varepsilon\right)\right),\\
\qquad\qquad \textit{if } \mathcal{P}_2 \textit{ is } \varepsilon\textit{-additive-approximate short}
\end{cases}$$

APPENDIX (Continued)

Corollary 10 **(Hausdorff distance between approximate short paths)** *Suppose that $\mathcal{P}_1$ and $\mathcal{P}_2$ are two paths between two nodes. Then, the Hausdorff distance $d_H(\mathcal{P}_1, \mathcal{P}_2)$ between these two paths can be bounded as follows:*

$$d_H(\mathcal{P}_1, \mathcal{P}_2) \stackrel{\text{def}}{=} \max\left\{ \max_{v_1 \in \mathcal{P}_1} \min_{v_2 \in \mathcal{P}_2} \left\{ d_{v_1,v_2} \right\}, \max_{v_2 \in \mathcal{P}_2} \min_{v_1 \in \mathcal{P}_1} \left\{ d_{v_1,v_2} \right\} \right\} \leq \eta_{\mathcal{P}_1,\, u_0 \overset{s}{\leftrightsquigarrow} u_1} + \zeta_{\mathcal{P}_2,\, u_0 \overset{s}{\leftrightsquigarrow} u_1}$$

Corollary 11 *Suppose that there exists a node v on the shortest path between u_0 and u_1 such that $\min_{v' \in \mathcal{P}_2} \{d_{v,v'}\} \geq \gamma$. Then, the following is true.*

- *If $\mathcal{P}_2$ is a μ-approximate short path then*

$$\mu > \frac{2^{\frac{\gamma}{6\delta^+ + 1}}}{12\gamma - \left(24 + o(1)\right)\left(6\delta^+ + 1\right)} - \frac{1}{3} \Rightarrow \mu = \Omega\left(\frac{2^{\gamma/\delta^+}}{\gamma}\right)$$

- *If $\mathcal{P}_2$ is a ε-additive-approximate short path then*

$$\varepsilon > \frac{2^{\frac{\gamma}{6\delta^+ + 1}}}{\left(48\delta^+ + \frac{17}{2}\right)} - \log_2(48\delta^+ + 8) \qquad \Rightarrow \varepsilon = \Omega\left(\frac{2^{\gamma/\delta^+}}{\delta^+} - \log \delta^+\right)$$

In particular, assuming real world networks have small constant values of δ^+, the asymptotic dependence of μ and ε on γ can be summarized as:

$$\text{both } \mu \text{ and } \varepsilon \text{ are } \Omega(2^{c\gamma}) \text{ for some constant } 0 < c < 1$$

Proof of Theorem 9. *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be a shortest path and another path, respectively, between two nodes u_0 and u_1. Note that any "sub-path" of a μ-approximate short path is* also *a μ-approximately short path,* i.e.*, $u_i \overset{\mathcal{P}}{\leftrightsquigarrow} u_j$ is also a μ-approximate short path, and similarly any sub-path of a ε-additive-approximate short*

APPENDIX (Continued)

path is also *a ε-additive-approximate short path. μ-approximate shortest paths also restrict the "span" of a path-chord of the path,* i.e., if $(u_0, u_1, \ldots, u_k)$ is a μ-approximate short path and $\left\{u_i, u_j\right\} \in E$ then $|j - i| \leq \mu$.

(a) *Let v and v' be two nodes on $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively, such that $\alpha = d_{v,v'} = \max\limits_{v'' \in \mathcal{P}_1} \min\limits_{v''' \in \mathcal{P}_2} \{d_{v'',v'''}\}$. Let $v_\ell \in u_0 \overset{\mathcal{P}_1}{\leftrightsquigarrow} v$ and $v_r \in u_1 \overset{\mathcal{P}_1}{\leftrightsquigarrow} v$ be two nodes defined by*

$$d_{v_\ell,v} = \begin{cases} 2\alpha + 1, & \text{if } d_{u_0,v} > 2\alpha + 1 \\ d_{u_0,v}, & \text{otherwise} \end{cases}$$

$$d_{v_r,v} = \begin{cases} 2\alpha + 1, & \text{if } d_{u_1,v} > 2\alpha + 1 \\ d_{u_1,v}, & \text{otherwise} \end{cases}$$

By definition of α, there exists two nodes $\widetilde{v_\ell}$ and $\widetilde{v_r}$ on the path $\mathcal{P}_2$ such that $d_{v_\ell,\widetilde{v_\ell}}, d_{v_r,\widetilde{v_r}} \leq \alpha$. Consider the $\mathcal{P}_3 = \widetilde{v_\ell} \overset{\mathcal{P}_2}{\leftrightsquigarrow} \widetilde{v_r}$ that is the part of path $\mathcal{P}_2$ from $\widetilde{v_\ell}$ to $\widetilde{v_r}$. Note that

$$d_{\widetilde{v_\ell},\widetilde{v_r}} \leq d_{\widetilde{v_\ell},v_\ell} + d_{v_\ell,v_r} + d_{v_r,\widetilde{v_r}} \leq 6\alpha + 2$$

Thus, we arrive at the following inequalities

$$\ell(\mathcal{P}_3) \leq \begin{cases} (6\alpha + 2)\mu, & \text{if } \mathcal{P}_2 \text{ is } \mu\text{-approximate short} \\ 6\alpha + 2 + \varepsilon, & \text{if } \mathcal{P}_2 \text{ is } \varepsilon\text{-additive-approximate short} \end{cases}$$

Now consider the path $\mathcal{P}_4 = v_\ell \overset{\mathcal{S}}{\leftrightsquigarrow} \widetilde{v_\ell} \overset{\mathcal{P}_2}{\leftrightsquigarrow} \widetilde{v_r} \overset{\mathcal{S}}{\leftrightsquigarrow} v_r$ obtained by taking a shortest path from v_ℓ to $\widetilde{v_\ell}$ followed by the path $\mathcal{P}_3$ followed by a shortest path from v_r to $\widetilde{v_r}$. Note that

$$\ell(\mathcal{P}_4) \leq \begin{cases} (6\alpha + 2)\mu + 2\alpha, \text{ if } \mathcal{P}_2 \text{ is } \mu\text{-approximate short} \\ 6\alpha + 2 + \varepsilon + 2\alpha = 8\alpha + 2 + \varepsilon, \\ \qquad \text{if } \mathcal{P}_2 \text{ is } \varepsilon\text{-additive-approximate short} \end{cases}$$

APPENDIX (Continued)

We claim that $\min_{\widetilde{v}\in\mathcal{P}_4}\{d_{v,\widetilde{v}}\} = \alpha$. *Indeed, if* $\widetilde{v} \in \mathcal{P}_3$ *then, by definition of* α, $\min_{\widetilde{v}}\{d_{v,\widetilde{v}}\} = \alpha$. *Otherwise, if* $\widetilde{v} \in v_\ell \overset{S}{\leftrightsquigarrow} \widetilde{v_\ell}$, *then by triangle inequality* $d_{v_\ell,v} \le d_{v,\widetilde{v}} + d_{\widetilde{v},v_\ell} \Rightarrow d_{v,\widetilde{v}} \ge 2\alpha + 1 - d_{\widetilde{v},v_\ell} > \alpha$. *Similarly, if* $\widetilde{v} \in \widetilde{v_r} \overset{S}{\leftrightsquigarrow} v_r$, *then by triangle inequality* $d_{v_r,v} \le d_{v,\widetilde{v}} + d_{\widetilde{v},v_r} \Rightarrow d_{v,\widetilde{v}} \ge 2\alpha + 1 - d_{\widetilde{v},v_r} > \alpha$. *Since* $v_\ell \overset{\mathcal{P}_1}{\leftrightsquigarrow} v_r$ *is a shortest path between* v_ℓ *and* v_r *and* v *is a node on this path, by Theorem 7,* $\alpha \le (6\delta^+ + 2)\,(\lfloor \log_2 \ell(\mathcal{P}_4) \rfloor - 1)$. *Thus, we have the following inequalities:*

• *If* $\mathcal{P}_2$ *is a* μ*-approximate short path then*

$$\begin{aligned}
& \ell(\mathcal{P}_4) \\
\le\ & (6\alpha + 2)\mu + 2\alpha \\
=\ & (6\mu + 2)\alpha + 2\mu \\
\le\ & (6\mu + 2)\,(6\delta^+ + 2)\,(\log_2 \ell(\mathcal{P}_4) - 1) + 2\mu \\
\le\ & (6\mu + 2)\,(6\delta^+ + 2)\,(\log_2((6\mu + 2)\alpha + 2\mu) - 1) + 2\mu \\
\Rightarrow\ & \alpha \le (6\delta^+ + 2)\,(\log_2((3\mu + 1)\alpha + \mu))
\end{aligned} \tag{A.7}$$

• *If* $\mathcal{P}_2$ *is a* ε*-additive-approximate short path then*

$$\begin{aligned}
\ell(\mathcal{P}_4) \le\ & 8\alpha + 2 + \varepsilon \\
\le\ & 8\,(6\delta^+ + 2)\,(\log_2 \ell(\mathcal{P}_4) - 1) + 2 + \varepsilon \\
\le\ & 8\,(6\delta^+ + 2)\,(\log_2(8\alpha + 2 + \varepsilon) - 1) + 2 + \varepsilon \\
\Rightarrow\ & 8\alpha + 2 + \varepsilon \\
& \le 8\,(6\delta^+ + 2)\,(\log_2(8\alpha + 2 + \varepsilon) - 1) + 2 + \varepsilon \\
\equiv\ & \alpha \le (6\delta^+ + 2)\left(\log_2\left(4\alpha + 1 + \tfrac{\varepsilon}{2}\right)\right)
\end{aligned} \tag{A.8}$$

APPENDIX (Continued)

Both (Equation A.7) *and* (Equation A.8) *are of the form* $\alpha \le a \log_2 (b\,\alpha + c) \equiv 2^{\alpha/a} \le b\,\alpha + c$ *where*

$$a = 6\,\delta^{+} + 2 \ge 1 \quad \textit{for both} \text{ (Equation A.7) } \textit{and} \text{ (Equation A.8)}$$

$$b = \begin{cases} 3\,\mu + 1 \ge 4 & \textit{for} \text{ (Equation A.7)} \\ 4 & \textit{for} \text{ (Equation A.8)} \end{cases} \qquad c = \begin{cases} \mu \ge 1 & \textit{for} \text{ (Equation A.7)} \\ 1 + \frac{\varepsilon}{2} \ge 1 & \textit{for} \text{ (Equation A.8)} \end{cases}$$

Thus, α *is at most* z_0 *where* z_0 *is the largest positive integer value of* z *that satisfies the equation:*

$$2^{z/a} \le b\,z + c$$

In the sequel, we will use the fact that $\log_2 (x\,y + 1) \ge \log_2 (x + y)$ *for* $x, y \ge 1$. *This holds since*

$$x \ge 1 \;\&\; y \ge 1 \;\Rightarrow\; y\,(x - 1) \ge x - 1 \;\equiv\; x\,y + 1 \ge x + y$$

We claim that $z_0 \le \eta = a \log_2 (2\,a\,b \log_2 (a\,b\,c) + c)$. *This is verified by showing that* $2^{\eta/a} \ge b\,\eta + c$ *as follows:*

$$2^{\eta/a} = 2^{\log_2 \left(2\,a\,b \log_2 (a\,b\,c) + c\right)} = 2\,a\,b \log_2 (a\,b\,c) + c$$

$$b\,\eta + c = a\,b\,(\log_2 (2\,a\,b \log_2 (a\,b\,c) + c)) + c$$

APPENDIX (Continued)

$$2^{\eta/a} > b\,\eta + c$$

$$\equiv \quad 2\,a\,b\,\log_2(a\,b\,c) + c \geq a\,b\,(\log_2(2\,a\,b\log_2(a\,b\,c) + c)) + c$$

$$\equiv \quad 2\,\log_2(a\,b\,c) \geq \log_2(2\,a\,b\log_2(a\,b\,c) + c)$$

$$\Leftarrow \quad 2\,\log_2(a\,b\,c) \geq \log_2(2\,a\,b\,c\log_2(a\,b\,c) + 1)$$

since $2\,a\,b\log_2(a\,b\,c) \geq 1$ *and* $c \geq 1$

$$\equiv \quad (a\,b\,c)^2 \geq 2\,a\,b\,c\log_2(a\,b\,c) + 1$$

$$\Leftarrow \quad a\,b\,c \geq \log_2(a\,b\,c) + 1$$

and the very last inequality holds since $a\,b\,c \geq 4$. *Thus, we arrive at the at the following bounds:*

• *If* $\mathcal{P}_2$ *is a* μ*-approximate short path then*

$$\eta = \left(6\,\delta^+ + 2\right)\log_2\left(\left(6\,\mu + 2\right)\left(6\,\delta^+ + 2\right)\log_2\left[\,(6\,\delta^+ + 2)\;(3\,\mu + 1)\,\mu\right] + \mu\right)$$

• *If* $\mathcal{P}_2$ *is a* ε*-additive-approximate short path then*

$$\eta = \left(6\,\delta^+ + 2\right)\log_2\left(8\left(6\,\delta^+ + 2\right)\log_2\left[\,(6\,\delta^+ + 2)\;(4 + 2\varepsilon)\;\right] + 1 + \frac{\varepsilon}{2}\right)$$

(b) *Let the ordered sequence of nodes in the path* $\mathcal{P}_3 = v_1 \overset{\mathcal{P}_2}{\leftrightsquigarrow} v_1'$ *be a (length)* maximal *sequence of nodes such that:*

$$\forall\, v' \in \mathcal{P}_3 :\; \min_{v \in \mathcal{P}_1}\{\,d_{v,v'}\} > Z_{\mathcal{P}_1,\mathcal{P}_2}$$

Consider the following set of nodes belonging to the two paths $u_0 \overset{\mathcal{P}_2}{\leftrightsquigarrow} v_1$ *and* $v_1' \overset{\mathcal{P}_2}{\leftrightsquigarrow} u_1$*:*

$$\mathcal{S}_\ell = \bigcup\left\{v' \in u_0 \overset{\mathcal{P}_2}{\leftrightsquigarrow} v_1 \;\middle|\; \exists\, v \in \mathcal{P}_1 :\; d_{v,v'} = \min_{v'' \in \mathcal{P}_2}\{\,d_{v,v''}\}\right\}$$

$$\mathcal{S}_r = \bigcup\left\{v' \in v_1' \overset{\mathcal{P}_2}{\leftrightsquigarrow} u_1 \;\middle|\; \exists\, v \in \mathcal{P}_1 :\; d_{v,v'} = \min_{v'' \in \mathcal{P}_2}\{\,d_{v,v''}\}\right\}$$

APPENDIX (Continued)

Since $u_0 \in \mathcal{S}_\ell$ and $u_1 \in \mathcal{S}_r$, it follows that $\mathcal{S}_\ell \neq \emptyset$ and $\mathcal{S}_r \neq \emptyset$. Note that

$$\bigcup \left\{ v \in u_0 \overset{\mathcal{P}_1}{\leftrightsquigarrow} u_1 \,\middle|\, \exists\, v' \in \mathcal{S}_\ell \cup \mathcal{S}_r \colon d_{v,v'} = \min_{v'' \in \mathcal{P}_2} \{ d_{v,v''} \} \right\} = \bigcup_{v \in u_0 \overset{\mathcal{P}_1}{\leftrightsquigarrow} u_1} \{ v \}$$

Thus, there exists two adjacent nodes v_4 and v_4' on $\mathcal{P}_1$ such that both d_{v_4,v_3} and $d_{v_4',v_3'}$ is at most $Z_{\mathcal{P}_1,\mathcal{P}_2}$. Using triangle inequality it follows that

$$d_{v_3,v_3'} \leq d_{v_3,v_4} + d_{v_4,v_4'} + d_{v_4',v_3'} = 2\, Z_{\mathcal{P}_1,\mathcal{P}_2} + 1$$

giving the following bounds

$$\ell\left(v_3 \overset{\mathcal{P}_2}{\leftrightsquigarrow} v_3'\right) \leq \begin{cases} \mu\, d_{v_3,v_3'} \leq 2\,\mu\, Z_{\mathcal{P}_1,\mathcal{P}_2} + \mu, \\ \qquad \text{if } \mathcal{P}_2 \text{ is } \mu\text{-approximate short} \\ d_{v_3,v_3'} + \varepsilon \leq 2\, Z_{\mathcal{P}_1,\mathcal{P}_2} + 1 + \varepsilon, \\ \qquad \text{if } \mathcal{P}_2 \text{ is } \varepsilon\text{-additive-approximate short} \end{cases}$$

For any node v' on $\mathcal{P}_3$, we can always use the following path to reach a node on $\mathcal{P}_1$:

- *if $d_{v',v_3} \leq d_{v',v_3'}$ then we take the path $v' \overset{\mathcal{P}_2}{\leftrightsquigarrow} v_3 \overset{\text{s}}{\leftrightsquigarrow} v_4$ of length at most $\left\lfloor \frac{\ell\left(v_3 \overset{\mathcal{P}_2}{\leftrightsquigarrow} v_3'\right)}{2} \right\rfloor + Z_{\mathcal{P}_1,\mathcal{P}_2}$ to reach the node $v = v_4$ on $\mathcal{P}_1$;*
- *otherwise we take the path $v' \overset{\mathcal{P}_2}{\leftrightsquigarrow} v_3' \overset{\text{s}}{\leftrightsquigarrow} v_4'$ of length at most $\left\lfloor \frac{\ell\left(v_3 \overset{\mathcal{P}_2}{\leftrightsquigarrow} v_3'\right)}{2} \right\rfloor + Z_{\mathcal{P}_1,\mathcal{P}_2}$ to reach the node $v = v_4'$ on $\mathcal{P}_1$.*

This gives the following worst-case bounds for $d_{v,v'}$:

$$d_{v,v'} \leq \begin{cases} \left\lfloor (\mu + 1)\, Z_{\mathcal{P}_1,\mathcal{P}_2} + \frac{\mu}{2} \right\rfloor, \text{ if } \mathcal{P}_2 \text{ is } \mu\text{-approximate short} \\ \left\lfloor 2\, Z_{\mathcal{P}_1,\mathcal{P}_2} + \frac{1+\varepsilon}{2} \right\rfloor, \text{ if } \mathcal{P}_2 \text{ is } \varepsilon\text{-additive-approximate short} \end{cases}$$

APPENDIX (Continued)

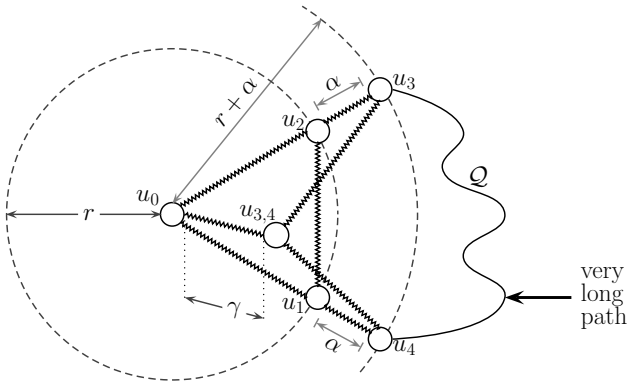


Figure 18: Illustration of the claims in Theorem 12 and Corollary 13.

❑

A.2.3 Theorem 12 and Corollary 13

Theorem 12 (see Fig. Figure 18 for a visual illustration) *Suppose that we are given the following:*

- *three integers* $\kappa \geq 4$, $\alpha > 0$, $r > \left(\frac{\kappa}{2} - 1\right)\left(6\,\delta^{+}_{\text{worst}}(G) + 2\right)$,
- *five nodes* u_0, u_1, u_2, u_3, u_4 *such that*
  - $u_1, u_2 \in B_r(u_0)$ *with* $d_{u_1,u_2} \geq \frac{\kappa}{2}\left(6\,\delta^{+}_{\text{worst}}(G) + 2\right)$,
  - $d_{u_1,u_4} = d_{u_2,u_3} = \alpha$.

Then, the following statements are true for any shortest path $\mathcal{P}$ *between* u_3 *and* u_4*:*

(*a*) *there exists a node* v *on* $\mathcal{P}$ *such that*

$$d_{u_0,v} \leq r - \left(\frac{3\kappa - 2}{12}\right)(6\,\delta^{+}_{\text{worst}}(G) + 2) = r - \mathrm{O}\Big(\kappa\,\delta^{+}_{\text{worst}}(G)\Big)$$

(*b*) $\ell(\mathcal{P}) \geq \left(\frac{3\kappa-2}{6}\right)\left(6\,\delta^{+}_{\text{worst}}(G) + 2\right) + 2\,\alpha = \Omega\left(\kappa\,\delta^{+}_{\text{worst}}(G) + \alpha\right)$.

Corollary 13 (see Fig. Figure 18 for a visual illustration) *Consider any path* $\mathcal{Q}$ *between* u_3 *and* u_4 *that does not involve a node in* $\cup_{r' \leq r}\mathcal{B}_{r'}(u_0)$*. Then, the following statements hold:*

APPENDIX (Continued)

(i) $\ell(Q) \geq 2^{\frac{\alpha}{6\,\delta^+_{\text{worst}}(G)+2} + \frac{\kappa}{4} + \frac{5}{6}} - 1 = 2^{\Omega\left(\frac{\alpha}{\delta^+_{\text{worst}}(G)} + \kappa\right)}$. *In particular, if* $\delta^+_{\text{worst}}(G)$ *is a constant then* $\ell(Q) = 2^{\Omega(\alpha+\kappa)}$ *and thus* $\ell(Q)$ *increases at least exponentially with both* α *and* κ.

(ii) *if* Q *is a* μ*-approximate short path then*

$$\mu \geq \frac{2^{\frac{\alpha}{6\,\delta^+_{\text{worst}}(G)+2} + \frac{\kappa}{4} - \frac{1}{6}}}{12\,\alpha + (3\,\kappa - 26 - \text{o}(1))\,(6\,\delta^+_{\text{worst}}(G) + 2)} - \frac{1}{3} = \Omega\left(\frac{2^{\Theta\left(\frac{\alpha}{\delta^+_{\text{worst}}(G)} + \kappa\right)}}{\alpha + \kappa\,\delta^+_{\text{worst}}(G)}\right)$$

In particular, if $\delta^+_{\text{worst}}(G)$ *is a constant then* $\mu = \Omega\left(\frac{2^{\Theta(\alpha+\kappa)}}{\alpha+\kappa}\right)$ *and thus* μ *increases at least exponentially with both* α *and* κ.

(iii) *if* Q *is a* ε*-additive-approximate short path then*

$$\varepsilon > \frac{2^{\frac{\alpha}{6\,\delta^+_{\text{worst}}(G)+2} + \frac{\kappa}{4} - \frac{1}{6}}}{48\,\delta^+_{\text{worst}}(G) + \frac{17}{2}} - \log_2(48\,\delta^+_{\text{worst}}(G) + 16)$$

In particular, if $\delta^+_{\text{worst}}(G)$ *is a constant then* $\varepsilon = \Omega\left(2^{\Theta(\alpha+\kappa)}\right)$ *and thus* ε *increases at least exponentially with both* α *and* κ.

A.2.3.1 Proof of Theorem 12

Consider the shortest-path triangle $\Delta_{\{u_0,u_3,u_4\}}$ *and let* $u_{0,3}, u_{0,4}$ *and* $u_{3,4}$ *be the Gromov product nodes of* $\Delta_{\{u_0,u_3,u_4\}}$ *on the sides (shortest paths)* u_0 *to* u_3, u_0 *to* u_4 *and* u_3 *to* u_4, *respectively. Thus,* $d_{u_0,u_{0,3}} = d_{u_0,u_{0,4}}$, *and* $\beta = d_{u_3,u_{3,4}} = \left\lfloor \frac{d_{u_0,u_3} + d_{u_3,u_4} - d_{u_0,u_4}}{2} \right\rfloor = \left\lfloor \frac{d_{u_3,u_4}}{2} \right\rfloor$ *since* $d_{u_0,u_3} = d_{u_0,u_4} = r + \alpha$.

We first claim that $d_{u_0,u_{0,3}} < r = d_{u_0,u_2}$. *Suppose for the sake of contradiction that* $d_{u_0,u_{0,3}} = d_{u_0,u_{0,4}} \geq r$. *Then, by Theorem 5 we get* $d_{u_1,u_2} \leq 6\,\delta^+_{\text{worst}}(G) + 2$ *which contradicts the assumption that* $d_{u_1,u_2} \geq \frac{\kappa}{2}\left(6\,\delta^+_{\text{worst}}(G) + 2\right)$ *since* $\kappa \geq 4$.

Thus, assume that $d_{u_0,u_{0,3}} = d_{u_0,u_{0,4}} = r - x$ *for some integer* $x > 0$. *By Theorem 5,* $d_{u_{0,3},u_{0,4}} \leq 6\,\delta^+_{\text{worst}}(G) + 2$. *Let* $d_{u_{0,3},u_{0,4}} = 6\,\delta^+_{\text{worst}}(G) + 2 - y$ *for some integer* $0 < y \leq 6\,\delta^+_{\text{worst}}(G) + 2$ *and* $d_{u_1,u_2} = \frac{\kappa}{2}\left(6\,\delta^+_{\text{worst}}(G) + 2\right) + z$

APPENDIX (Continued)

for some integer $z \geq 0$*. Consider the* 4*-node condition for the four nodes* $u_1, u_2, u_{0,3}, u_{0,4}$*. The three relevant quantities for comparison are:*

$$q_{\parallel} = d_{u_1,u_2} + d_{u_{0,3},u_{0,4}} = \left(\tfrac{\kappa}{2} + 1\right)\left(6\,\delta^+_{\text{worst}}(G) + 1\right) + z - y$$

$$q_{=} = d_{u_{0,3},u_2} + d_{u_{0,4},u_1} = \left(d_{u_0,u_2} - d_{u_0,u_{0,3}}\right) + \left(d_{u_0,u_1} - d_{u_0,u_{0,4}}\right) = 2x$$

$$\begin{aligned} q_{\backslash\backslash} = d_{u_{0,3},u_1} + d_{u_{0,4},u_2} &\leq \left(d_{u_{0,3},u_{0,4}} + d_{u_{0,4},u_1}\right) + \left(d_{u_{0,3},u_{0,4}} + d_{u_{0,3},u_2}\right) \\ &= 12\,\delta^+_{\text{worst}}(G) + 4 - 2y + 2x \end{aligned}$$

We now show that $x > \left(\frac{3\kappa-2}{12}\right)\left(6\,\delta^+_{\text{worst}}(G) + 2\right)$*. We have the following cases.*

- *Assume that* $q_{\backslash\backslash} \leq \min\{q_{\parallel},\, q_{=}\}$*. This implies*

$$\begin{aligned} & \left|q_{\parallel} - q_{=}\right| \leq 2\,\delta^+_{\text{worst}}(G) \\ \equiv \quad & \left|\left(\tfrac{\kappa}{2} + 1\right)\left(6\,\delta^+_{\text{worst}}(G) + 2\right) + z - y - 2x\right| \leq 2\,\delta^+_{\text{worst}}(G) \\ \Rightarrow \quad & x \geq \frac{\left(\frac{\kappa}{2} + 1\right)\left(6\,\delta^+_{\text{worst}}(G) + 2\right) + z - y - 2\,\delta^+_{\text{worst}}(G)}{2} \\ & \geq \left(\tfrac{3\kappa-2}{12}\right)\left(6\,\delta^+_{\text{worst}}(G) + 2\right) + \tfrac{1}{6} \end{aligned}$$

APPENDIX (Continued)

- *Otherwise, assume that* $q_{=} \leq \min\{q_{\parallel},\, q_{\backslash\backslash}\}$. *This implies*

$$
\begin{aligned}
& \left|q_{\parallel} - q_{\backslash\backslash}\right| \leq 2\,\delta^{+}_{\text{worst}}(G) \\
\Rightarrow \quad & q_{\backslash\backslash} \geq q_{\parallel} - 2\,\delta^{+}_{\text{worst}}(G) \\
\Rightarrow \quad & d_{u_{0,3},u_1} + d_{u_{0,4},u_2} \geq \left(\tfrac{\kappa}{2}+1\right)\left(6\,\delta^{+}_{\text{worst}}(G)+2\right) + z - y - 2\,\delta^{+}_{\text{worst}}(G) \\
\Rightarrow \quad & (d_{u_{0,3},u_{0,4}} + d_{u_{0,4},u_1}) + (d_{u_{0,3},u_{0,4}} + d_{u_{0,3},u_2}) \geq d_{u_{0,3},u_1} + d_{u_{0,4},u_2} \\
& \qquad \geq \left(\tfrac{\kappa}{2}+1\right)\left(6\,\delta^{+}_{\text{worst}}(G)+2\right) + z - y - 2\,\delta^{+}_{\text{worst}}(G) \\
\Rightarrow \quad & 2x + 2\left(6\,\delta^{+}_{\text{worst}}(G) + 2 - y\right) \\
& \qquad \geq \left(\tfrac{\kappa}{2}+1\right)\left(6\,\delta^{+}_{\text{worst}}(G)+2\right) + z - y - 2\,\delta^{+}_{\text{worst}}(G) \\
\Rightarrow \quad & x \geq \left(\tfrac{3\kappa-2}{12}\right)\left(6\,\delta^{+}_{\text{worst}}(G)+2\right) + \tfrac{1}{6}
\end{aligned}
$$

- *Otherwise, assume that* $q_{\parallel} \leq \min\{q_{=},\, q_{\backslash\backslash}\}$. *This implies*

$$
\begin{aligned}
& \left|q_{=} - q_{\backslash\backslash}\right| \leq 2\,\delta^{+}_{\text{worst}}(G) \\
\equiv \quad & \left|\,2x - (d_{u_{0,3},u_1} + d_{u_{0,4},u_2})\,\right| \leq 2\,\delta^{+}_{\text{worst}}(G) \\
\Rightarrow \quad & 2x \geq d_{u_{0,3},u_1} + d_{u_{0,4},u_2} - 2\,\delta^{+}_{\text{worst}}(G) \\
& \qquad \geq (d_{u_1,u_2} - d_{u_{0,4},u_1}) + (d_{u_1,u_2} - d_{u_{0,3},u_1}) - 2\,\delta^{+}_{\text{worst}}(G) \\
\equiv \quad & 2x \geq \kappa\left(6\,\delta^{+}_{\text{worst}}(G)+2\right) + 2z - 2x - 2\,\delta^{+}_{\text{worst}}(G) \\
\Rightarrow \quad & x \geq \left(\tfrac{3\kappa-2}{12}\right)\left(6\,\delta^{+}_{\text{worst}}(G)+2\right) + \tfrac{\delta^{+}_{\text{worst}}(G)}{2} + \tfrac{1}{6}
\end{aligned}
$$

Using Theorem 5, it now follows that

$$
d_{u_0,u_{3,4}} \;\leq\; d_{u_0,u_{0,3}} + d_{u_{0,3},u_{0,4}} \;\leq\; (r - x) + (6\,\delta^{+}_{\text{worst}}(G)+2) \;<\; r - \left(\frac{3\kappa-2}{12}\right)(6\,\delta^{+}_{\text{worst}}(G)+2)
$$

APPENDIX (Continued)

This proves part (***a***) *with* $u_{3,4}$ *being the node in question. To prove part* (***b***)*, note that*

$$|\mathcal{P}| = 2\beta \geq 2(r+\alpha) - 2d_{u_0,u_{3,4}} \geq 2\alpha + \left(\frac{3\kappa-2}{6}\right)(6\,\delta^+_{\text{worst}}(G)+2)$$

A.2.3.2 Proof of Corollary 13

Consider such a path Q *and consider the node* $u_{3,4}$ *on the shortest path between* u_3 *and* u_4*. Since every node of* Q *is at a distance strictly larger than* $r+\alpha$ *from* u_0*, by Theorem 12 the following holds for every node* $v \in Q$

$$d_{u_{3,4},v} \;\geq\; (r+\alpha) - d_{u_0,u_{3,4}} \;=\; (r+\alpha) - \left(r - \left(\frac{3\kappa-2}{12}\right)(6\,\delta^+_{\text{worst}}(G)+2)\right) \;=\; \alpha + \left(\frac{3\kappa-2}{12}\right)(6\,\delta^+_{\text{worst}}(G)+2)$$

Thus, by Corollary 8 (with $\gamma = \alpha + \left(\frac{3\kappa-2}{12}\right)\left(6\,\delta^+_{\text{worst}}(G)+2\right)$*), we get*

$$\ell\,(Q) \geq 2^{\frac{\gamma}{6\,\delta^+_{\text{worst}}(G)+2}+1} - 1 = 2^{\frac{\alpha}{6\,\delta^+_{\text{worst}}(G)+2} + \frac{\kappa}{4} + \frac{5}{6}} - 1$$

If Q *is a* μ*-approximate short path, then by Corollary 11:*

$$\mu \;>\; \frac{2^{\frac{\gamma}{6\,\delta^+_{\text{worst}}(G)+2}}}{12\,\gamma - \left(24+\text{o}(1)\right)\left(6\,\delta^+_{\text{worst}}(G)+2\right)} - \frac{1}{3} \;=\; \frac{2^{\frac{\alpha}{6\,\delta^+_{\text{worst}}(G)+2} + \frac{\kappa}{4} - \frac{1}{6}}}{12\,\alpha + (3\,\kappa - 26 - \text{o}(1))\,(6\,\delta^+_{\text{worst}}(G)+2)} - \frac{1}{3}$$

If Q *is a* ε*-additive-approximate short path, then by Corollary 11:*

$$\varepsilon \;>\; \frac{2^{\frac{\gamma}{6\,\delta^+_{\text{worst}}(G)+2}}}{48\,\delta^+_{\text{worst}}(G) + \frac{17}{2}} - \log_2\left(48\,\delta^+_{\text{worst}}(G)+16\right) \qquad = \frac{2^{\frac{\alpha}{6\,\delta^+_{\text{worst}}(G)+2} + \frac{\kappa}{4} - \frac{1}{6}}}{48\,\delta^+_{\text{worst}}(G) + \frac{17}{2}} - \log_2\left(48\,\delta^+_{\text{worst}}(G)+16\right)$$

APPENDIX (Continued)

A.2.4 Lemma 14

Lemma 14 **(equivalence of strong and weak domination; see Fig. Figure 9 for a visual illustration)** *If* $\lambda \geq \left(6\,\delta^{+}_{\text{worst}}(G) + 2\right)\log_2 n$ *then*

$$\mathfrak{M}_{u,\rho,\lambda} \stackrel{\text{def}}{=\!=} \mathbb{E}\left[\begin{array}{l|l} \textit{number of pairs of nodes} & v \text{ is selected} \\ v,y \text{ such that } v,y \text{ is} & \textit{uniformly ran-} \\ \textbf{weakly } (\rho,\lambda)\textit{-dominated} & \textit{domly from} \\ \textit{by } u & \cup_{\rho<j\leq\lambda}\mathcal{B}_j(u) \end{array}\right]$$

$$= \mathbb{E}\left[\begin{array}{l|l} \textit{number of pairs of} & v \text{ is selected} \\ \textit{nodes } v,y \textit{ such that} & \textit{uniformly ran-} \\ v,y \textit{ is } \textbf{strongly } (\rho,\lambda)\text{-} & \textit{domly from} \\ \textit{dominated by } u & \cup_{\rho<j\leq\lambda}\mathcal{B}_j(u) \end{array}\right]$$

Proof. *Suppose that* v,y *is weakly* (ρ,λ)*-dominated by u,* i.e.*, there exists a shortest path* $v \overset{\mathcal{P}}{\rightsquigarrow} y$ *between* $v,y \in \mathcal{B}_{\rho+\lambda}(u)$ *such that for some node* $v' \in v \overset{\mathcal{P}}{\rightsquigarrow} y$ *we have* $v' \in \mathcal{B}_{\rho}(u)$*. Let* $v \overset{\mathcal{Q}}{\rightsquigarrow} y$ *be any other path between v and y that does* not *contain a node from* $\mathcal{B}_{\rho}(u)$*. Then, by Corollary 13***(i)** *(with* $\kappa = 4$*) we have*

$$\ell(\mathcal{Q}) \geq 2^{\frac{\lambda}{6\,\delta^{+}_{\text{worst}}(G)+2} + \frac{11}{6}} - 1 \geq 2^{\log_2 n + \frac{11}{6}} - 1 > n - 1$$

which contradicts the obvious bound $\ell(\mathcal{Q}) < n$*. Thus,* no *such path* $\mathcal{Q}$ *exists and* v,y *is strongly* (ρ,λ)*-dominated by u.* ❑

APPENDIX

PROOF OF THEOREMS IN CHAPTER 4

B.1 Theorem 2

A proof of Theorem 2 is implicit in [18]. For the benefit of the reader, we provide a selfcontained proof of Theorem 2 here using elementary graph theory.

(a) *Let $u \overset{s}{\leftrightsquigarrow} v$ be a maximal shortest path in G. Suppose that we select neither u nor v in a solution of solution of Str-Met-Dim on G. Then there exists no node x in our solution of Str-Met-Dim on G such that $x \blacktriangleright \{u, v\}$, implying our solution of Str-Met-Dim on G is not a valid solution and thereby showing $sdim(G) \geq MNC(\widehat{G})$. To prove $sdim(G) \leq MNC(\widehat{G})$, suppose that we select at least one end-point of every maximal shortest path in G. Consider any pair of nodes u and v. If at least one of u or v, say u, is selected in a solution of Str-Met-Dim on G, then $u \blacktriangleright \{u, v\}$. Otherwise, $u \overset{s}{\leftrightsquigarrow} v$ is not a maximal shortest path, and let $x \overset{s}{\leftrightsquigarrow} y$ be a maximal shortest path containing u and v. Then, we have selected at least one of x or y, say x, in a solution of Str-Met-Dim on G, and $x \blacktriangleright \{u, v\}$.*

(b) *It follows from the construction of $\widetilde{G}$ that $diam(\widetilde{G}) = 2$ since any pair of nodes has a shortest path of length at most 2 between them via y. Note that, for any pair of nodes u and v, $\mathsf{Nbr}(u) = \mathsf{Nbr}(v)$ in G if and only if $\mathsf{Nbr}(u) = \mathsf{Nbr}(v)$ in $\overline{G}$. To show $sdim(\widetilde{G}) \leq \kappa + MNC(G)$, let $S \subset V$ be the set of nodes in a minimum node cover of G of cardinality $MNC(G)$. Consider the set of $\kappa + MNC(G)$ nodes in $S = S \cup \{x_1, x_2, ..., x_\kappa\}$ as a possible solution of Str-Met-Dim on $\widetilde{G}$. To show that this is indeed a valid solution, consider any pair of nodes u and v in $\widetilde{G}$. Then the following simple case analysis suffices:*

- *Suppose that at least one of u and v is x_i for some i. Then,$S' \ni x_i \blacktriangleright \{u, v\}$.*

APPENDIX (Continued)

- *Otherwise, suppose that one of u and v, say u, is y (and thus $v \in V$). Select a node $x_i \in S'$ such that $\{x_i, v\} \notin \widetilde{E}$. Then the shortest path of length 2 from x_i to v formed by the edges $\{x_i, y\}$ and $\{y, v\}$ shows that $S \ni xi \blacktriangleright \{u, v\}$.*
- *Otherwise, if $\{u, v\} \in E$ then at least one of u and v, say u, is in S and $u \blacktriangleright \{u, v\}$.*
- *Otherwise, $\{u, v\} \notin E$. Thus, $\{u, v\} \in \widetilde{E}$. If at least one of u and v, say u, is in S then $u \in S$ and $u \blacktriangleright \{u, v\}$. Otherwise, both of u and v are not in S , and there are the following two sub-cases to consider.*
  - *At least one of u and v, say u, is u_i for some i. Then the shortest path of length 2 from x_i to v formed by the edges $\{x_i, u_i\}$ and $\{u_i, v\}$ shows that $S \ni x_i \blacktriangleright \{u, v\}$.*
  - *Otherwise, $\mathsf{Nbr}(u) \neq \mathsf{Nbr}(v)$ in G, which implies that there exists a node $u' \in V$ such that u is adjacent to exactly one of u and v, say u. Thus, $\{u, u\} \notin \widetilde{E}$ but $\{v, u\} \in \widetilde{E}$. Note that $u \notin S$ and $\{u, u\} \in E$ implies u is in S . Then the shortest path of length 2 from u to u formed by the edges $\{u, v\}$ and $\{v, u\}$ shows that $S \ni u \blacktriangleright \{u, v\}$.*

To show $sdim(\widetilde{G}) \geq \kappa + MNC(G)$, let $S \subset \widetilde{V}$ be the set of $sdim(\widetilde{G})$ nodes in an optimal solution of Str-Met-Dim on $\widetilde{G}$. Consider the set of nodes in $S = S \{x_1, x_2, ..., x_\kappa, y\}$ as a possible solution of the node cover problem of G. We first show that S is in fact a valid node cover of G. Since $diam(\widetilde{G}) = 2$, any shortest path in G is of length at most 2. Consider an edge $\{u, v\} \in E$ and suppose that both u and v are not in S (and thus also not in S). Since $\{u, v\} \notin \widetilde{E}$, the length of any shortest path between u and v is exactly 2, and thus no node $x \in \widetilde{V}$ $\{u, v\}$ can strongly resolve the pair of nodes u and v, resulting in a contradiction that S is a solution of Str-Met-Dim on $\widetilde{G}$. Thus, S is a node cover of G and $MNC(G) \leq |S|$. To show that $|S| = |S|\kappa$, note that:

- *Every x_i must belong to S since otherwise no node in S can strongly resolve the pair of nodes x_i and x_j for any $j \neq i$.*

APPENDIX (Continued)

- *Since every x_i belongs to S, the node y does not need to belong to S.*

APPENDIX

SUPPLEMENTAL INFORMATION

TABLE XXI: Details of 11 biological networks studied

	name	brief description	# nodes	# edges	reference
1.	*E. coli* transcriptional	*E. coli* transcriptional regulatory network of direct regulatory interactions between transcription factors and the genes or operons they regulate	311	451	(37)
2.	Mammalian signaling	Mammalian network of signaling pathways and cellular machines in the hippocampal CA1 neuron	512	1047	(38)
3.	*E. coli* transcriptional	*E. coli* transcriptional regulatory network of direct regulatory interactions between transcription factors and the genes or operons they regulate	418	544	♯
4.	T-LGL signaling	Signaling network inside cytotoxic T cells in the context of the disease T cell large granular lymphocyte leukemia	58	135	(39)
5.	*S. cerevisiae* transcriptional	*S. cerevisiae* transcriptional regulatory network showing interactions between transcription factor proteins and genes	690	1082	(40)
6.	*C. elegans* metabolic	Network of biochemical reactions (*C. elegans* metabolism)	453	2040	(9)
7.	Drosophila segment polarity (6 cells)	1-dimensional 6-cell version of the gene regulatory network among products of the segment polarity gene family that plays an important role in the embryonic development of *Drosophila melanogaster*	78	132	(41)
8.	ABA signaling	Guard cell signal transduction network for abscisic acid (ABA) induced stomatal closure in plants	55	88	(42)
9.	Immune response network	Network of interactions among immune cells and pathogens in the mammalian immune response against two bacterial species	18	42	(43)
10.	T cell receptor signaling	Network for T cell activation mechanisms after engagement of the TCR, the CD_4/CD8 co-receptors and CD28.	94	138	(44)
11.	Oriented yeast PPI	An oriented version of an unweighted PPI network constructed from *S. cerevisiae* interactions in the BioGRID database	786	2445	(45)

♯ Updated version of the network in (37); see `www.weizmann.ac.il/mcb/UriAlon/Papers/networkMotifs/coli1_1Inter_st.txt`

APPENDIX (Continued)

TABLE XXII: Details of 9 social networks studied

	name	brief description	type	# nodes	# edges	reference
1.	Dolphin social network	Social network of frequent associations between 62 dolphins in a community living off Doubtful Sound in New Zealand	undirected, unweighted	62	160	(46)
2.	American College Football	Network of American football games between Division IA colleges during the regular Fall 2000 season	undirected, unweighted	115	612	(47)
3.	Zachary Karate Club	Network of friendships between 34 members of a karate club at a US university in the 1970s	undirected, unweighted	34	78	(48)
4.	Books about US politics	Network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller `amazon.com`; edges between books represent frequent copurchasing of books by the same buyers.	undirected, unweighted	105	442	‡
5.	Sawmill communication network	A communication network within a small enterprise: a sawmill. All employees were asked to indicate the frequency with which they discussed work matters with each of their colleagues on a five-point scale ranging from less than once a week to several times a day. Two employees were linked in the network if they rated their contact as three or more.	undirected, unweighted	36	62	(49)
6.	Jazz Musician network	A social network of Jazz musicians	undirected, unweighted	198	2742	(50)
7.	Visiting ties in San Juan	Network for visiting relations between families living in farms in the neighborhood San Juan Sur, Costa Rica, 1948	undirected, unweighted	75	144	(51)
8.	World Soccer Data, Paris 1998	Members of the 22 soccer teams which participated in the World Championship in Paris in 1998 had contracts in 35 countries. Counts of which team exports how many players to which country are used to generate this network.	directed, weighted	35	118	†
9.	Les Miserables	Network of co-appearances of characters in Victor Hugo's novel "Les Miserables". Nodes represent characters as indicated by the labels and edges connect any pair of characters that appear in the same chapter of the book. The weights on the edges are the number of such coappearances.	undirected, weighted	77	251	(52)

‡ V. Krebs, unpublished manuscript, found on Krebs' website `www.orgnet.com`.

† Dagstuhl seminar: *Link Analysis and Visualization*, Dagstuhl 1-6, 2001.
(see `http://vlado.fmf.uni-lj.si/pub/networks/data/sport/football.htm`)

APPENDIX (Continued)

Biological details of source, target and central nodes (u_{source}, u_{target} and u_{central}) used in Table VIII and Table IX

Network 1: E. coli transcriptional

Node name	Node type	Details
fliAZY	u_{source}	Contains fliA gene (sigma factor), fliZ (possible cell-density responsive regulator of sigma) and fliY (periplasmic cystine-binding protein)
fecA	u_{source}	Ferric citrate, outer membrane receptor
arcA	u_{target}	Aerobic respiration control, transcriptional dual regulator
aspA	u_{target}	Component of aspartate ammonia-lyase
crp	u_{central}	Component of CRP transcriptional dual regulator (DNA-binding transcriptional dual regulator)
CaiF	u_{central}	DNA-binding transcriptional activator
sodA	u_{central}	Component of superoxide dismutases that catalyzes the dismutation of superoxide into oxygen and hydrogen peroxide

APPENDIX (Continued)

Network 4: T-LGL signaling network

Node name	Node type	Details
PDGF	u_{source}	Platelet-derived growth factor is one of the numerous growth factors, or proteins that regulates cell growth and division.
IL15	u_{source}	Interleukin 15 is a cytokine.
Stimuli	u_{source}	Antigen Stimulation
apoptosis	u_{target}	process of programmed cell death
IL2	u_{central}	Interleukin 2 is a cytokine signaling molecule in the immune system
Ceramide	u_{central}	A waxy lipid molecule within the cell membrane which can participate in variety of cellular signaling like proliferation and apoptosis
GZMB	u_{central}	A serine proteases that is released within cytotoxic T cells and natural killer cells to induce apoptosis within virus-infected cells, thus destroying them
NFKB	u_{central}	nuclear factor kappa-light-chain-enhancer of activated B cells, a protein complex that controls the transcription of DNA
MCL1	u_{central}	Induced myeloid leukemia cell differentiation protein Mcl-1

CITED LITERATURE

1. *Watts, D. J. and Strogatz, S. H.: Collective dynamics of small-worldnetworks. nature, 393(6684):440, 1998.*

2. *Barábasi, A. L. and Albert, R.: Emergence of scaling in random networks. Science, 286:509–512, 1999.*

3. *Girvan, M. and Newman, M. E.: Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12):7821–7826, 2002.*

4. *Travers, J. and Milgram, S.: The small world problem. Phychology Today, 1(1):61–67, 1967.*

5. *Erdös, P. and Rènyi, A.: On random graphs i. Publ. Math. Debrecen, 6:290–297, 1959.*

6. *Erdös, P. and Rényi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci, 5(1):17–60, 1960.*

7. *Albert, R. and l. Barabási, A.: Statistical mechanics of complex networks. Reviews in Modern Physics, 74:47–97, 2002.*

8. *Newman, M. E. J.: Scientific collaboration networks: Ii. shortest paths, weighted networks, and centrality. Physical Review E, 64:016132, 2001.*

9. *Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and l. Barabasi, A.: The large-scale organization of metabolic networks. Nature, 407:651–654, 2000.*

10. *Redner, S.: How popular is your paper? an empirical study of the citation distribution. The European Physical Journal B-Condensed Matter and Complex Systems, 4(2):131–134, 1998.*

11. *Chen, Q., Chang, H., Govindan, R., and Jamin, S.: The origin of power laws in internet topologies revisited. In INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, volume 2, pages 608–617. IEEE, 2002.*

12. *Albert, R., Jeong, H., and Barabási, A.-L.: Internet: Diameter of the world-wide web. nature, 401(6749):130, 1999.*

13. *Barabási, A.-L. and Pósfai, M.: Network science. Cambridge university press, 2016.*

14. *Newman, M. E. J.: Networks: An Introduction. Oxford University Press, 2010.*

15. *Colizza, V., Flammini, A., Serrano, M. A., and Vespignani, A.: Detecting rich-club ordering in complex networks. Nature Physics, 2:110–115, 2006.*

16. *Latora, V. and Marchior, M.: A measure of centrality based on network efficiency. New Journal of Physics, 9:188, 2007.*

17. *Albert, R., DasGupta, B., Gitter, A., Gürsoy, G., Hegde, R., Pal, P., Sivanathan, G. S., Sontag, E., and New, A.: Computationally efficient measure of topological redundancy of biological and social networks. Physical Review E, 84(3):036117, 2011.*

18. *Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., and Grafton, S. T.: Dynamic reconfiguration of human brain networks during learning. Proc Natl Acad Sci USA, 108(18):7641–7646, 2011.*

19. *Gromov, M.: Hyperbolic groups. Essays in group theory, 8:75–263, 1987.*

20. *Jonckheere, E. A. and Lohsoonthorn, P.: Geometry of network security. Proceedings of the American Control Conference, 2:976–981, 2004.*

21. *Jonckheere, E., Lohsoonthorn, P., and Bonahon, F.: Scaled gromov hyperbolic graphs. Journal of Graph Theory, 57(2):157–180, 2007.*

22. *Ariaei, F., Lou, M., Jonckeere, E., Krishnamachari, B., and Zuniga, M.: Curvature of sensor network: clustering coefficient. EURASIP Journal on Wireless Communications and Networking, 213185, 2008.*

23. *Narayan, D. and Saniee, I.: Large-scale curvature of networks. Physical Review E, 84:066108, 2011.*

24. *Papadopoulos, F., Krioukov, D., Boguna, M., and Vahdat, A.: Greedy forwarding in dynamic scale-free networks embedded in hyperbolic metric spaces. In Proceedings of the IEEE Conference on Computer Communications, pages 1–9. IEEE Press, 2010.*

25. *Jonckheerea, E., Loua, M., Bonahona, F., and Baryshnikova, Y.: Euclidean versus hyperbolic congestion in idealized versus experimental networks. Internet Mathematics, 7(1):1–27, 2011.*

26. *Bogun, M., Papadopoulos, F., and Krioukov, D.: Sustaining the internet with hyperbolic mapping. Nature Communications, 1(62), 2010.*

27. *De Montgolfier, F., Soto, M., and Viennot, L.: Treewidth and hyperbolicity of the internet. In Network Computing and Applications (NCA), 2011 10th IEEE International Symposium on, pages 25–32. IEEE, 2011.*

28. *Robertson, N. and Seymour, P. D.: Graph minors. i. excluding a forest. Journal of Combinatorial Theory Series B, 35(1):39–61, 1983.*

29. *Bodlaender, H. L.: Dynamic programming on graphs with bounded treewidth. In International Colloquium on Automata, Languages, and Programming, pages 105–118. Springer, 1988.*

30. *Chepoi, V. and Estellon, B.: Packing and covering δ-hyperbolic spaces by balls. In Lecture Notes in Computer Science, pages 59–73. Springer, 2007.*

31. *Chepoi, V., Dragan, F., Estellon, B., Habib, M., and Vaxès, Y.: Diameters, centers, and approximating trees of delta-hyperbolicgeodesic spaces and graphs. In Proceedings of the twenty-fourth annual symposium on Computational geometry, pages 59–68. ACM, 2008.*

32. *Chepoi, V., Dragan, F. F., Estellon, B., Habib, M., Vaxès, Y., and Xiang, Y.: Additive spanners and distance and routing labeling schemes for δ-hyperbolic graphs. Algorithmica, 62(3-4):713–732, 2012.*

33. *Gavoille, C. and Ly, O.: Distance labeling in hyperbolic graphs. In International Symposium on Algorithms and Computation, pages 1071–1079. Springer, 2005.*

34. *Abraham, I., Balakrishnan, M., Kuhn, F., Malkhi, D., Ramasubramanian, V., and Talwar, K.: Reconstructing approximate tree metrics. In Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing, pages 43–52, New York, 2007. ACM Press.*

35. *Krauthgamer, R. and Lee, J. R.: Algorithms on negatively curved spaces. In null, pages 119–132. IEEE, 2006.*

36. *Roe, J.: Index theory, coarse geometry , and topology of manifolds. Conference Board of the Mathematical Sciences Regional Conference Series, 90, 1996.*

37. *Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U.: Network motifs in the transcriptional regulation network of escherichia coli. Nature Genetics, 31:64–68, 2002.*

38. *Ma'ayan, A., Jenkins, S. L., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, B., Eungdamrong, N. J., Weng, G., Ram, P. T., Rice, J. J., Kershenbaum, A., Stolovitzky, G. A., Blitzer, R. D., and Iyengar, R.: Formation of regulatory patterns during signal propagation in a mammalian cellular network. Science, 309 (5737):1078–1083, 2005.*

39. *Zhang, R., Shah, M. V., Yang, J., Nyland, S. B., Liu, X., Yun, J. K., Albert, R., and Loughran, T. P.: Network model of survival signaling in large granular lymphocyte leukemia. Proceedings of the National Academy of Sciences, 2008.*

40. *Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., and Alon, D. U.: Network motifs: simple building blocks of complex networks. Science, 298:824–827, 2002.*

41. *Dassow, G. V., Meir, E., Munro, E. M., and Odell, G. M.: The segment polarity network is a robust developmental module. Nature, 406:188–192, 2000.*

42. *Li, S., Assmann, S. M., and Albert, R.: Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling. PLoS Biology, 4(10):e312, 2006.*

43. *Thakar, J., Pilione, M., Kirimanjeswara, G., Harvill, E. T., and Albert, R.: Modeling systems-level regulation of host immune responses. PLoS Computational Biology, 3(6):e109, 2007.*

44. *Saez-Rodriguez, J., Simeoni, L., Lindquist, J. A., Hemenway, R., Bommhardt, U., Arndt, B., u. Haus, U., Weismantel, R., Gilles, E. D., Klamt, S., and Schraven, B.: A logical model provides insights into t cell receptor signaling. PLoS Computational Biology, 3(8):e163, 2007.*

45. *Gitter, A., Klein-Seetharaman, J., Gupta, A., and Bar-Joseph, Z.: Discovering pathways by orienting edges in protein interaction networks. Nucleic Acids Research, 39(4), 2011.*

46. *Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., and Dawson, S. M.: The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology, 54(4):396–405, 2003.*

47. *Girvan, M. and Newman, M. E. J.: Community structure in social and biological networks. Proc Natl Acad Sci USA, 99(12):7821–7826, 2002.*

48. *Zachary, W. W.: An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 33:452–473, 1977.*

49. *Michael, J. H. and Massey, J. G.: Modeling the communication network in a sawmill. Forest Products Journal, 47:25–30, 1997.*

50. *Gleiser, P. and Danon, L.: Community structure in jazz. Advances in Complex Systems, 6(4):565–573, 2003.*

51. *Loomis, C. P., Morales, J. O., Clifford, R. A., and Leonard, O. E.: Turrialba: Social systems and the introduction of change. (The Free Press, Glencoe, IL, page p. 45 and 78, 1953.*

52. *Knuth, D. E.: The Stanford GraphBase: A Platform for Combinatorial Computing. AcM Press New York, 1993.*

53. *Jonckheere, E., Lohsoonthorn, P., and Ariaei, F.: Scaled gromov four-point condition for network graph curvature computation. Internet Mathematics, 7(3):137–177, 2011.*

54. *Kannan, R., Tetali, P., and Vempala, S.: Markov-chain algorithms for generating bipartite graphs and tournaments. Random Structures and Algorithms, 14:293–308, 1999.*

55. *Alberts, B.: Molecular biology of the cell. New York, Garland Publishers, 1994.*

56. *Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., b. Tagne, J., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A.: Transcriptional regulatory networks in saccharomyces cerevisiae. Science, 298(5594):799–804, 2002.*

57. *Bridson, M. R. and Haefliger, A.: Metric Spaces of Non-Positive Curvature. Springer, 1999.*

58. *Albert, R.: Scale-free networks in cell biology. Journal of Cell Science, 118:4947–4957, 2005.*

59. *Burt, R. S.: Structural Holes: The Social Structure of Competition. Harvard University Press, 1995.*

60. *Borgatti, S. P. and Holes:, S.: Unpacking burt's redundancy measures. Connections, 20(1):35–38, 1997.*

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

62. *Gupta, P., Janardan, R., Smid, M., and DasGupta, B.: The rectangle enclosure and point-dominance problems revisited. International Journal of Computational Geometry & Applications, 7(5):437–455, 1997.*

63. *Slater, P. J.: Leaves of trees. Congr. Numer, 14(549-559):37, 1975.*

64. *Harary, F. and Melter, R. A.: On the metric dimension of a graph. Ars Combinatoria, 2:191–195, 1976.*

65. *Sebo, A. and E. Tannier, O.: Metric generators of graphs. Mathematics of Operations Research, 29(2):383–393, 2004.*

66. *Oellermann, O. R. and J. Peters-Fransen, T.: strong metric dimension of graphs and digraphs. Discrete Applied Mathematics, 155:356–364, 2007.*

67. *Rodriguez-Velazqueza, J. A., Yerob, I. G., Kuziaka, D., and Oellermannc, O. R.: On the strong metric dimension of cartesian and direct products of graphs. Discrete Mathematics, 335:8–19, 2014.*

68. *Yi, E.: On strong metric dimension of graphs and their complements. Acta Mathematica Sinica,English Series, 29(8):1479–1492, 2013.*

69. *Cook, S. A.: The complexity of theorem-proving procedures. In Proceedings of the third annual ACM symposium on Theory of computing, pages 151–158. ACM, 1971.*

70. *Karp, R. M.: Reducibility among combinatorial problems. In Complexity of computer computations, pages 85–103. Springer, 1972.*

71. *Khot, S.: On the power of unique 2-prover 1-round games. In ACM Symposium on Theory of Computing, pages 767–775, 2002.*

72. *Impagliazzo, R. and Paturi, R.: Complexity of k-sat. In Computational Complexity, 1999. Proceedings. Fourteenth Annual IEEE Conference on, pages 237–240. IEEE, 1999.*

73. *Khuller, S., Raghavachari, B., and Rosenfeld, A.: Landmarks in graphs. Discrete Applied Mathematics, 70(3):217–229, 1996.*

74. *Vazirani, V. V.: Approximation algorithms. Springer Science & Business Media, 2013.*

75. *Fomin, F. V. and Kaski, P.: Exact exponential algorithms. Communications of the ACM, 56(3):80–88, 2013.*

76. *Chen, J., Kanj, I. A., and Xia, G.: Improved upper bounds for vertex cover. Theoretical Computer Science, 411(40-42):3736–3756, 2010.*

77. *Khot, S. and Regev, O.: Vertex cover might be hard to approximate to within 2-e. Journal of Computer and System Sciences, 74(3):335–349, 2008.*

78. *Dinur, I. and Safra., S.: On the hardness of approximating minimum vertex cover. Annals of Mathematics, 162(1):439–485, 2005.*

79. *Impagliazzo, R., Paturi, R., and Zane, F.: Which problems have strongly exponential complexity? Journal of Computer and System Sciences, 63(4):512–530, 2001.*

80. *Cygan, M., Fomin, F. V., Kowalik, Ł., Lokshtanov, D., Marx, D., Pilipczuk, M., Pilipczuk, M., and Saurabh, S.: Parameterized algorithms, volume 3. Springer, 2015.*

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

82. *Chatterjee, T., DasGupta, B., Mobasheri, N., Srinivasan, V., and Yero, I. G.: On the computational complexities of three privacy measures for large networks under active attack. [cs.CC], 2016.*

83. *Zhang, C. and Gao, Y.: On the complexity of k-metric antidimension problem and the size of k-antiresolving sets in random graphs. In COCOON 2017, LNCS, eds. Y. Cao and J. Chen, pages 555–567. Springer, 10392, 2017.*

84. *Backstrom, L., Dwork, C., and Kleinberg, J.: Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In Proceedings of the 16th international conference on World Wide Web, pages 181–190. ACM, 2007.*

85. *Mauw, S., Trujillo-Rasua, R., and Xuan, B.: Counteracting active attacks in social network graphs. In Proceedings of the 30th IFIP Annual Conference on Data and Applications Security and Privacy, pages 233–248. 9766, 2017.*

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

87. *Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C.: Introduction to algorithms. The, MIT Press, 2001.*

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

89. *Albert, R., DasGupta, B., and Mobasheri, N.: Topological implications of negative curvature for biological and social networks. Physical Review E, 89:032811, 2014.*

90. *Newman, M. E. J.: The structure and function of complex networks. SIAM Review, 45:167–256, 2003.*

91. *Holme, P., Kim, B. J., Yoon, C. N., and Han, S. K.: Attack vulnerability of complex networks. Physical Review E, 65:056109, 2002.*

92. *Sachs, A.: Completeness interconnectedness and distribution of interbank exposures - a parameterized analysis of the stability of financial networks. Quantitative Finance, 14(9):1677–1692, 2014.*

93. *Gai, P. and Kapadia, S.: Contagion in financial networks. In Proc. R, pages 2401–2423, 466(2120, 2010. Soc. A.*

94. *Markose, S., Giansante, S., Gatkowski, M., and Shaghaghi, A. R.: Too interconnected to fail: financial contagion and systemic risk in network model of cds and other credit enhancement obligations of us banks. Technical report, Economics Discussion Papers, Department of Economics, University of Essex, 683, 2009.*

95. *Callaway, D. S., Newman, M. E. J., Strogatz, S. H., and Watts, D. J.: Network robustness and fragility: percolation on random graphs. Physical Review Letters, 85:5468–5471, 2000.*

96. *Amini, H., Cont, R., and Minca, A.: Resilience to contagion in financial networks. Mathematical Finance, 26(2):329–365, 2016.*

97. *Cont, R., Moussa, A., and Santos, E. B.: Network structure and systemic risk in banking systems. In Handbook on Systemic Risk, Cambridge, eds. J. Fouque and J. Langsam, pages 327–368. University Press, 2013.*

98. *Wagner, A.: Estimating coarse gene network structure from large-scale gene perturbation data. Genome Research, 12:309–315, 2002.*

99. *Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F., and Arenas, A.: Self-similar community structure in a network of human interactions. Physical Review E, 68:065103, 2003.*

100. *Enron email network, available from uc berkeley enron email analysis website.*

101. *Paranjape, A., Benson, A. R., and Leskovec, J.: Motifs in temporal networks. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017.*

102. *Panzarasa, P., Opsahl, T., and Carley, K. M.: Patterns and dynamics of users' behavior and interaction: network analysis of an online community. Journal of the American Society for Information Science and Technology, 60(5):911–932, 2009.*

103. *Hamsterster friendships network dataset.*

104. *Demaine, E. D., Reidl, F., Rossmanith, P., Villaamil, F. S., Sikdar, S., and Sullivan, B. D.: . Structural Sparsity of Complex Networks: Bounded Expansion in Random Models and Real-World Graphs, 1406:2587, 2014.*

105. *Maiya, A. S. and Berger-Wolf, T. Y.: Expansion and search in networks. In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 239–248. ACM, 2010.*

106. *Achiam, Y., Yahav, I., and Schwartz, D. G.: Why not scale free? simulating company ego networks on twitter. In Conference on Advances in Social Networks Analysis and Mining, San, ed. I. International, pages 174–177, CA, 2016. Francisco.*

107. *B. DasGupta, B. and Mobasheri, N.: On optimal approximability results for computing the strong metric dimension. Discrete Applied Mathematics, 221:18–24, 2017.*

108. *Gast, M., Hauptmann, M., and Karpinski, M.: Inapproximability of dominating set on power law graphs. Theoretical Computer Science, 562:436–452, 2015.*

109. *Hauptmann, M., Schmied, R., and Viehmann, C.: Approximation complexity of metric dimension problem. Journal of Discrete Algorithms, 14:214–222, 2012.*

110. *Khanin, R. and Wit, E.: How scale-free are biological networks. Journal of Computational Biology, 13(3):810–818, 2006.*

111. *Stumpf, M. P. H., Wiuf, C., and May, R. M.: Subnets of scale-free networks are not scale-free: Sampling properties of networks. Proceedings of the National Academy of Sciences, 102(12):4221–4224, 2005.*

112. *Zito, M.: Greedy algorithms for minimisation problems in random regular graphs. In Proc. 9th Annuual European Symposium on Algorithms, pages 525–536, 2001.*

VITA

Name *Nasim Mobasheri*

Education ***B.Sc.**, Electrical Engineering*
Sharif University of Technology, Tehran, Iran, 2012
***Ph.D.**, Computer Science*
University of Illinois at Chicago, Chicago, Illinois, United States, 2018

Experience ***Research/Teaching Assistant**, University of Illinois at Chicago, 2012-2018*
***Teaching Assistant**, Sharif University of Technology, 2011-2012*

Publications *[As per custom in my research area, authors appear alphabetically in order of last name]*

Reka Albert, Bhaskar Dasgupta and **Nasim Mobasheri**, Some perspectives on network modeling in therapeutic target prediction, *Biomedical Engineering and Computational Biology, 5, 17-24 (2013).*

Reka Albert, Bhaskar Dasgupta and **Nasim Mobasheri,** Topological implications of negative curvature for biological and social networks, *Physical Review E, 89(3), 032811(2014).*

Bhaskar Dasgupta and **Nasim Mobasheri**, On optimal approximability results for computing the strong metric dimension, *Discrete Applied Mathematics, 221, 18-24 (2017).*

Bhaskar Dasgupta, Marek Karpinski, **Nasim Mobasheri** *and Farzaneh Yahyanejad,* Node Expansions and Cuts in Gromov-hyperbolic Graphs, *Algorithmica, in press.*

Tanima Chatterjee, Bhaskar DasGupta, **Nasim Mobasheri**, *Venkatkumar Srinivasan and Ismael G. Yero,* On the Computational Complexities of Three Privacy Measures for Large Networks Under Active Attack,*(2016).*

Bhaskar Dasgupta, **Nasim Mobasheri** *and Ismael G. Yero,* On analyzing and evaluating privacy measures for social networks under active attack, *(2017).*

Bhaskar Dasgupta and **Nasim Mobasheri**, optimal approximability results for computing the strong metric dimension, *in 8th Slovenian International Conference on Graph Theory, Kranjska Gora, Slovenia, (2015).*

American Physical Society Reuse and Permissions License

09-Jul-2018

This license agreement between the American Physical Society ("APS") and Nasim Mobasheri ("You") consists of your license details and the terms and conditions provided by the American Physical Society and SciPris.

Licensed Content Information

License Number:	**RNP/18/JUL/005927**
License date:	09-Jul-2018
DOI:	10.1103/PhysRevE.89.032811
Title:	Topological implications of negative curvature for biological and social networks
Author:	Réka Albert, Bhaskar DasGupta, and Nasim Mobasheri
Publication:	Physical Review E
Publisher:	American Physical Society
Cost:	USD $ 0.00

Request Details

Does your reuse require significant modifications: No

Specify intended distribution locations:	United States
Reuse Category:	Reuse in a thesis/dissertation
Requestor Type:	Author of requested content
Items for Reuse:	Whole Article
Format for Reuse:	Electronic

Information about New Publication:

University/Publisher:	University of Illinois at Chicago
Title of dissertation/thesis:	Geodesic-based Properties in Complex Networks
Author(s):	Nasim Mobasheri
Expected completion date:	Jul. 2018

License Requestor Information

Name:	Nasim Mobasheri
Affiliation:	Individual
Email Id:	nmobas2@uic.edu
Country:	United States

American Physical Society
Reuse and Permissions License

TERMS AND CONDITIONS

The American Physical Society (APS) is pleased to grant the Requestor of this license a non-exclusive, non-transferable permission, limited to Electronic format, provided all criteria outlined below are followed.

1. You must also obtain permission from at least one of the lead authors for each separate work, if you haven't done so already. The author's name and affiliation can be found on the first page of the published Article.
2. For electronic format permissions, Requestor agrees to provide a hyperlink from the reprinted APS material using the source material's DOI on the web page where the work appears. The hyperlink should use the standard DOI resolution URL, http://dx.doi.org/{DOI}. The hyperlink may be embedded in the copyright credit line.
3. For print format permissions, Requestor agrees to print the required copyright credit line on the first page where the material appears: "Reprinted (abstract/excerpt/figure) with permission from [(FULL REFERENCE CITATION) as follows: Author's Names, APS Journal Title, Volume Number, Page Number and Year of Publication.] Copyright (YEAR) by the American Physical Society."
4. Permission granted in this license is for a one-time use and does not include permission for any future editions, updates, databases, formats or other matters. Permission must be sought for any additional use.
5. Use of the material does not and must not imply any endorsement by APS.
6. APS does not imply, purport or intend to grant permission to reuse materials to which it does not hold copyright. It is the requestor's sole responsibility to ensure the licensed material is original to APS and does not contain the copyright of another entity, and that the copyright notice of the figure, photograph, cover or table does not indicate it was reprinted by APS with permission from another source.
7. The permission granted herein is personal to the Requestor for the use specified and is not transferable or assignable without express written permission of APS. This license may not be amended except in writing by APS.
8. You may not alter, edit or modify the material in any manner.
9. You may translate the materials only when translation rights have been granted.
10. APS is not responsible for any errors or omissions due to translation.
11. You may not use the material for promotional, sales, advertising or marketing purposes.
12. The foregoing license shall not take effect unless and until APS or its agent, Aptara, receives payment in full in accordance with Aptara Billing and Payment Terms and Conditions, which are incorporated herein by reference.
13. Should the terms of this license be violated at any time, APS or Aptara may revoke the license with no refund to you and seek relief to the fullest extent of the laws of the USA. Official written notice will be made using the contact information provided with the permission request. Failure to receive such notice will not nullify revocation of the permission.
14. APS reserves all rights not specifically granted herein.
15. This document, including the Aptara Billing and Payment Terms and Conditions, shall be the entire agreement between the parties relating to the subject matter hereof.

AUTHOR AND USER RIGHTS

INTRODUCTION

Elsevier requests transfers of copyright, or in some cases exclusive rights, from its journal authors in order to ensure that we have the rights necessary for the proper administration of electronic rights and online dissemination of journal articles, authors and their employers retain (or are granted/transferred back) significant scholarly rights in their work. We take seriously our responsibility as the steward of the online record to ensure the integrity of scholarly works and the sustainability of journal business models, and we actively monitor and pursue unauthorized and unsubscribed uses and re-distribution (for subscription models).

In addition to authors' scholarly rights, anyone who is affiliated with an institution with a journal subscription can use articles from subscribed content under the terms of their institution's license, while there are a number of other ways in which anyone (whether or not an author or subscriber) can make use of content published by Elsevier, which is free at the point of use or accessed under license.

Author Rights

As a journal author, you have rights for a large range of uses of your article, including use by your employing institute or company. These rights can be exercised without the need to obtain specific permission.

How authors can use their own journal articles

Authors publishing in Elsevier journals have wide rights to use their works for teaching and scholarly purposes without needing to seek permission.

Table of Author's Rights

	Preprint version (with a few exceptions- see below *)	Accepted Author Manuscript	Published Journal Articles
Use for classroom teaching by author or author's institution and presentation at a meeting or conference and distributing copies to attendees	Yes	Yes	Yes
Use for internal training by author's company	Yes	Yes	Yes
Distribution to colleagues for their research use	Yes	Yes	Yes
Use in a subsequent compilation of the author's works	Yes	Yes	Yes
Inclusion in a thesis or dissertation	Yes	Yes	Yes
Reuse of portions or extracts from the article in other works	Yes	Yes with full acknowledgement of final article	Yes with full acknowledgement of final article
Preparation of derivative works (other than for commercial purposes)	Yes	Yes with full acknowledgement of final article	Yes with full acknowledgement of final article
Preprint servers	Yes	Yes with the specific written permission of Elsevier	No
Voluntary posting on open web sites operated by author or author's institution for scholarly purposes	Yes (author may later add an appropriate bibliographic citation, indicating subsequent publication by Elsevier and journal title)	Yes, with appropriate bibliographic citation and a link to the article once published	Only with the specific written permission of Elsevier
Mandated deposit or deposit in or posting to subject-oriented or centralized repositories	Yes under specific agreement between Elsevier and the repository	Yes under specific agreement between Elsevier and the repository**	Yes under specific agreement between Elsevier and the repository
Use or posting for commercial gain or to substitute for services provided directly by journal	Only with the specific written permission of Elsevier	Only with the specific written permission of Elsevier	Only with the specific written permission of Elsevier

** Voluntary posting of Accepted Author Manuscripts in the arXiv subject repository is permitted.