C02029: Doctor of Philosophy
March 2023

Privacy diffusion in online social media reconstruction, modelling and blocking

Xiangyu Hu

School of Computer Science
Faculty of Engg. & IT
University of Technology Sydney
NSW - 2007, Australia

Privacy diffusion in online social media reconstruction, modelling and blocking

by

Xiangyu Hu

Thesis submitted in fulfilment of the requirements
for the degree of

Doctor of Philosophy

under the supervision of

Tianqing Zhu

to

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

March 2023

Certificate of Original Authorship

I, ***Xiangyu Hu***, declare that this thesis is submitted in fulfilment of the requirements for the award of *Doctor of Philosophy*, in the *School of Computer Science Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note: Signature removed prior to publication.

Date: 2023/03/11

Abstract

Social media has become a ubiquitous tool for spreading news and messages. It enables communication between individuals, and is a convenient platform for people to connect, interact, communicate or share information with others on the Internet. However, as users share their personal information, privacy information can also be revealed, which can spread through the network, making it crucial to study how private information propagates across social media. Many studies have used information diffusion models to examine how information flows through social networks. However, these models are theoretical and may not behave in the same way as private information. This raises questions about the observed phenomena and differences between privacy information and normal news diffusion. To tackle these challenges, we identified four major research problems: 1) Delineating private information from the clutter of other information on social media, 2) Identifying the propagation paths of private information, 3) Determining the features of the information or the diffusion process that inform adequate protection mechanisms and 4) Modeling private information diffusion through online social media.

To address these issues, we collected spreading information on online social media to construct graph structures that show the propagation path of different information. We found that privacy information differs in propagation features and the size of star structures. We proposed a new information diffusion model that considers the probability of users receiving, forwarding, and holding interest in a message. Lastly, we designed two block mechanisms to congest the diffusion of privacy information in social media. The main contributions of this thesis are a new information diffusion model, insights into privacy information propagation features, and block mechanisms to protect privacy information in social media. The main innovations and contributions of this thesis are as follows: 1)We introduce a new model that simulates the auction process with preserving users' privacy in online social network. 2) We use the data from Twitter API to construct graph structures that depict the propagation path of different types of information in social media. 3) We discuss the problem of modeling privacy information propagation, together with the propagation features and parameters of privacy information. 4)We propose a novel mechanism to stop privacy diffusion in online social media and prevent the privacy of social media users from leakage. 5) We carry out a new privacy diffusion-blocking methods. We try to block privacy data diffusion by limiting the connection between users in online social networks.

DEDICATION

I dedicate my Thesis work to my friends, my supervisors and my family. A special feeling of gratitude to my loving parents whose words of encouragement and push for tenacity ring in my ears.

I also dedicate this dissertation to my many friends and UTS who have supported me throughout the process. I will always appreciate all they have done for helping me develop my technology skills, Lefeng Zhang for the many hours of theory analysis, and Sheng Shen for helping me have a good experience in Australia's life.

I dedicate this work and give special thanks to my supervisors wanlei Zhou and Tianqing Zhu for being there for me throughout the entire doctorate program. Both of them give a wonderful direction in the research area and help me a lot in complete my research.

ACKNOWLEDGMENTS

This project would not have been possible without the support of many people. Many thanks to my supervisors, Wanlei Zhou, who read my numerous revisions and helped make some sense of the confusion. Also gives many thanks to my co-supervisor, Tianqing Zhu, who help me to complete the thesis and give directions in researcher area. I want thanks for Dayong Ye, and Bo Liu, who offered guidance and support.

Thanks to the University of Technology, Sydney for awarding me a Dissertation Completion Fellowship, providing me with the financial means to complete this project. And finally, thanks to my parents, and numerous friends who endured this long process with me, always offering support and love.

LIST OF PUBLICATIONS

RELATED TO THE THESIS :

- Xiangyu. Hu, Tianqing. Zhu, Xuemeng. Zhai, Wanlei. Zhou and Wei. Zhao, "Privacy Data Propagation and Preservation in Social Media: a Real-world Case Study", in IEEE Transactions on Knowledge and Data Engineering, 2021. (Early acess)
- Xiangyu. Hu, Tianqing. Zhu, Xuemeng. Zhai, Wanlei. Zhou and Wei. Zhao, "Privacy Data Diffusion Modeling and Preserving in Online Social Network" in IEEE Transactions on Knowledge and Data Engineering, 2022. (Early acess)
- Xiangyu Hu, Dayong Ye, Tianqing Zhu, Angle Huo. A Differentially Private Auction Mechanism in Online Social Networks[J]. Journal of Systems Science and Systems Engineering, 2021, 30(4):14.
- Xiangyu Hu, Zhiping Jin, Lefeng Zhang, Andy Zhou, Ye, Dayong. Privacy preservation auction in a dynamic social network[J]. Concurrency Computat Pract Exper. 2020; e6058. https://doi.org/10.1002/cpe.6058

OTHERS :

- Tianqing. Zhu, Jin. Li, Xiangyu. Hu, Pin. Xiong and Wanlei. Zhou, "The Dynamic Privacy-Preserving Mechanisms for Online Dynamic Social Networks," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 6, pp. 2962-2974, 1 June 2022, doi: 10.1109/TKDE.2020.3015835.
- Xiangyu. Hu, Tianqing. Zhu, Xuemeng. Zhai, Wanlei. Zhou and Wei. Zhao, "Sparse representation for privacy data blocking: a graph edge based mechanism" (Ready to submit)

Table of Contents

List of Figures

Figure **Page**

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Social network, as its name implies, is a network formed by social activities on the Internet, which mainly comes from social networking. E-mail is the starting point of social networking. Early e-mail solved the problem of remote mail transmission. So far, it is also the most popular application on the Internet and the starting point of social networking. BBS has pushed social networking forward, from the simple point-to-point communication cost reduction to the point-to-point communication cost reduction. Instant messaging (IM) and blog (blog) are more like the upgraded versions of the previous two social tools. The former improves the instant effect (transmission speed) and simultaneous communication ability (parallel processing). The latter began to reflect the theories of sociology and Psychology - the information release node began to reflect a more vigorous individual consciousness because the scattered information in the time dimension began to be aggregated and then became the "image" and "character" of the information release node. With the quiet evolution of social networking, a person's image on the network tends to be complete. At this time, social networks appear.

The social media sites such as Twitter and Instagram have significantly changed how people communicate. However, as users share their personal information, private information can be revealed, even if unintentionally, which can spread through the network. For example, forgetting to turn off location sharing during a visit to a medical specialist may reveal private information about one's health that could be published or sold. Private information does not need to be sensitive; it can be any information about an individual published without the person's permission or awareness. Notably, such leaks present significant risks to social media users. For this reason, everyone from the individual concerned to social media managers to data collectors and curators is interested in taking measures to stop the propagation of private information in social media. In this thesis, we try to collect diffusion data from Twitter and reconstruct the information propagation process. We then analyze the basic structure of the paths and discover their topological features. Then we try to model how the privacy information diffuses through social media and the block mechanisms.

1.1 Social Network Sites

Online social media platforms such as Instagram, Twitter, and Whatsapp significantly disseminate real-time information. In addition, these social platforms provide real-time 'sensors' for social trends and incidents. A recent report has proved that most events that happen in the real world come up and spread firstly through these social media platforms rather than traditional media (e.g., newspapers or televisions).

Online social network has reshaped the communication and interaction habits of human beings in nowadays life. They express their views and broadcast news and information about ongoing events they witness or experience. Their expressions, discussions, and comments contain diverse topics, including politics, the economy, and entertainment. Conversely, Online social network has even affected the events happening worldwide. For example, millions turned to online social networks to update and follow significant events during political and social upheavals. As radical words rapidly and widely spread on Twitter, the response to oppressive regimes and low living standards in Tunisia eventually became a nationwide revolution known as the “Arab Spring". Moreover, the presidential election always creates a big buzz on social media. The heated discussions of Hillary Clinton's Email Controversy caused the most severe impediment to her election campaign. Moreover, Japan's nuclear crisis in 2011 caused a run on salt in China of a rumor that the iodine in salt can prevent the human body from radiation-related illnesses.

1.2 Auction mechanism in social media

Auctions in social media refer to the use of online platforms such as Facebook, Instagram, or Twitter to conduct auctions. Social media auctions are becoming increasingly popular as they offer an easy and convenient way to reach a large audience and facilitate bidding.

In social media auctions, sellers can post pictures or videos of the item for sale, along with a description and starting bid. Interested buyers can then place their bids in the comments section or by direct message. Social media auctions are often used for charity fundraisers, where individuals or organizations can auction off items or services to raise money for a cause. However, it is important to note that social media auctions may not always follow the same rules and regulations as traditional auctions, and buyers should exercise caution before participating.

Social media auctions have several advantages over traditional auctions. One of the main advantages is that they can reach a large audience quickly and easily. Social media platforms have millions of active users, which means that auctions can attract more potential buyers. Additionally, social media auctions are often less formal than traditional auctions, which can make them more accessible and less intimidating for buyers who may not be familiar with the auction process. Another advantage of social media auctions is that they can be conducted at any time and from any location. Traditional auctions are typically held in a specific location at a specific time, which can be inconvenient for buyers who live far away or have other commitments. Social media auctions can be conducted from anywhere, which means that buyers can participate from the comfort of their own homes.

However, social media auctions also have some potential drawbacks. Since they are not regulated in the same way as traditional auctions, there is a risk of fraud or misrepresentation. Buyers should be careful when participating in social media auctions, and should only bid on items from reputable sellers. Additionally, social media auctions may not have the same level of transparency as traditional auctions, which can make it difficult for buyers to determine the true value of an item. Therefore, design a reasonable auction mechanism for social media is a challenge in auction theory. Auction mechanism design is an interface that combines economics and artificial intelligence. It employs game theoretic tools to simulate agent interactions and explores the impact of institutional design on agent outcomes. The role of social networks in auctions has not been fully explored, as potential buyers who are not directly connected to the seller may be unaware of the auction without an effective mechanism for information propagation. This can result in a high-value buyer being excluded from the auction due to inadequate communication among agents. Meanwhile, as the auction information diffuses through the social media, the privacy leakage troubles the users in auction. Therefore, how to design a reliable auction mechanism which satisfy the individual rationality while preserving users' privacy is a challange in this research area.

1.3 Information cascade

Information cascade is a phenomenon described in behavioral economics and network theory, in which many people make the same decisions sequentially. It is similar to herding behavior but different from herding behavior. Information level connectivity is often considered a two-step process. First, to begin cascading, individuals must encounter scenarios with decisions, usually binary decisions. Second, external factors can influence this decision (usually by observing the actions of other individuals in similar situations and their results).

It has been indicated that information flows in a cascade on social networks. According to the authors, viral analysis of information cascades on social networks may lead to many valuable applications, such as identifying the most influential individuals in the network. This information can be used to maximize market efficiency or influence public opinion. Various structural and temporal characteristics of the network affect cascade virility. In addition, these models are widely applied to rumor propagation in social networks to investigate rumors and reduce their impact on online social networks.

Contrary to the research on the information cascade in social networks, the social impact model of belief communication believes that people have some concepts about their personal beliefs in their networks. Therefore, the social impact model relaxes the hypothesis of an information cascade; people only act on observable actions taken by others. In addition, the social impact model focuses on embedding people into social networks rather than queuing. Finally, the social impact model relaxes the assumption of the information cascade model, that is, people either complete an action or do not complete it, because it takes into account the sustained scale of the "intensity" that agents believe an action should complete.

Information cascades can also reorganize the social networks through which they pass. For example, although social relations on Twitter have maintained a low loss in any month, about 9% of social relations have changed. After the information cascade (such as sharing viral tweets), there is usually a peak in attention and non-attention activities. As the level of Twitter sharing is connected to the network, users will adjust their social relations, especially those who have contact with the original author of viral Twitter. For instance, the author of viral Twitter will see a sudden decrease in previous followers and a sudden increase in new followers.

As part of this cascade-driven reorganization process, the information cascade can also create a variety of social networks in which people often associate with others with similar characteristics. Twitter cascades increase the similarity between connected users because users lose contact with more different users and add new contacts to similar users. The information cascade generated by media news reports may also contribute to political polarization by classifying social networks according to political lines: Twitter users who follow and share polarized news reports often lose social contact with users with opposite ideologies.

1.4 Information diffusion in social media

The vigorous development of SNS (Social Network Service) has shifted the way humans use the Internet from simple information retrieval to the construction and maintenance of online social relationships, as well as the creation, communication, and sharing of online information. SNS seeps into every aspect of social life and profoundly impacts it. According to the statistical report published by official institutes, the number of SNS users in China reached 762 million in 2020, accounting for 95.6% of Internet users. Facebook's monthly active users hit 2 billion, doubling in less than five years. Millions of users make SNS powerful in many fields (e.g., business marketing and social governance). However, the risks of social networks must be addressed. Scholars have proven that fake news spreads faster and more comprehensively than real news on social networks. Suppose malicious people utilize social networks to spread harmful information, such as terrorism and rumors. In that case, it will cause a massive menace to social stability, such as the Chinese salt scramble in 2011, the Arab Spring revolution, and the manipulation of political elections. Facing the opportunities and challenges brought by SNS, many scientists are devoted to studying information diffusion in SNS, and information diffusion modeling is the basis of these studies. Information diffusion models aim at capturing the dynamics of information diffusing in social work. Therefore, it is significant to model the information diffusion process in academia and practice. Academically, diffusion modeling involves multiple disciplines (e.g., statistics, complex networks, sociology, and machine learning). Moreover, diffusion-related data itself is easily accessible, which in turn promotes the development of related disciplines. Practically, they are the basis of many downstream applications such as popularity prediction, influence maximization (IM), source identification, network inference, and social recommendation. Employing these applications in various social tasks (e.g., marketing, rumor source identification, and trending topics detection) makes society run more efficiently. The influence in social networks is diffused from user to user, which can be initiated by a set of seed (initial) users. The notable study of influence diffusion can be traced back to Kempe's work, where the influence maximization (IM) problem was formulated as a monotone submodular maximization problem. The main purpose of this work is to find a subset of users as the seed set that maximizes the follow-up adoptions (influence spread). They proposed two diffusion models that were accepted widely in subsequent research, called the Independent Cascade model (IC model) and the Linear Threshold model (LT model). Besides, they proved IM is NP-hard and implemented the Greedy algorithm by Monte Carlo (MC) simulations. When opposite points of view, negative and positive information, from different cascades, are spread simultaneously on the same social network, users are more inclined to accept the information arriving at them first. Therefore, one solution to blocking rumor spread is to launch a positive cascade to compete with misinformation [3], [4]. Since the budget for positive seeds is limited, a classical rumor blocking (RB) problem is formulated, which spreads a positive cascade by selecting a positive seed set to prevent the spread of misinformation as much as possible.

Studying the diffusion of information can be divided into two aspects. The first aspect is to build a propagation model to simulate how the information propagates through social media [39, 57, 75]. In such aspect, some researchers built a diffusion model with similar topological structures to the real information diffusion in social media [63, 72, 114, 118]. On the other hand, some researchers focused on modeling different information diffusion in online social networks [57, 60, 72, 92]. In 2019, X. Wu *et al.* proposed an adaptive diffusion of the sensitive information model. In this thesis, they minimized the sensitive information diffusion while preserving the diffusion of non-sensitive information by modeling the sensitive information diffusion size as the reward of a bandit [112][30][21]. The second aspect in this area is to discover the real diffusion path of information in social media and to model them into a complex network, then find the difference of the topological structure from other kinds of complex networks [99, 122, 123, 125, 126]. The closest studies on this subject fall into two categories: sensitive information diffusion and propagation modeling. Studies on how sensitive information spreads typically focus on rumormongering [89, 91, 94, 127, 139]. And, most often, the solution to stopping their spread is to counter the rumor with the truth. However, while this might work to constrain hearsay, it is not a suitable remedy for privacy breaches. Private information may differ significantly from news or rumors [135]. First, the paths by which private information propagates might have a unique structure. For example, normal news is often published by traditional media outlets, which have many followers. At the same time, private data is likely to stem from users who do not exert a substantial impact on the platform in isolation. Second, not everyone is interested in a given individual's private information, so, unlike news or a celebrity rumor, it is probably wrong to assume that the information will spread throughout the entire network. How private information spreads through social media needs to be studied empirically and compared with what we know about the diffusion of other types of information. In this thesis, we followed the second aspect and improved the methods which rebuilt the propagation path of information in social media.

In the current study, researchers have shown that sensitive information differs from the normal news in social media [130, 140]. In 2018, S.Vosoughi *et al.* investigated the differential diffusion of verified true and false news in social media and proved that false news in social media is often more novel than real news in social media [103]. However, the study of the difference between privacy information and normal news is still a gap in this area. Analyzing sensitive information in online social networks is not new [59, 141]; however, the study of privacy information propagation in this area is very limit. T. Zhu [143] first proposed the dynamic privacy propagation model in social media. In this thesis, Dr. Zhu assumed that privacy information is different from other information, e.g., the normal news or the rumors on social media, because the privacy information is accumulated. Therefore Dr.Zhu creatively adopted an accumulation mechanism into the propagation model. In her work, the privacy information of one user in social media has a specific value. Part of this user's privacy information will be spread to other users on social media at once with a certain probability. Herefore, we can take a more profound analysis of the privacy propagation in social media with a specific definition of privacy information and a significant dataset. In addition, she proposed two categories of privacy preservation mechanisms, *i.e.*, the centralized and the decentralized mechanisms, to prevent propagation in social media. In this thesis, we use decentralized mechanisms to stop privacy information propagation in online social networks.

1.5 Information diffusion model

Although there are limited research outcomes related to privacy propagation in social media, the information diffusion in social media has been well researched [28, 49, 80, 84]. In real-world scenarios, information diffuses between the number of users through online social networks. Therefore, any complex network nodes could potentially serve as information disseminators in an information diffusion model. The users in the information diffusion model could be divided into four categories during the information propagation process: 1) Susceptible users, the users with no information but who are interested in receiving the information from other users, 2) Infected (I), the users with some specific information and who are ready to infect other users intentionally, 3) Inactive (A), the users with specific information but are not intentionally willing to distribute it and 4) Defence (D), the users holds none of the information, nor will it accept any. In a typical diffusion model [54, 55], all nodes are inactive at the beginning. Therefore, we transform some nodes into active statues at first, and these nodes will affect other nodes in the network. In the diffusion process, the inactive nodes will be transformed into an active statues by their active neighbor. The success of node u activating node v only depends on the propagation probability of information from u to v, and each edge has its propagation probability. When no other nodes are activated, the diffusing process will be terminated. Figure 1 is an example of this diffusion process model in social media. In the research area, information diffusion models obey the following rules in online social networks:

- **Information diffusion mechanism**: In online social networks, users will spread their general information to their nodes through the follower-followee relationship in an online social network. That is to say, users in social media will spread their general information to and only to their neighborhood users.

- **Diffusion probability:** Every user will spread their general information to their neighborhood users in a certain probability. This probability is determined by each user's interest in this information.

- **Users' statues:** To represent the state and the transition of different users during the information diffusion process, we divided the users into three statuses in the diffusion process: 1) Susceptible(S), the users who have no information and are interested in obtaining this information, 2) Infected(I), the users who have information and are ready to spread it, 3) Inactive(A), the users who do not want to participate in spreading the information.

With these three basic rules of information diffusion, researchers have built several information diffusion models over the past decade, such as SI, SIS models, etc. [106, 132].

Previous works have focused on modeling the diffusion of normal information or sensitive information (like rumors) [84–86, 97]. At the same time, it is hard to model the privacy diffusion in social media for the following reasons:

Firstly, previous studies have shown that the privacy information and normal news propagation path differ significantly in topological structures. However, it is hard to analyze the difference between them deeply.

Secondly, modeling the diffusion structures of different information in online social media is a foreseeable area. A novel information diffusion model should be discovered to estimate the extent of the new privacy diffusion.

Thirdly the third challenge is how to constrain the diffusion of privacy information. In this thesis, we formulate this challenge into a problem constraining the diffusion size of privacy information while users have a good user experience in social media.

Current information diffusion mechanisms, such as SI models, did not consider the diffusion features of privacy information in online social networks. Previous work on stopping sensitive information focused on limiting the diffusion ability of critical users, which can hardly be applied directly forward, stopping the privacy information diffusion.

1.6 Information diffusion blocking mechanism

The structure of the social network is defined as a complex network, in which the node represents the users and edges represent the connected relationships [50] As privacy diffusion is similar to rumor propagation, we can start by analyzing current rumor-blocking methods. Currently, the study of stopping illegitimate texts, which include rumors, fake news, and privacy information propagation in online social media, has mainly focused on the block of rumor information [136][14][38]. These methods often model the relationships of social media users into the complex network and simulate the blocking mechanism through the network. They could be divided into the following two categories:

The first is limiting the users' behaviors. These behaviors include commenting, sharing, and retweeting [14] [17]. The users limited in these methods are the users with strong influences in the network. These high influences presented in the complex networks are the nodes with high degree, high intermediate centrality, or high CI value, which represent the influence of a node in complex networks [16][58]. The second method is to delete the users from social media and block the diffusion paths. The third method is diffusing the truth news against the rumor. In recent years, some researchers believed that limiting the user's behaviors could not stop information diffusion in social media. Thus, they proposed several new methods to block the rumor information spreading [97]. By applying such a method, the rumors will collapse [98] [38].

However, these two types of methods are not suitable for stopping the privacy information propagation efficiently for the following reasons:

Firstly, the diffusion of privacy information is different from rumors. Users who diffuse the privacy information might be the normal users who do not strongly influence online social media [41]. Therefore, limiting the high-influence users is less efficient in blocking privacy information on social media.

Secondly, limiting the user's behaviors may violate the spirit of f́reedom of speech' which is the footstone of the modern online social network [98]. Meanwhile, limiting nodes in social media will drop many unnecessary connections, which may decrease the utility of social media.

Thirdly, unlike rumors or fake news, privacy information is accurate information published without the person's permission or awareness. As a result, diffusing the truth information against the privacy information has little impact on privacy diffusion.

The above challenges make privacy data blocking a tricky question in online social media. Moreover, blocking privacy diffusion differs from rumors as we must accurately catch privacy information from social media. Current research studies on blocking privacy diffusion mainly focused on limiting the important nodes [69]. This method will drop a large number of connection relationships when deleting the nodes from the network. Most dropped nodes in these methods are nodes with a high degree. The deleting action will drop the connections between these dropping nodes and their neighborhood. These dropping connections are primarily unnecessary and cause a lousy utility for users in social media.

In this thesis, to study privacy efficiently, we carry out the following research issues: 1) We first analyze how to design an auction mechanism while preserving the users' privacy in in social media 2) we then apply the diffusion of privacy in auction to analyze the privacy diffusion in the whole social media. To achieve this we need to collect the data on privacy diffusion in social media and identify the features of privacy propagation. 3) we propose a novel information diffusion model to simulate the privacy information diffusion in social media, and 4) we give two mechanisms to block the privacy information diffusion in social media. The main contributions are shown as the following:

- We propose a privacy-preserving auction mechanism in online social networks. To address the issue of how auction information propagates in social networks, we first build a new model to simulate the process in online social network auctions. Considering the time delay in social networks (agents in social networks often need time to react), we introduce time intervals into our auction model. Privacy preserving auctions have been studied in centralised environments but have not been studied in online social networks which are decentralised environments. Our privacy preserving auction mechanism is based on differential privacy theory. Differential privacy is a promising privacy model which has been successfully applied to many real-world applications. Differential privacy offers a rigorous guarantee that the analytical outputs of two datasets are almost the same if the two datasets differ by at most one data record. By using differential privacy mechanisms on a social network, the seller's valuation information can be protected. Moreover, differential privacy can also guarantee that the values of users' bids can be hidden while the selection of the winner of an auction is not affected. This is because selecting a winner in an auction can be interpreted as querying an interesting record in a dataset. Also, an auction with differentially private modified bids can be interpreted as a neighbouring auction of the original auction

- We take a first look at this problem and, in doing so, identify three significant challenges with studying real-world cases. 1) How does one delineate private information from the clutter of other information on social media? Topological features may be a good starting point for this. 2) How can one identify the propagation paths of private information? With no previous research to rely on, this is green field exploration. However, with the private information identified, we can trace its flows and see whether there are unique patterns to its diffusion. 3) Once identified, what are the features of the information or the information's diffusion process that will help to inform adequate protection mechanisms? This last question is the ultimate goal in the pursuit of privacy preservation. To tackle these challenges, we began by collecting datasets from Twitter API. We collected several spreading information in online social media and used them to construct the spreading process of the information in social media into the graph structures. These graph structures show the propagation path of different information in social media. We found two key differences in the features and flows of normal news and private data. First, normal information often diffuses fastest at the first hop of the propagation path, while private information diffuses fastest at the second or third hop. Second, normal news emanates from highly impactful nodes; private information does not have a prominent center of origin.

- We propose a new approach to model and analyze the privacy information diffused through the online social network. This model considers three key parameters: the influence of neighborhood users in online social media, the interest of different users, and the time of information publishing. We propose a novel mechanism to stop the propagation of privacy information in online social media and prevent the privacy of social media users from leakage. This new information diffusion model is illustrated to simulate the diffusion process of information in social media by considering the following three parameters: 1) The probability of users receiving this message, 2) The probability that users tend to forward this message, 3) The interest the users hold for this message We propose a novel mechanism to stop the propagation of privacy information in online social media and prevent the privacy of social media users from leakage. Our mechanism can effectively restrict the privacy propagation in our diffusion model. Substantial experiments have been conducted on real data from our collection and other public sources. The results of our experiment are evaluated in terms of the affection and business performance of users' experience. It shows that our mechanism can constrain the privacy propagation in social media and provide an improved experience.

- We block the privacy data diffusion by limiting the connection between users in the online social network. Compared with limiting nodes, this method has the following advantages: 1) it will not limit users' behavior in social media which provides a more comfortable experience, 2) it will limit only some meaningful connection relationships in social media, which means that we will not drop the unnecessary connections in the nodes limiting method. The current study has yet to give the parameter to evaluate the importance of edges in a complex network. Finding these important nodes in the network is the critical problem in this thesis. Considering the difficulties and complexities in solving such a problem, in this thesis, we will first deeply analyze the topological structure in the complex network and discover the important edges in the network by using a newly proposed complex network analysis approach. This method will decompose the complex network into several atoms according to each node in the network. We will use these decomposed atoms to discover the important edges in social media. Our mechanism can effectively restrict privacy propagation in our diffusion model. Substantial experiments have been conducted on real data from our collection and other public sources. The results are evaluated regarding the affection and the business performance of the user's experience. It shows that our mechanism can constrain privacy propagation in social media and provide an improved experience.

The rest of this thesis are organised as the following. We first give some preliminary knowledge which used in the thesis in chapter 2. Then the main contributions of this thesis are divided into two parts. Part I introduce a new auction mechanism in social media to give a illustration of how the auction privacy information diffusion in online social media and design an algorithm to preserve users' privacy during the auction. This part contains chapter 3 and chapter 4. Part II tries to analyse how the privacy information diffusion in social media based on the illustration of auction privacy diffusion in social media and design a new diffusion model to simulate this diffusion and propose some mechanisms to block their diffusion in social media. This part contains chapter 5, chapter 6 and chapter 7. In Section 8, This paper concludes with a summary of the material covered and our intentions for future work.

CHAPTER 2

Preliminary

2.1 Privacy preserving in social media

Privacy preserving in social media refers to the protection of users' personal information and sensitive data when using social media platforms. With the widespread use of social media, users often share a large amount of personal information on these platforms, including their location, interests, and personal preferences. This information can be used by advertisers, marketers, and other entities to target users with personalized ads or other content.

To protect users' privacy in social media, several approaches have been proposed, including data encryption, differential privacy, and homomorphic encryption. These techniques aim to secure users' data while still allowing social media platforms to provide personalized content and services to users.

One approach to privacy-preserving in social media is through the use of privacy-preserving protocols for data sharing and communication. For example, secure multiparty computation (MPC) enables multiple parties to compute a function over their respective inputs while keeping their inputs private. MPC can be used for tasks such as private messaging or data sharing on social media platforms. Another approach is through the use of privacy-enhancing technologies (PETs), such as data anonymization and pseudonymization. These techniques aim to hide users' identities and prevent the disclosure of their personal information to unauthorized parties. For example, social media platforms can use techniques such as differential privacy to protect users' data while still providing personalized recommendations and content.

Overall, privacy preserving in social media is an important area of research that aims to protect users' personal information while still allowing them to benefit from the services and features provided by social media platforms.

2.2 Complex network

In the real world, many complex systems in nature could be represented by different kinds of topological graph. In the past half century, researchers have tried to use different methodologies to study and model these topological graphs, like regular graphs and random graphs [70, 71]. However, in recent years, researchers find that neither of these methodologies could present the real topological structures for the complexity of these graphs in real world complex system. Therefore, these kinds of topological graphs are called complex networks. The theory of complex network has been applied into many research area like the power grid system [104], biology system [6, 7, 102], economics [83] and social media [8]. Different kinds of complex networks have different topological structures. In recent years, researchers try to decompose the complex network into several small structures to find their difference [115][20][137] and apply them into the deep learning methods [81][108][77][90]. In social media, complex networks are often used to represent the users' relationship and we will use it to model our information propagation path in social media.

Scientists are concerned about how to describe the topological structure of this kind of network in the last one hundred years and the study of this problem has gone through three periods. At first, they used some regular structure network like a globally coupled network, nearest-neighbor coupled network and a star coupled network to represent their topological structure. In 1958, Erdos.P and Renyi.A proposed a new method to construct a new model of these networks[26]. Because the connection between two nodes in this model is determined by a probability, scientists call this model a random network. It was thought to be the best way to describe the topological structure of these networks until 1990s. Research found that most real networks are neither the regular structure network nor the random network for it has different statistical characteristics with these two kinds of networks. These networks were called the complex network.

Unfortunately, Scientists still have not yet given a precise definition of complex networks. It is called a complex network because of the following two reasons (1) it is the topology aggregation of a real complex system, (2) it seems to be more complicated than the regular structure network and random network because researchers could easily generate these two kinds of network but it is still very difficult to generate a network which satisfy the real statistical characteristics.

A large number of experimental researchers indicate that the complex network has a small-world phenomenon[107]. It means that these networks have small characteristic path lengths. In a real interpersonal relationship, it is known as six degrees of separation in that one person could reach out any other in the world within 6 people (and have some connection to that person). Meanwhile, the degree of the nodes in the complex network satisfy the power-law distribution and the number of nodes with certain degree could be approximated by a power law function[4]. This phenomenon is called the scale-free phenomenon.

Researchers have used several methods to describe the topological properties of complex network. These methods could be divided into two categories, one uses some basic statistical parameters like degree distribution, average path length and propagation depth and width to describe the basic topological properties of the network. The degree of a node is the edges the node have in complex network. This parameter shows the connection relationships of each node. Degree distribution shows the number of nodes in each different degree in a complex network. This metric shows the connection mode in node level in the complex network. The average shortest path is the average value of shortest distance between each pair of nodes in complex network. The depth of a node i is the number of hops from node i node to the root node. The width is a parameter reflecting the number of nodes which are in the same depth forming a cascade layer in the network. The propagation depth of a propagation path network is the maximum depth of the nodes from the original nodes in the cascade. The other one is to study the sub-structures of the network for the reason that the overall structure of the network is too large to give a clear descriptions and find their difference. These sub-structures are called high-order structures. In this thesis, we use two categories of complex networks to analyze the diffusion of different information in social media. To represent them, we use two categories of graph models in social media: the follower-followee graph and the information propagation graph.

In the research area of complex network, researchers often use a graph model $G = V, E$ to represent the relationships between each entities in complex network. In such graph models G, a set of nodes $V = \{v_1, v_2, ..., v_n\}$ is used to represent the users in the online social media, and the edges $E = \{e_{ij} = (v_i, v_j) | v_i, v_j \in V, i \neq j\}$ between each pair of nodes denote the connection between different users in online social media. Two nodes are called adjacent if they share a common edge.

2.3 Social Network Data Model

Social relationship plays a significant role in human society and its evolution reflects the development of human civilization. In early time, the social relationships were limited by regions for people can only use the face-to-face language and actions to express their ideas. The birth of writing languages and letters has promoted the transformation of social relations. Through these letters, people can communicate with other people across different regions, socialize with distant friends. However, it is still not inflexible since the inefficiency of letters in communications.

In last century, the social relationship meets a great breakthrough and development with the invention of Internet. It improves the efficiency of human communication greatly and has developed diversified social communication tools like Email, Blog and social media. These tools not only enrich people's social relations, carry more social information, but also solve the real time problem in social relations, so that people can receive messages from distant friends in real time and interact with them. In these communication tools, Twitter is a representative example. People can not only communicate with others but also can be used as a service platform for current news, content sharing, commodity recommendation and so on. The number of registered users in Twitter has exceeded over 500 million in 2012, which is a widely popular social tool at present. Therefore, using Twitter as the representative of social tools for research and analysis can excavate the social relations of contemporary human beings, and the research on human social relations has universal significance.

Twitter has many active users and a large number of high-quality tweets, so twitter contains a lot of valuable social relations and social content, which is worth exploring. As the initial source and final receiving unit of social content, Twitter account plays a vital role in the process of content production, dissemination and reception. The basic attribute information of users who use Twitter, including gender, age, geographical location, personal description can be used to mine users' social relations and social content through Twitter account. For example, tweets published through twitter accounts can form a basic judgment on users' personality characteristics, personal preferences and life attitudes. The social relationship of users can be studied and analyzed through the attention relationship, forwarding relationship, reference relationship and @ relationship of Twitter account.

Due to the large number of twitter accounts and the different social relationships between these accounts, twitter accounts may contain a variety of direct or indirect relationships. These social relationships form a huge social network, which can be abstracted into a large complex network, researchers could study twitter social networks through the theoretical knowledge of complex networks. In these complex networks, different twitter accounts may have different concerns and preferences. Some accounts may be very interested in sports events, and some accounts may focus on military news. Users usually become closer due to the common concerns and preferences. The categories of these accounts can be summarized by analyzing and abstracting these accounts with close social relationships. The account classification method based on complex network is to abstract the social relationship of accounts into complex network, and then classify unknown accounts through the theory and method of complex network.

Online users use the service provided by the social network sites to connect and interact with other users through personal relationships, interaction and the information provided by the social media service. Social media appliances collect data like this from users and this data is stored in the dataset. The service providers analyze this data and provide the corresponding services to different users.

The data from the SNSs is modeled to a complex network $G(V,E)$, where V is a set of nodes V = {v_1,v_2,...,v_n} which represent the users in the SNS, and E is a set of edges E = {e_{ij}=(v_i,v_j) v_i,v_j∈V} which represent the connection between different users in the network. These connections may vary from site to site. The network may be directed and undirected, it depends on the connection relationship.

Like other kinds of complex networks, the social network satisfies the small-world and scale-free phenomenon. However, it is different from other types of networks like technological and biological networks in the following aspects (1) the social network shows distinctly different patterns of correlation between the degrees of adjacent vertices, with degrees being positively correlated assortative mixing in most social networks and negatively correlated disassortative mixing in most nonsocial networks, (2) social networks in general have a far higher degree of clustering than the other network. This phenomenon suggests a possible candidate theory that social networks contain groups or a so-called community structure.

Social media devices collect social network operators and store data from service users in order to share data with a wide range of third-party consumers (such as researchers) to study disease transmission and risk. As the collected data usually contains sensitive information, network operators may release the risk of privacy disclosure when sharing a complete social network map or sub map with third-party users (such as advertisers, marketers and healthcare professionals). How to preserve these sensitive information is still a problem to be solved.

In social media, users usually obtain information from other users who they followed. Therefore, the information distributed along the follower-followee relationship. To simulate the diffusion process in social media, we first give a definition of follower-followee relationship network in social media.

Definition (Follower-followee relationship network). The nodes in the follower-followee network represent users. The edges are directed and represent one user's followership with the other users in social media [32].

Based on the follower-followee relationship network, the process of information diffusion happens through the directed edges between two nodes in the network. This diffusion process forms the information diffusion graph. The definition of information diffusion network are shown as the following:

Definition (Information diffusion network). The nodes in the information diffusion network represent users. The edges between node i and j are directed and represent the information are diffused from node i to j in social media [32].

The adjacency matrix is a kind of square matrix which used to represent the finite network. Each element of the matrix indicate whether the pairs of nodes are adjacent or not in the complex network. For example, an element $A_{ij} = 0$ indicates that there is no edges between node i and j while it show there is an edges when $A_{ij} \neq 0$. The diagonal elements of the matrix are all zero, Since most research on complex network only consider the topological structure of complex network. The adjacency matrix are normally binary in research area, *.i.e* the non-zero element are 1 in adjacency matrix.

Degree is the basic features of nodes in complex network. The degree of one single node i in complex network is the number of nodes which connected directly to node i. Average_degree is the average value of all nodes in complex network:

$$Average\_degree = \frac{\sum_{i=1}^{n} D_i}{n} \tag{2.1}$$

where D_i is the degree of each node in the graph, n is the number of nodes in the graph.

For an edges connected 2 nodes in network, the sum of all the nodes' degree is 2 times the number of edges. Therefore, the product of average degree and number of nodes is also 2 times the number of edge.

2.4 A case study of diffusion process mode in a graph

In real-world scenarios, information diffuses between the amount of users through the online social networks [62, 87, 116],. Therefore, any nodes of a complex network could potentially serve as information diffusion in a information diffusion model and the users in the information diffusion model could be divided into 4 categories during the information propagation process: 1) Susceptible users, the users with no information but who are interested in receiving the information from other users, 2) Infected (I), the users with some certain information and who are ready to intentionally infect other users, 3) Inactive (A), the users with certain information but are not intentionally willing to distribute it and 4) Defence (D), the users holds none of the information, nor will it accept any. In a typical diffusion model, all nodes are inactive at the beginning. Therefore, we transform some nodes into active statues at first and these nodes will affect other nodes in the network. In the diffusion process, the inactive nodes will be transformed into active statue by their active neighbor. The success of node u activating node v only depends on the propagation probability of information from u to v, and each edge has its own propagation probability. When no other nodes are activated, the diffusing process will be terminated. Figure 1 is an example of this diffusion process model in social media.

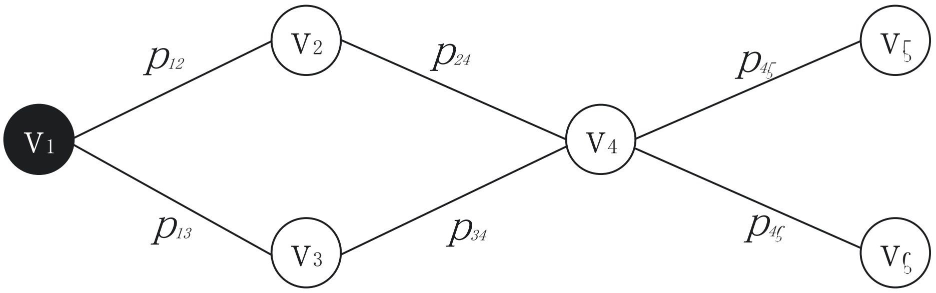


Figure 2.1: An example of the information diffusion model in social media.

Figure 2.1 is an example of the information diffusion model in social media. Among the nodes in the graph, the black nodes are the nodes with information (infected nodes) and the white nodes are the nodes who do not receive the information (susceptible nodes or inactive nodes). In this figure, v_i is the node in social media and p_{ij} is the diffusing probability between node $i.$ and $j.$ In the diffusion model, the information diffused through the social media among the edges between the nodes in social media. That is to say, the information will first diffuse from node v_1 to v_2 and v_3. Then nodes v_2 and v_3 will diffuse it to their neighborhood in the network. Meanwhile, different pairs of nodes have different diffusing probability, *.i.e* for any nodes i, j, u, v in the graph, p_{ij} $\neq$p$_{uv}$.

The information diffusion models have the following principles: 1) information can only spread from the active nodes to the inactive nodes, *.i.e*, users can only obtain and send the information from users who have forward it, 2) it is assumed that there are t time rounds of information propagation. After t rounds of time, the information propagation will be terminated and 3) Any inactive node that could be transformed into an active node. From then on, this active node has the opportunity to spread the information to its neighborhood users as long as the information it holds has not expired. Specifically, we will choose a set of nodes as the active node and the source nodes of information before the diffusion process. we define $p_{u,v}$ as the probability that nodes u successfully diffused information to nodes v in the follower-followee relationship network, then nodes v will spread the received information to its neighbors nodes. Moreover, in each round, each active source node has an opportunity to spread information to each of its neighbors, because in reality, social media users may repeatedly check their followees' messages which published a few days or even months ago.

The most widely known diffusion model is the SI model [11, 75, 109, 131, 134]. It is a classical diffusion model which was first used in the epidemic process. It is based on taking the continuous-time limit of difference equations for the evolution of the average number of individuals in each compartment. Usually, a differential equation is applied to describe the population over a certain time frame [114]. For the diffusion process is highly related to the topological structure of the graphs. It is highly used in information diffusion process.

2.5 Differential privacy

Differential privacy is a privacy model that aims to protect the privacy of individuals while allowing for statistical analysis of their data. It was first proposed by Cynthia Dwork in 2006 [24] as a way to address the challenges of privacy in the era of big data.

The basic idea of differential privacy is to add random noise to the data so that the statistical properties of the data remain the same while individual data points are protected. In other words, differential privacy guarantees that the output of a computation will not be affected significantly by the inclusion or exclusion of any single data point.

To achieve differential privacy, data is modified by adding random noise to the original data. The amount of noise added is determined by a parameter called the privacy budget, which represents the maximum amount of information that can be revealed about any individual in the data set. The higher the privacy budget, the less noise needs to be added, but the less privacy is protected.

Differential privacy has become increasingly important as more and more personal data is being collected and analyzed for various purposes, such as research, marketing, and public policy. It has been successfully applied to many real-world applications, including data mining, machine learning, and social network analysis.

Generally, differential privacy is a promising privacy preservation model. It theoretically ensures that the ability of an adversary to inflict harm to any individual in a dataset is essentially the same, independent of whether any individual opts in to, or opts out of the dataset which has been used in many research area, such as machine learning [144], feature selection [128], cyber physical systems [145] and social network [145]. Compared to previous privacy models, differential privacy can successfully resist background attacks and provide a provable privacy guarantee. Differential privacy is a privacy preserving model, which was originally developed for tabular data to give strong guarantees for preserving users' privacy without relying on the background knowledge, computing power or subsequent behavior of the opponent [113]. The definition of differential privacy is shown as follows [142]:

A randomised mechanism M gives (ϵ;δ)-differential privacy for every set of outputs Œ©, and for any neighbouring datasets of D and D$^{'}$, if mechanism M satisfies:

$$Pr[M(D)]\in\Omega[M(D^{'})\in\Omega]+\delta \tag{2.2}$$

If $\delta = 0$, the randomised mechanism M gives ϵ-differential privacy, which is the strictest definition. The (ϵ;δ)-differential privacy provides freedom to violate the strict ϵ-differential privacy for some low probability events. δ-differential privacy is usually called pure differential privacy, while (ϵ;δ)-differential privacy with $\delta \geq 0$ is called approximate differential privacy.

The reliability of Differential privacy is the form of query and the resulting disturbance. It helps preserve users' privacy information. One way to guarantee privacy in differential privacy is to apply random noise *.i.e* the Laplace distribution, the normal distribution and the variance depends into the output of the queries. This is achieved by using methods on $delta$and the sensitivity of the query. The global sensitivity of the query provides the maximum amount of noise that needs to be introduced into the real response to ensure privacy. However, some categories of queries in this area can produce considerable global sensitivity. These categories of requires the development of the smoothing sensitivity concept to reduce the amount of noise that must be introduced into the queries while still satisfy the differential privacy guarantee. Blocki textit et al. proposed the concept of restricted sensitivity to improve the accuracy of distinguishing private data analysis. Unlike the global sensitivity method, which quantifies all possible data sets, this limited sensitivity method quantifies a class of limited data sets.

2.6 Differential privacy model in OSN

The using of differential privacy in graphs data studies algorithms for computing accurate graph statistics while preserving differential privacy of individual nodes and edges [45]. Particularly, in social networks, the using of such method is to guarantee that an adversary is not able to obtain the information like whether a person is in the social network $G = (V, E)$ or whether two persons are connected in the social network $G = (V, E)$.

Formally, for any double neighboring graphs G and $G^{'}$ and the algorithm A and the set $S \subseteq Range(A)$ of possible outcomes form the algorithm A, it provides ϵ-differential private satisfies the following formula [1]:

$$Pr[A(G) \epsilon S] \leq exp(\epsilon) \star Pr[A(G') \epsilon S] \tag{2.3}$$

where ϵ is a value set to control the trade-off between privacy and accuracy and the probability Pr is computed over the random coin tosses of the algorithm.

Based on the differential privacy, the third-party entities could perform complex analysis on the graphn data without inferring the privacy information of a significant users. Therefore, various common graph data analysis, such as clustering analysis for identifying online communities and users impact analysis for identifying influential users in social media, have been applied to the social media data graph. These analysis include the number of edges, the number of sub graphs, centrality measurement and degree distribution. The number of triangles in one graph could also be applied by these methods.

In the network graph data, differential privacy try to achieve the guarantee that an adversary from the third-party users will not be able to obtain the information that an user x appears in a graph G or an user X and a person Y are connected to each other in the original graph G. The first kind of methods is called interactive setting. In such methods, a third-party users put forward queries to graph data providers. The providers will use ϵ- priority private algorithm a to response the quires. Algorithm A could modify the query or the response to preserve the privacy information of social media users. Given possible outputs of two graph data D and D$^{'}$, the two graph data have difference in only one record, but the output pf the responses are the largest due to this different record. The second method is called non interactive setting. The original data is not used to answer queries. The second kind of method are the non-interactive setting. The data from the database is not used to response to the queries. On the contrary, the graph data providers calculate and publish useful statistical data about the graph data in the form of summary or synthetic database under differential privacy. Published summary or composite databases are customized to answer specific types of queries. In such situation, the third party users can conduct deeply analysis, such as estimating the distance between social media users through the published graph data or using synthetic databases without being able to infer any private information about OSN users [45].

The semantic interpretation of differential privacy depends on the meaning of adjacent databases. Because differential privacy ensures that the output of the algorithm cannot be used to distinguish adjacent databases, it is the differences between neighboring databases that are preserved . In the definition of differential privacy, neighboring databases refer to the addition or deletion of a data record. In the hospital example, a patient's private information is encapsulated in a single record. Therefore, differential privacy ensures that the output of the algorithm does not reveal the patient's medical history.

The semantic interpretation of differential privacy rests on the definition of neighboring databases. Since differential privacy guarantees that the output of the algorithm cannot be used to distinguish between neighboring databases, what is being protected is precisely the difference between neighboring databases. In the definition of differential privacy, a neighboring database means that the addition or removal a single record of the data. In the hospital example, the patient's private information is encapsulated within a single record. So differential privacy ensures that the output of the algorithm does not disclose the patient's medical history.

In the network database, the relationship between different individuals may be the source of adjacent databases. However, the correspondence between private data and database records is not clear. In order to adapt to the differential privacy of graphs data, we must select a definition for neighboring graphs and understand the privacy semantics of the selection. In this article, we will describe two alternatives that provide different degrees of privacy preserving.

The input of a social network should be as a graph, G = (V,E), where V is a set of n entities and E is a set of edges. Edges are undirected pairs (u, v) such that u and v are members of V. (Results are easily extended to handle directed edges.) While the meaning of an edge depends on the domain, it could connote friendship, email exchange, sexual relations, we assume that it represents a sensitive relationship that should be kept private. The focus of the present work is concerned with graph structure, so the inclusion of attributes on nodes or edges is left for future work[96].

In recent years, many efforts have been made for the differential private release in the graph based dataset. Algorithms that satisfy the guarantees the need of differential privacy have been developed to publish statistics about the graph dataset. The existing differential privacy used for the preserving the graph data privacy could generally be divided into two categories: the node privacy and edge privacy.

In the following subsections, we will discuss the node level privacy and the edge level privacy categories in detail.

2.6.1 Edge level differential privacy

The first adaptation of differential privacy to graphs is mathematically similar to the definition for tables. As we have described in the previous character, in the definition of differential privacy, the most important thing is to find the neighboring database. In the graph database, we call them neighboring graphs. Neighboring graphs are defined as graphs that differ by one record. Given a graph G, one can produce a neighboring graph $G^{'}$ by either adding/removing an edge in E, or by adding/removing an isolated node in V. Restricting to isolated nodes ensures that the change to V does not require additional changes to E to make it consistent with V . Formally, G and $G^{'}$ are neighbors if $|V \oplus V^{'}|+|E \oplus E^{'}|= 1$[47].

From the previous definition, we could easily understand that the edge level differential privacy strictly satisfied the formula $|V \oplus V^{'}|+|E \oplus E^{'}|= 1$. However, the Node level differential privacy could not for the reason that when one node is removed from a graph, the edges should also be removed from it.

For the reason that the edge-level differential privacy is a weaker concept than the node-level differential privacy, it has been studied more extensively. Intuitively, edge privacy related to the requirement that the changing of the edges in graph data, but the overall relationship pattern can be published public.

Definition **edge neighbors**: the edge neighbors graph dataset refers to two graphs data G and G$^{'}$ if one dataset could transfer into another by adding or deleting one edges in the data.

Given two graph datasets G = (V, E) and G$^{'}$=(V$^{'}$,E$^{'}$), a query Q which satisfies edge level differential privacy should satisfy differential privacy for all pairs of graphs G and G for some x∈V such that: 1) V=V$^{'}$; 2)E$^{'}$=E-E$_x$,|E$_x$|=K

2.6.2 Node level differential privacy

A second adaptation of differential privacy to graphs provides much stronger privacy protection. In node-differential privacy, two graphs are neighbors if they differ by at most one node and all of its incident edges. Formally, G and G$^{'}$ are neighbors if |V⊕V$^{'}$| = 1

The node level differential privacy are used to limit the inference of whether an user is in the graph data or not. Node-level differential privacy provides preservation not only to the nodes but also the adjacent edges of nodes in graph. The definition of node-level differential privacy could be described as:

For any two neighboring graphs dataset G and G$^{'}$ and for the output sets S⊆Range(A) of algorithm A, it provides ϵ-node differential privacy only if the algorithm A satisfies the following formula:

Pr[A(G)∈S]≤exp(ϵ)⋆Pr[A(G$^{'}$))∈S]

where ϵ is a value set by the graph data providers to make a the trade-off results between privacy and accuracy. The probability Pr is computed over the random coin tosses of the algorithm.

Generally speaking, an adversary has a very limited possibility to infer their wanted information from the output of A(G) and the output of A(G$^{'}$). Therefore, the differential privacy model indicates that adding or deleting any single node will not significantly interfere with the output distribution of graph data.

This means, graph G and G$^{'}$ will produce similar distributions over the statistics released. Note that the information of each person on the graph corresponds to a specific node and all edges adjacent to the node

Note that each person's information in the graph dataset are related to a specific node and all edges connected this nodes to other node in the graph [10]. Therefore, we should conceal both the node and its edges in thegraph[79].

Social networks are often represented as a graph $G = V,E$. where the nodes $V = \{v_1, v_2, ..., v_n\}$ represent the users, and the edges $E = \{e_{ij} = (v_i, v_j) | v_i, v_j \in V, i \neq j\}$ represent the connections between them. Two nodes are adjacent if they share a common edge.

In this thesis, we use two types of graphs to represent information propagation across the network: a follower-followee graph and an information propagation graph [33]. The difference between them lies in the way directed edges are formed. In the follower-followee graph, two nodes share a directed edge if one user is a follower of the other, whereas, in the information propagation graph, a directed edge occurs when one user spreads information to another.

2.7 Graph Convolutional Network

The Graph Convolutional Network(GCN), which is a deep learning model used for graph-structured data such as social networks, citation networks, and biological networks. GCN is based on the convolutional neural network (CNN) and can operate on the graph structure. It was first introduced in 2017 by Kipf and Welling.

GCN applies a modified version of the convolution operation that can be used on graph data by leveraging the graph's adjacency matrix. The adjacency matrix represents the relationship between nodes in the graph. The convolution operation is applied on the feature representation of each node and its neighbors, with the weights being shared across different nodes. By doing this, GCN can learn a low-dimensional embedding for each node that can be used for various downstream tasks such as node classification, link prediction, and community detection.

The basic architecture of GCN consists of multiple graph convolutional layers, each followed by a non-linear activation function. The output of the last layer is usually fed into a softmax function to predict the node class or generate node embeddings. GCN can also be combined with other neural network architectures, such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs), to handle more complex graph data.

GCN has become a popular model in the field of graph neural networks due to its ability to learn expressive node embeddings that capture the graph structure and features of the nodes. It has been successfully applied in various domains such as social network analysis, recommendation systems, and bioinformatics.

The GCNs have two possible types of input:

- A feature description vector x_i for every node i can be summarized in a $N \times D$

feature matrix X, where N is the number of nodes and D is the number of input features

- A representation of the graph structure in matrix form. Typically, the matrix would be an adjacency matrix A (or some function thereof).

The output of GCN is an N × F feature matrix where F is the number of output features per node).

In the GCN model, every neural network layer can then be written as a non-linear function.

$$H^{l+1} = f(H^l, A) \tag{2.4}$$

In which $H^0 = X$ and $H^L = Z$ are the input and output of the neural network. L is the number of layers. The GCN models then differ only in how f() is chosen and parameterised. And the layer-wise propagation in GCN models follows the following rule:

$$f(H^l, A) = \sigma(AH^lW^l) \tag{2.5}$$

where W^l is a weight matrix for the l-th layer of the neural network and $\sigma(\cdot)$ is a non-linear activation function like the ReLU.

Additionally, a Fourier transform is used to find and filter the basic features of nodes and edges that might help to classify the different types of nodes in the graph.

2.8 Sparse representation

Sparse representation aims to find sparse solutions of linear system equations and is widely used in fields such as image processing, signal processing, machine learning and medical imaging [15],. In the work of Wright et al. [117], sparse representation was applied to face recognition. Yang et al. [25] proposed a single image super-resolution coding method based on signal sparse representation. In recent years, sparse representation has been widely used in image sharpness evaluation [56], 3D shape estimation [138], and wireless communication [78]. At the same time, for different types of sparse representation, scholars have also put forward a variety of solving algorithms. In the early stage, olshausen *et al.* [2] applied sparse coding to neurobiology and proposed a method based on the maximum likelihood function

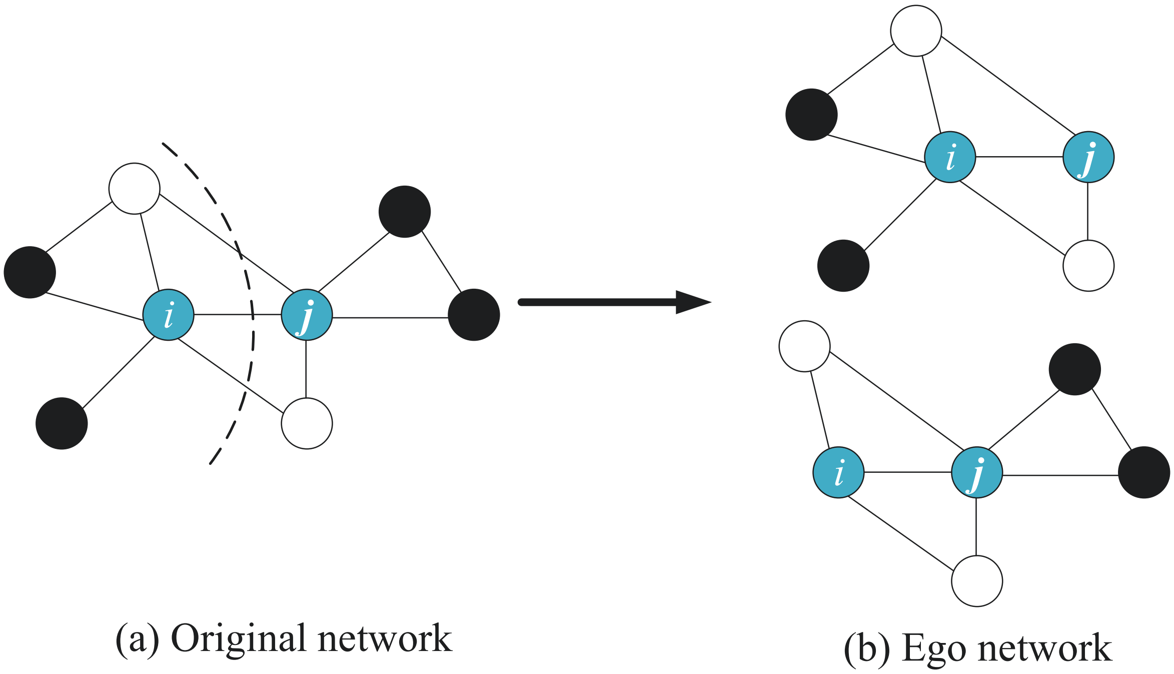


Figure 2.2: An example of the ego networks.

Among the current sparse representation algorithms, K-SVD algorithm is one of the most classical universal algorithms in sparse representation, which aims to find the optimal representation dictionary under strict sparse constraints. The algorithm is an iterative algorithm, which iterates between the sparse coding based on the current dictionary and the updated dictionary, making the sparse coding more suitable for the data itself. The update of dictionary column vector is combined with the update of sparse representation to accelerate the convergence speed. K-SVD algorithm is widely used to solve various sparse representation problems. Similar to image processing, the network can also be segmented or sampled into many small slices, that is, self-centered network. Therefore, the dictionary learning method, for example

Network representation learning are used to learn the potential low dimensional distribution vector representation of the entire network graph, while the original network structure of the network are remained [15, 110, 124]. In the network representation learning area, we use the embedding method like random walk to embed the nodes in a graph into a low-dimensional vector space. In general, the adjacency matrix are often used by researchers to represent a complex network since the adjacency matrix contains the whole connection relationships between different users in network graph and . However, the adjacent matrix required a large computational cost to represent the huge networks. As a result, the sparse representation has been used in this area to compress the matrix.

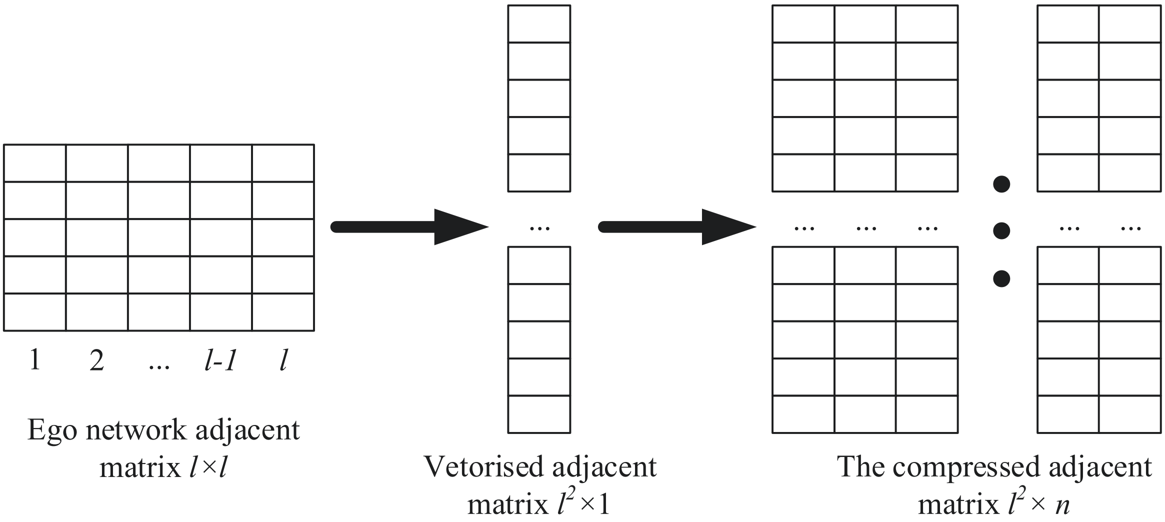

1
2
...
l-1
l
Ego network adjacent matrix $l \times l$
Vetorised adjacent matrix $l^2 \times 1$
The compressed adjacent matrix $l^2 \times n$

Figure 2.3: The process of compressing the adjacent matrix.

In order to compress the matrix, ego network is applied to solve this problem. Fig. 1 is an example of generating an ego network. The ego network is a simple local sub structure of the network, which can be obtained without any compression, and retains most of the connection information in the small-scale structure. The ego network consists of the local node textit I, the neighbors (called alters) of the node text I, and their connections. The size of ego network refers to the number of nodes. As shown in Fig. 1b, the size of the ego network is 6, because there are six nodes in the ego network, including ego. The size of self network is determined by the degree of self. The number and change of edges between self reflect the density of self network.

Compression requires the data block to be the same size as the input. To solve this problem, we adds non nodes to the ego network with nodes smaller than to ensure that they have the same size. In other words, it will add zero to the adjacency matrix to expand the size of the ego network to meet the requirements. Without losing the statistical significance of the complex network, the compression mechanism selects the average degree as the sample size of the sample self network because it is their basic statistical parameter. Sample size l refers to the size of the sample ego network. Generally, the sample size is not larger than that of the ego network.

Figure 2.3 shows the process of compressing the adjacent matrix, sub-figure (A) is the adjacent matrix of ego network with the size of $l \times l$. Sub-figure (B) is the vectorized adjacent matrix of ego network. Sub-figure (C) is the compressed adjacent matrix the adjacent matrix of an ego network L is a square matrix with the size of $l \times l$. This mechanism vectorized this adjacent matrix into a $l^2 \times 1$ vector by arranging each column vector of matrix L into one column. This process will conduct n times until each ego network from the original network has been vectorized. Then it splices these n vectorized adjacent matrix into one matrix. This is a matrix with the size of $l^2 \times n$ and contains the whole information of each ego network in the complex network. This matrix is called as the ego network matrix. For a huge network, this matrix is obviously smaller than the $n \times n$ adjacent matrix.

Based on the sampling of the ego network, network sparse then regard the networks as signals enables dictionary learning to decompose the network into several pieces of sub-graphs. This step could be considered as a problem of matrix decomposition. In this step, decomposing the ego network matrix uses the famous the K-means (K-SVD) algorithm to create a dictionary for sparse representations. It decompose the ego network matrix into a dictionary matrix and a sparse coding matrix. Like the compressed sensing in the image processing, the original adjacent matrix could be represented by using the dictionary matrix and sparse coding matrix.

In the decomposing step, the column vectors of dictionary matrix could be inverse transformated to network structures. These small structures are part of the original network and could reconstruct the original network by self replication and linear superposition. They are the so-called atoms in the complex network.

Part I

Preservation of Auction Privacy in Social Media

Chapter 3

A privacy preserving auction mechanism in online social media

3.1 Introduction

Auction is one of the fundamental research problems in various domains, such as economy [44], cloud computing [27], multi-agent systems and social networks [129]. Unlike other approaches, auction-based approaches offer a fair opportunity for each user to obtain her expected resources by bidding on these resources, which could be physical items or online services. Due to the fairness property of auction, it has become a common paradigm for item sale in online social networks and has attracted much attention [53, 129].

Existing auction mechanisms in online social networks, however, have a common limitation that the privacy of users has been overlooked. The privacy of users typically include 1) users' neighborhood information, e.g., the number of neighbors and the identities of neighbors, and 2) the values of users' bids. Therefore, a privacy-preserving auction mechanism in online social networks is highly desired. However, designing such a privacy-preserving auction mechanism is challenging due to the following two reasons.

1. Hiding users' neighborhood information usually implies that the number of neighbors of each user has to be changed. However, arbitrary changing the number of neighbors may lead to some users being isolated, i.e., having zero neighbors. These users, thus, are excluded from the auction, which is unfair to them.

2. Hiding the values of users' bids typically means modifying these values. However, such modification may result in that the users who offer the highest bids fail in an auction, which is unfair to these users.

In this chapter, we propose a privacy-preserving auction mechanism in online social networks. Privacy-preserving auction has been studied in centralized environments [18], but has not been studied in online social networks which are decentralized environments. Our privacy-preserving auction mechanism is based on the differential privacy technique. Differential privacy is a promising privacy model [22, 142] which has been successfully applied to many real-world applications [121, 128]. Differential privacy offers rigorous privacy guarantee that the analytical outputs of two datasets are almost the same if the two datasets differ in at most one data record. By using differential privacy mechanisms on a social network, each user's neighborhood information can be protected while isolated users can be avoided by properly setting the values of parameters in differential privacy mechanisms. Moreover, differential privacy can also guarantee that the values of users' bids can be preserved while the selection of the winner of an auction is not affected. This is because selecting a winner in an auction can be interpreted as querying an interesting record in a dataset. Also, the auction with differentially private modified bids can be interpreted as a neighboring auction of the original auction. As mentioned before, differential privacy can guarantee that querying results of two neighboring datasets are almost the same. Thus, it can also guarantee that selecting results from two neighboring auctions are almost the same.

In summary, this chapter has the following two contributions.

1. To the best of our knowledge, we are the first to overcome the common limitation of existing mechanisms by adopt differential privacy for privacy-preserving auction in online social networks.

2. Comparing to the latest mechanisms [53, 129] which are studied only in small networks and lack of experiments, we experimentally demonstrate that our mechanism can work efficiently in large networks with various structures.

3.2 Related work

Auction has been a common paradigm for item sale in online social networks [51] [3] [25]. The current main auction mechanisms in online social network are reviewed as follows.

Jackson et al developed the first auction mechanism in online social network based on the social network analysis [44]. Borgatti et al analyzed the theory of auction in online social networks [12]. Myerson et al proposed a mechanism by adding a reserve price to the original VCG (*Vickrey-Clarke-Groves* [19] [100] [37]) mechanism [68]. Goldberg et al. considered how to optimize the revenue of selling multiple homogeneous items [34] [35]. These auction mechanisms, however, only consider the neighbors of a seller as buyers, which may exclude a number of potential buyers who also exist in the social network but are not neighbors of the seller.

In 2017, Li et al developed an auction mechanism named IDM which not only considered the neighbors of a seller but also other potential buyers in a social network. The buyers who are not neighbors of the seller can be searched by creating diffusion critical sequences. They then extended their mechanism by accommodating the simultaneous sale of multiple items in a social network [129] [9]. They have theoretically proved that their mechanism has not only incentive compatibility but also individual rationality. However, their mechanism still has two shortages. First, they did not consider users' privacy leakage in the auction process, such as users' neighbors and values of bids. Actually, privacy-preserving auction has already been studied. Nguyen considered the protocol privacy problem in 2000 [74]. Ghosh used differential privacy preserving agents' privacy in auction [31]. However, they do not take the network topology into account, which make their methods not applicable to online social networks. Second, their mechanism was developed only in a small network with less than 15 users. In a real social network, the number of users is typically thousands. Hence, their mechanism may not be efficient in real social networks.

In this chapter, to overcome the two shortages of the existing mechanisms, we propose a privacy-preserving auction mechanism in large online social networks based on differential privacy.

3.3 Methodology

In this section, we will describe our mechanism in detail. We first give the model of social networks. Second, we describe the IDM. Then we will introduce our mechanism.

3.3.1 IDM method and its privacy leakage

3.3.1.1 The details of IDM

In this section, before we describe IDM, we first give the definitions of *diffusion critical node* and *diffusion critical sequence*.

Definition 1 *diffusion critical node*: Given a connected graph $G = (V,E)$ and three nodes v_i, v_j and v_k, we say that v_k is the *diffusion critical node* of path from node v_i to v_j if all the paths from node v_i to v_j have to pass through node v_k.

Definition 2 *diffusion critical sequence*: Given a connected graph $G = (V,E)$ and two nodes v_i and v_j, let x_1, x_2,...x_n be the *diffusion critical node* from node v_i to v_j. We define $C_i = \{x_1, x_2, ...x_n, v_j\}$ as the *diffusion critical sequence* of the path from node v_i to v_j, which is an ordered set of all diffusion critical nodes of v_j and v_j itself.

Definition 3 *Information Diffusion Mechanism (IDM)*: Let node v_B denote the highest valuation report among all the node in the graph G. To simplify the representation, let $C_m = \{x_2, ...x_n, v_B\}$ be the diffusion critical sequence of node v_B. The IDM allocates the item to the first buyer v_i in the diffusion critical sequence C_m whose bid is the highest bid when diffusion critical node i + 1 does not participate in the auction. The winner w pays the highest bid without her participation. Each buyer in $C_w \backslash w$ (winner's diffusion critical node) is rewarded the payment increase (not social welfare increase) due to her diffusion action. If $i \in C_w \backslash w$ keeps the item by herself, she will pay something say x, while if she gives it to $i+1$ and $i+1$ keeps it, $i+1$ will pay a different amount say y, the difference between y and x is rewarded to i.

Let's give an example of IDM in the Figure 1. The nodes denotes the seller in the agent who want to sell goods in the graph. There is a number which represent the valuation of the agent in each node apart from node s. It is clear that L is the buyer with the highest valuation and the diffusion critical sequence of L is $C_L = \{C, I, L\}$. According to the allocation policy, buyer I wins the auction because $v_I = 12$ is the highest valuation when L does not participate in the auction. Finally, according to the payment policy, buyer I should pay 12, and buyer C is rewarded 1 = 12-11 and the seller receive 11.

3.3.1.2 Privacy leakage in IDM

In the IDM, The number in each circles of the figure is each agent's valuation. According to the IDM, agent I wins this auction and gets the item. Therefore, to select the winner during an auction, a seller must know which agents are involved in the auction. In this situation, each agent involved in the auction has to send their neighborhood information

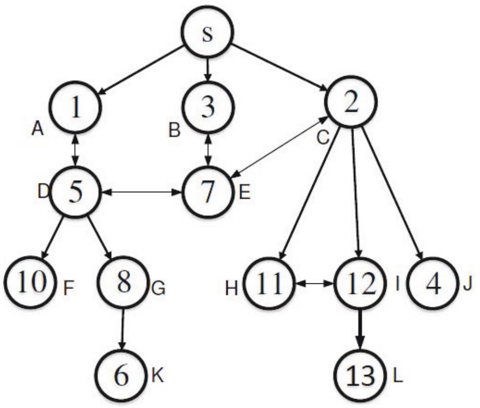


Figure 3.1: An example of the IDM auction.

to the seller. An agent's neighborhood information, however, is private and should be protected. Moreover, each agent involved in the auction has to send her valuation to the seller through other agents. Therefore, agents' valuations may be revealed to others.

3.3.2 Our mechanism

Our mechanism adopts the differential privacy technique to protect agents' neighborhood and valuation information by adding virtual nodes to each agent's neighborhood and Laplace noise to each agent's valuations.

3.3.2.1 Adding virtual node

As we have discussed before, every agent involved in an auction should report her neighborhood information to the seller. So to preserve the neighborhood information of agents in the social network, we try to add some virtual nodes to the neighborhood of each node in the social network. This step is to confuse the adversary about agents' neighborhood information in the social network.

In this step, we use the Laplace mechanism [142] to calculate the number of the nodes which should be added to the neighborhood of each agent in the social network. The Laplace mechanism adds independent noise to the true answer. We use $Lap(b)$ to represent the noise sampled from a Laplace distribution with a scaling of b.

***Definition* 6** *The Laplace Mechanism*

For a function $f : D \to R$ over a dataset D, the mechanism M provides the ϵ-differential privacy

$$M(D) = f(D) + lap(b)$$
$$b = \Delta f/\epsilon$$

where Δf is the sensitivity of differential privacy, which is a parameter determining how much perturbation is required for a particular query in a mechanism.

In our mechanism, for we should add virtual nodes to the neighborhood of each agent in the network, we use node level differential privacy. The sensitivity Δf of each agent is the number of each agent's neighbors. The number of nodes added to the neighborhood of each agent i in the social network is a random number chosen by the Laplace mechanism. Parameter b is :

$$b = De_i/\epsilon$$

where De_i is the degree of node i and ϵ is the budget of differential privacy. In our experiment, the value of ϵ is valid from 0.1 to 1.

3.3.2.2 Adding noise to valuation

Apart from the neighborhood information, the valuation information of agents involved in an auction has also to be reported to the seller. To preserve this information, each agent's valuation information should be added a Laplace noise to confuse other agents. The valuation in an agent i's bid is calculated as follows:

$$V_i^L = V_i + Lap(b)$$

Where V_i is the real valuation of agent i, V_i^L is the valuation with Laplace noise which is sent to the seller. The noise added to the valuation is chosen by the Laplace mechanism. The scaling parameter b is $\Delta f/\epsilon$. The value of Δf is determined by the perturbation in a particular query shown as follows:

$$\Delta f = Max\{i \in V | Avg_V - v_i\}$$

where V is the set of all the agents' valuations in the social network. Avg_V is the average value of all the element in set V. Δf is the maximum difference between Avg_V and v_i.

The privacy budget of this section has some difference with the privacy budget in virtual nodes. As the total number of agents in the online social network is N, the privacy budget in this section is ranged from $0.1N$ to N. In this section, we use $N\epsilon$ to represent it and ϵ is ranged from 0.1 to 1. The parameter b is:

$$b = Max\{i \in V | Avg_V - v_i\}/N\epsilon$$

The valuation of virtual nodes is determined by the agent it belongs to and the Laplace mechanism. The value of each virtual node V$_d$ is:

$$V_d = V_a + Lap(b)$$

Where V_a is the valuation of the agent which this virtual node belongs to.

In our mechanism, the valuation of a node i, which could be a real node or a virtual node, may be less than 0 due to the introduction of Laplace noise. If this situation happened, we set the valuation to 0.

3.3.3 The auction in our mechanism and privacy leakage

After adding the virtual nodes to the social network and Laplace noise on each agent's valuation, the seller will select the winner and make payments to the agents on the path between the seller to the winner using the IDM method. If the winner is a virtual node, then we regard the agent which the virtual node belongs to as the winner.

In the IDM, the price which the winner pays to the seller is based on the *The Second Price Sealed Auction (SBSP)*. However due to the noise added into each agent's valuation in our mechanism, we can only use the *The First Price Sealed Auction (SFSP)* in our mechanism.

In our social network model, the nodes represent the individual agents and the edges represent the relationship between two individual agents in the social network. During the auction process, the information in social network is travels between different agents in the social network, so the auction process exposes users to the risk of leaking private information. In this subsection, we will describe two kinds of information leakage during the auction process and introduce the differential privacy mechanism in our model to solve this problem.

The first kind of privacy information which should be preserved is the valuation information of each agent. In our auction process, each agent in the social network should send their valuation information to the agent where the auction information comes from. This action may cause privacy leakage. We use the previous figure at time t = 2 as an example. When nodes E and G send their bids back to the seller, the information must pass nodes B and C, so their valuation will leak to these agents. The number of valuations the seller received is another type of private information. For example, if agents know the seller only received a few valuations in the first round of an auction, they may collude with each other and depress the valuation collectively.

3.4 Theoretical analysis of our mechanism

In this section, we will study the properties of our mechanism. We first give some definitions of the model of auction, and then presents the analysis.

3.4.1 Buyers categories and payment

In the auction, the agents are categorized into four kinds:

1.*The winner*:The agent who get the item w.

2.*On-path buyers*:All w's diffusion critical nodes.

3.*Unlucky buyers*:If $i \neq m$, then all d_{i+1} are unlucky buyers.

4.*Normal buyers*: all buyers who are not classified in any of the other three status.

Definition 7 *Payment of agent*

In our mechanism, the payment is the same as the IDM. It could be described as follows.

Assuming that the buyer $w \in C_W$ wins the item, the payment p_i of other agents in the network is defined as:

$$p_i = V_{-d_i} - V_{-d_i+1}(if\, i \in C_W \backslash w)$$
$$p_i = V_{-d_i}(if\, i = w)$$
$$p_i = 0 \text{ (otherwise)}$$

3.4.2 Concept

In an auction, the goal of the seller is to get more agents from the network to participate in the auction and thus to maximize her revenue. However, the seller cannot directly inform all agents other than her neighborhoods, and her neighbors do not have the motivation to spread the information without utility. This subsection will give some concept which both the sellers and the buyers have to obey in the auction.

Definition 8: *individually rational*

A mechanism is individually rational if for each buyer, her utility is non-negative when she truthfully reports her valuation, no matter to whom she tells the auction information and what the others do.

Definition 9: *weakly budget balanced*

the seller's revenue generated by M is defined by the sum of all buyers' payments Rev^M. We say a mechanism is *weakly budget balanced* if $Rev^M \geq 0$.

3.4.3 Theoretical analysis

Theorem 1: In the proposed mechanism, each users is guaranteed $(\frac{1}{De_i} + N)\epsilon$-differential privacy.

Proof: To analyze the privacy guarantee, we apply a composite property of the privacy budget: the sequential composition [64], which accumulates privacy budget of each step when a series of private analysis are performed sequentially on a dataset.

In the proposed mechanism, each user consumes privacy budget in both changing neighborhood and modifying valuations. During changing neighborhood, each user consumes privacy budget $\frac{\epsilon}{De_i}$, whereas during modifying valuation, each user consumes privacy budget ϵN. Thus, based on the sequential composition, each user is guaranteed $(\frac{1}{De_i} + N)\epsilon$-differential privacy.

Lemma 1: p_i is independent of v_i if i is not the winner.

Proof:if i is an unlucky buyer, her utility is always zero which is independent of v_i. If i is an on-path buyer, (i.e., $i \in C_W \backslash w$), the payment is $p_i = v_{-d_i} - v_{-d_i+1}$ (if $i \in C_W \backslash w$). Since $i \notin v_{-d_i}$, v_{-d_i} does not depend on v_{-d_i}. Since i is not the winner in the auction, $v_{-d_i+1} \neq v_i$. Therefore, the payment of these agent is independent from v_i. From the above proof, the payment of an agent p_i is independent of their valuation v_i.

Lemma 2: This mechanism is individually rational.

Proof: if i is an unlucky buyer his payment is 0 according to the payment policy. If he is an on-path buyer, his utility is $v_{-d_i} - v_{-d_i+1}$. Since $d_i \supset d_{i+1}$, which leads to $v_{-d_i+1} > v_{-d_i}$. Therefore for an on-path buyer, his utility is bigger than 0. if he is a normal buyer, his utility is 0. Therefore, for an agent in the acution, when he truthfully reports his valuation, his utility is non-negative, our mechanism is individually rational.

Lemma 3: This mechanism is Weakly budget analysis.

Proof: According to the payment policy, the seller's revenue in our mechanism is $\sum_{i \in C_w \backslash w}(v_{d_i} - v_{d_{i+1}}) + v_{d_{w+1}} = v_{d_1}$, it is the valuation of the first agent in C_w and this value is non-negative. So our mechanism is Weakly budget analysis

3.5 Theorem Environments

We evaluated the performance of our privacy of the IDM through an extensive set of experiments in this section. Our evaluation modify the entire process of auctions including sell and bid in the social network. In the following parts of this chapter, we will first present the experiment settings, and then discuss the experimental results.

3.5.1 Datasets

The datasets we used is from SNAP [52]. It is a paper-citation network which contains 4158 nodes and 13428 edges (we just deleted some node and edges from this network to make it more suitable for the setting of our mechanism). Apart from this real network, we also constructed a random network and a scale-free network which have the same nodes and similar edges with the real network to compare with the result of the real network.

3.5.2 Evaluation process

Our evaluation performed the entire process of an auction. This evaluation is divided into the following steps:

1)Confirm the sellers: in each auction, the agents in social network must have a seller to sell their goods. In our evaluation, for each auction, we randomly choose a node from all the nodes in the network to play the role of seller.

2)Agents bid: In the IDM, every agent apart from the seller should give a valuation to the item which sell by the seller. We assume that every agent's valuation obey a normal distribution as the followings.

$$V_i = N(\mu, \sigma^2)$$

In our experiment, the valuation of all the agents in the auction is chosen from the normal distribution, the value of μ and σ we will discuss in the next subsection.

3) Use the mechanism we describe in chapter 3.3 to add the virtual nodes and noise for every agent and its valuation in the auction.

4) Find the agent who get the item in the auction through the IDM algorithm.

The above actions are steps in one auction. To test the effective of our mechanism, we should repeat these steps for several times to get the result. In the experiment, we used

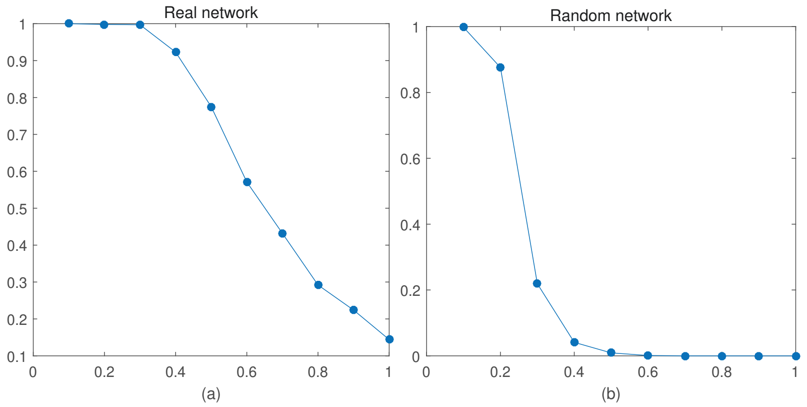


Figure 3.2: The result of *wrong-buyer rate* in real network, random network and scale-free network.

3.5.3 Metrics

To compare the difference, we also process the evaluation of the IDM on those three networks. The effectiveness of the proposed method was evaluated by comparing the difference of the result sets with the IDM. We evaluated the accuracy of our method in terms of three parameters: wrong-buyer rate, disappoint-price rate and the average-differenceÔºö

wrong-buyer rate:Let $V = \{v_1, v_2, ...v_n\}$ be the winner of the auction by the IDM mechanism, n is the number of experiments and $V^{'} = \{v^{'}_1, v^{'}_2, ...v_n\}$ be the winner of our mechanism. Wrong-buyers rate measures the difference of these two sets in the experiment:

$$wrong-buyerrate = \frac{Wrong-buyers}{n}$$
$$wrong-buyerrate = Num|V \cap V^{'}|$$

Disappoint-price:Sometimes sellers may receive a price which make them disappointed in the auction. In our evaluation, we assume that the threshold is the average valuation Œº is the disappoint-price $V_{dis} = \{v_i | v_i < \mu, v_i \in V^{'}\}$:

$$\textit{Disappoint-price rate} = \frac{disappoint-number}{n}$$
$$disappoint-number = Num|V_{dis}|$$

Average-difference:This parameter is the difference between the average value of V and $V^{'}$:

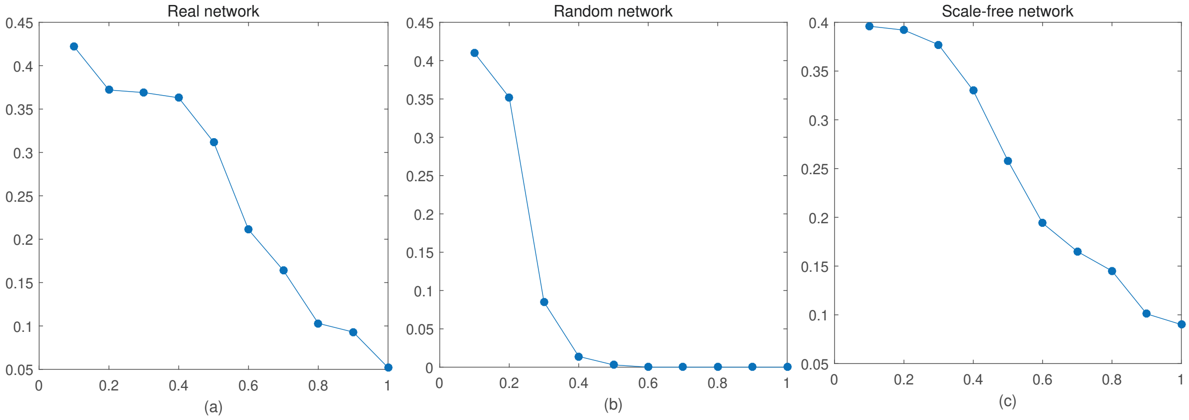


Figure 3.3: The result of *Disappoint-price* in real network, random network and scale-free network.

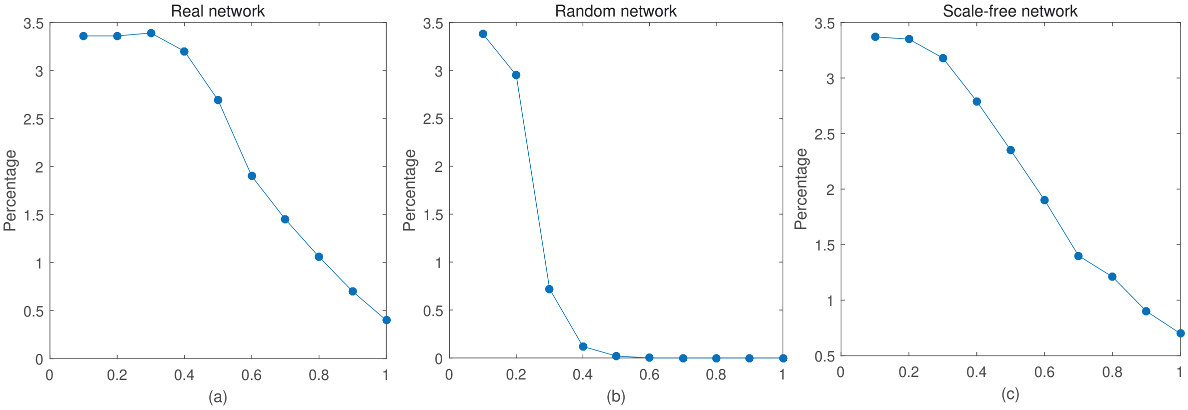


Figure 3.4: The result of *Average-difference* in real network, random network and scale-free network.

$$Average - difference = \frac{Difference}{n}$$

$$Difference = Avg_v - Avg_{v'}$$

3.5.4 Experiment setup

We set the average valuation value μ is 1000 and to compare the result of different parameters, the value of σ is ranged from 10 to 100 (Which means that the value of σ^2 is ranged from 100 to 10000). We run this evaluation for 1000 times and the result of them is shown in the following chapter.

3.5.5 Performance of our mechanism

3.5.5.1 Result of three metrics

We examined the performance of the proposed method in relation to the three kinds of networks we mentioned before. We first set the average valuation of agents $\mu = 1000$ and the variance $\sigma^2 = 100$ in the social network. the value of parameter ϵ ranged from 0.1 to 1.

The wrong-buyer rate corresponding to different values of ϵ for the three kinds of networks real network, random network and scale-free network are shown in Figure 2(a), 2(b), 2(c). As we have shown in the Fig a, the wrong-buyer rate in the real network decreased with the increasing of parameter ϵ: the wrong-buyer rate is almost 1 when ϵ is 0.1 and it increased to 0.14 when ϵ is at the value of 1. The similar situation also happened in the scale-free network that the wrong-buyer rate is decreased from 1 to 0.24. An interesting result happened in the random network which has some difference with the result of real network and scale-free network. Although the change of wrong-buyer rate has the same trend with the result in the real network and scale-free network, it decreased more quick than the other two types of networks. The wrong-buyer rate decreased to 0 when ϵ is only at the value of 0.7. This phenomenon also happened in the other two kinds of metrics which we will show in the following.

The Disappoint-price rate corresponding to different values of ϵ on the three kinds of network is shown in Figure 3(a),3(b),3(c). The result is similar to the result of metric wrong-buyers rate. The result of Disappoint-price rate in real network decreased from 0.422 to 0.052 while ϵ is ranged from 0.1 to 1. In the scale-free network, the Disappoint-price rate is 0.396 when ϵ is 0.1 and it decreased to 0.09 when ϵ at the value of 1. The situation in the random network, however, has a similar observation with metric wrong-buyer rate that with the increasing of ϵ, it quickly decreased from 0.41 to 0 when the value of ϵ is only at 0.06.

The change of Average-difference corresponding to different values of ϵ in the three kinds of network is shown Figure 4(a),4(b),4(c). In the real network, the Average-difference decreased from 3.3% to 0.4% with the increase of ϵ from 0.1 to 1. In scale-free network, the Average-difference decreased from 3.3% to 0.7% with the increase of ϵ from 0.1 to 1. The result in random network shows the same trend with the other two metrics that it decrease quickly than the real network and scale-free network with the increase of ϵ. The Average-difference decreased from 3.3% to 0.

From the above result, we could come to the conclusion that the three metrics we

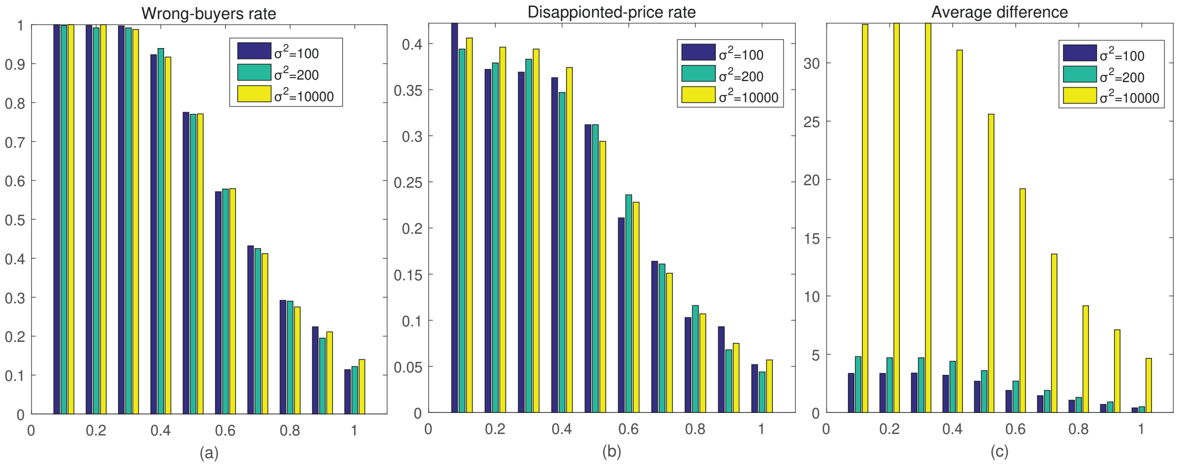


Figure 3.5: The result of the three metrics we mentioned in Section 5.3 in real network. A

mentioned before is at a low difference with the result of the original IDM when the value of ϵ is at 1. This means that our mechanism perform well in the evaluation. And this mechanism could provide privacy preserving for the agents in online social network auction. So that using differential privacy in online social network is feasible and effective.

Another phenomenon from the above result of evaluation is that all the three metrics of our experiment decreased with the increasing of differential privacy budget ϵ. The real network and scale-free network perform a similar trend in the evaluation, however, all the metrics of random network decreased quicker than the other two networks. This shows that the scale-free network could be a better model in the complex system in nature.

Even though the proposed mechanism can preserve agents' privacy in online social network auction, the effectiveness of any method could be affected by the parameter in the experiment. In the next section, we will use a set of controlled trials to show that the valuation of agents in online social network auction has only effect the metrics of average-difference. The Wrong-buyer rate and Disappoint-price rate are independent from the valuation of agents.

3.5.5.2 Influence of agents in the auction

In this section, we further examined the performance of the proposed method in relation to the valuation agent bid in the auction. We varied the valuation of agent by changing the normal distribution of valuation in the second steps of our evaluation process. We

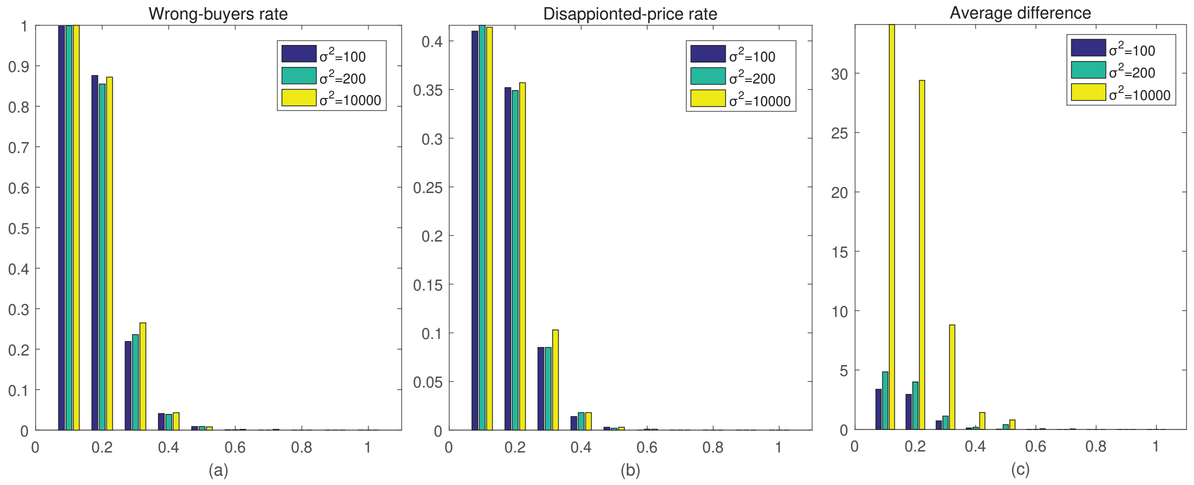


Figure 3.6: The result of the three metrics in random network.

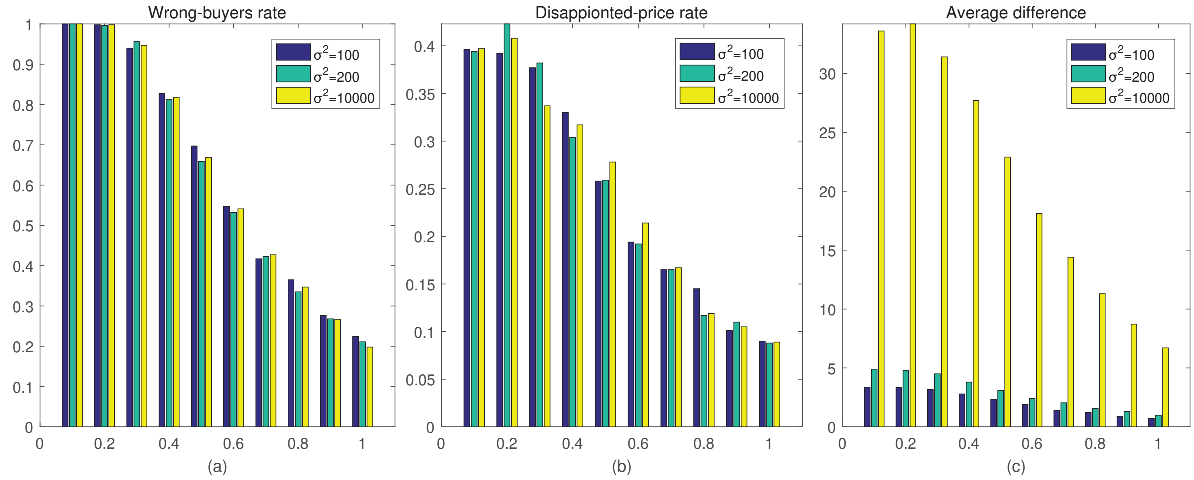


Figure 3.7: The result of the three metrics in scale-free network.

varied the value of σ^2 in 100, 200 and 10000. The result of the three value of σ^2 are shown in the following Figures.

Figure 5 shows the result of the three metrics we mentioned before in the real network. Sub-Figure 5(a) and Sub-Figure 5(b) is the result of Wrong-buyer rate and Disappoint-price rate in real network when the value of ϵ is ranged from 0.1 to 1. We could see that the result is very closed when σ^2 is 100, 200 and 10000 in all values of ϵ in these two metrics. Only in Average-difference which shown in the Sub-Figure 5(c), the result has big difference with the change of σ^2. The value is growing with the increasing of σ^2.

Similar situation happened in the other two networks which we have shown in the Figure 6 and Figure 7. Both these two figures' Sub-Figures show that the result of Wrong-buyer rate and Disappoint-price rate is very close among all the value of ϵ. The result has a big difference in the metrics of Average-difference. As the result of Average-difference does not effect the auction process, we could come to the conclusion that our mechanism is stable in the auction process.

From the above the figures, we can see that the three value of σ^2 in the evaluation have the same trend in the three kinds of networks: in the real network and scale-free network, the Wrong-buyer rate decreased with the increase of ϵ and they did not decrease to 0 when $\epsilon = 1$. In the random network, the Wrong-buyer rate decreased to 0 when ϵ $= 0.6$. This result shows that the valuation of agents could not effect the result of our auction mechanism. It is independent from the agents' bid and only the Laplace noise has effect on the result.

3.6 Illustrations and Photographs

In this chapter, we overcome a common limitation of existing auction mechanisms in online social networks by proposing a privacy-preserving auction mechanism. Our mechanism is based on the differential privacy technique. By using differential privacy, both the users' neighborhood information and their values of bids can be preserved without sacrificing the fairness and performance of users.

In the future, we intend to improve our mechanism from two ways. First, our mechanism is partially based on the information diffusion mechanism. This mechanism, however, has a drawback that the diffusion critical sequences may not be found in specific social networks, where a few nodes have high degrees. To address this drawback, a new information diffusion mechanism is needed. Second, our experiments were conducted on the basis of a paper-citation social network. In the future, after a new information diffusion mechanism is developed, we will evaluate our mechanism on other social networks, e.g., Twitter and Facebook.

CHAPTER 4

PRIVACY PRESERVATION AUCTION IN A DYNAMIC SOCIAL NETWORK

4.1 Introduction

The growing popularity of users in online social network gives a big opportunity for online auction. The famous *Information Diffusion Mechanism* (IDM) is an excellent methods even meet the incentive compatibility and individual rationality. Although the existing auction in online social network has considered the buyers' information has not known by the seller, current mechanism still can not preserve the information such as prices. In this section, we propose a novel mechanism which modeled the auction process in online social network and preserved users' privacy by using differential privacy mechanism. Our mechanism can successfully process the auction and at the same time preserve clients' price information from neighbours. we achieved these by adding Laplace noise for its valuation and the number of valuation seller received in the auction process. We also formulate this mechanism on the real network to show the feasibility and effective of the proposed mechanism.

Auction is a fundamental research problem in various domains, such as economics [43], cloud computing [27], multi-agent systems [111] and social networks [129]. Unlike other approaches, auction-based approaches offer a fair opportunity for each user to obtain their expected resources by bidding on these resources, for example physical items or online services. Due to the fairness property of auctions, it has become a common paradigm for item sale in online social networks and has attracted much attention [53, 129].

There are two categories of interest in current research in the area. One category is building a mechanism for online social network auctions and analysing the theory. For example, Jackson *et al.* developed the first auction mechanism in online social networks in their analysis of social and economic networks in 2010 and Borgatti *et al.* performed an analysis on auction theory in online social networks. Another is to consider the effects of social links connecting these buyers to other potential buyers who are unknown to the seller and join them into the auction. Li *et al.* proposed the very first model which considers buyers unknown to the seller in an auction in 2017. They developed an auction mechanism named IDM which is not only incentive compatibility but also individual rationality. Moreover, they considered multiple sellers in online social network which used a similar payment policy in 2018. Existing auction mechanisms in online social networks, however, have two common limitations. The first limitation is the auction process. The process of one auction are often ignored. For instance, they do not consider how information propagates within an auction, how agents reply their valuation information to sellers through social network platforms and when the seller closes the auction. These neglected problems prevent the current work from providing an ideal platform to perform auctions. The second limitation is that user privacy has been overlooked. User privacy in this case typically includes the value of user bids and the number of valuations the seller received during the auction. Therefore, a privacy-preserving auction mechanism in online social networks is highly desired. However, designing such a mechanism is challenging for the following two reasons.

1. Hiding the values of users' bids typically means modifying these values. However, such modification may result in a situation in which a user who offer the highest bid fails to win an auction, which is unfair to that user.

2. Hiding the count of valuations that sellers receive typically means changing the number of valuations received. However, such a change would result in the wrong information being provided to other agents in the social network regarding the popularity of the product or similar products among the social network's users.

In this chapter, we propose a privacy-preserving auction mechanism in online social networks. To address the issue of how auction information propagates in social networks, we first build a new model to simulate the process in online social network auctions. Considering the reaction time in social networks (agents in social networks often need time to react), we introduce time intervals into our auction model. Privacy-preserving auctions have been studied in centralised environments [18] but have not been studied in online social networks which are decentralised environments. Our privacy preserving auction mechanism is based on differential privacy theory. Differential privacy is a promising privacy model [22, 142] which has been successfully applied to many real-world applications. Differential privacy offers a rigorous guarantee that the analytical outputs of two datasets are almost the same if the two datasets differ by at most one data record. By using differential privacy mechanisms on a social network, the seller's valuation information can be protected. Moreover, differential privacy can also guarantee that the values of users' bids can be hidden while the selection of the winner of an auction is not affected. This is because selecting a winner in an auction can be interpreted as querying an interesting record in a dataset. Also, an auction with differentially private modified bids can be interpreted as a neighbouring auction of the original auction. As mentioned before, differential privacy can guarantee that querying results of two neighbouring datasets are almost the same. Thus, it can also guarantee that selecting results from two neighbouring auctions are almost the same.

In summary, this chapter has the following three contributions.

1. We overcome the common limitation of existing mechanisms by building a model to simulate the process of auctions in online social networks. our demonstrate that our mechanism can work efficiently in large networks with various structures.

2. We preserve the agents' privacy (including values of users' bids and the number of valuations received by the seller) by adopting differential privacy for auctions in online social networks.

4.2 Related Work

4.2.1 Auction mechanism in social netwok

The auction has been a common paradigm for item sale in online social networks [51] [3] [25]. The current main auction mechanisms in online social network are reviewed as follows.

Jackson *et al.* developed the first auction mechanism in online social network based on the social networks analysis [43]. Borgatti *et al.* analysed the theory of auctions in online social networks [12]. Myerson et al proposed a mechanism by adding a reserve price to the original VCG mechanism. Goldberg *et al.* considered how to optimise revenue when selling multiple homogeneous items [34] [35]. These auction mechanisms, however, only consider the neighbours of a seller as buyers, which may exclude a number of potential buyers who also exist in the social network but are not neighbors of the seller. This insufficient means that a potential buyer with a high valuation may not be aware of the auction since she is not directly connected to the seller and her neighbors do not wish to tell her the auction information in these auction mechanism. It is a fatal deficiency which purport that their mechanism is not suitable in the real social network model.

To solve this fatal deficiency, in 2017, Li *et al.* developed an auction mechanism named IDM which not only considered the neighbors of a seller but also other potential buyers in a social network. In this chapter, Li *et al.* has also analysed the classical VCG mechanism. The result shows that the VCG mechanism in online social network is neither incentive compatibility nor individual rationality. Then he introduced *diffusion critical node* into his paper and expanded this concept into *diffusion critical sequence* which captures how the auction information from the seller is propagated to buyer who gives the highest bids in the auction through its diffusion critical nodes. The buyers who are not neighbours of the seller can be searched by creating diffusion critical sequences. It is essentially a partial route of all the information diffusion paths from the seller to buyer i, which clearly shows to what extend each node in the sequence affects the information diffusion process. Using this diffusion critical sequences, Li *et al.* proposed their mechanism and theoretically proved that their mechanism has not only incentive compatibility but also individual rationality. Meanwhile, They then extended their mechanism by accommodating the simultaneous sale of multiple items in a social network [129] [9]. They guaranteed that the seller's revenue is better than not using the advertising and the seller does not need to pay if the advertising is not beneficial to her in this new mechanism.

4.2.2 Privacy preserving in social network

Social network comes from the social activity in online network. With the increasing number of users in the online social networks sites, the data of social network has gained much attraction in different area. These data are usually abstracted into graph structures in the research area. Social roles are abstracted into the nodes in graph, and their relationships constitute the edges of graphs. As a result, social networks constitute a special kind of graph, and many research methods of graph theory are used in social networks. One is the personal information of each users such as identity number, demographic information or ages that contains in the network, the other is the relationship between different individuals such as a patient and doctor, parent and children or the friendship which hide in the edges between different nodes.

To preserve privacy information of SNSs users, the data providers will anonymize the complex network before they publish the data to the public. The anonymization methods could be categorize the into two classes of technique. The first category is the random graph editing techniques and the other category of methods is the k-anonymization techniques. Random graph editing techniques achieve preserving users' privacy on SNSs by randomly adding, deleting or switching of some of the edges on the SNSs data. The K-Anonymization Approache is to construct an anonymized graph of the SNSs data G_1=(V_1,E_1), in which this anonymized graph G_1 satisfies the k-anonymized. In this situation, the adversary with such background knowledge could not re-identify a specific information in this graph with a probability greater than 1/k.

In recent years, Differential privacy mechanism is used in this area to find a general privacy preserving mechanism in social network data. In theory, it could ensures that regardless of whether the adversary chooses to opts in to, or opts out of the data set, his or her ability to lead privacy leakage of any individual in the data set is basically the same. Compared with the previous privacy preserving model, the differential privacy model can successfully preserve the privacy information of individual user from background attacks and provide verifiable privacy guarantees. The existing differential privacy mechanism for the SNSs data publish can be divided into two parts: the node-level differential privacy and edge-level differential privacy.

4.2.3 Summary

The current auction mechanism IDM which developed by Li *et al.* considered the neighbors of a seller but also other potential buyers in a social network. More than thatÔºåLi *et al.* also showed that their mechanism is not only incentive compatibility but also individual rationality. However, this mechanism exists several problems that we will show in the following. These problems makes it not excellent:

1. The IDM mechanism is only a payment policy of auction in online social network. That is to say it did not simulate the whole process of auction in their work.

2. The IDM mechanism did not consider users' privacy leakage in the auction process, such as users' neighbors and values of bids. In fact, privacy-preserving auctions have already been studied. Nguyen considered the protocol privacy problem in 2000 [74]. Ghosh *et al.* used differential privacy preserving agents' privacy in auction

[31]. However, they do not take the network topology into account, which prevent their methods from being applicable to online social networks.

3. The IDM mechanism was developed only in a small network with less than 15 users. However, the number of users is typically thousands in a real social network. Hence, their mechanism may not be efficient in real social networks.

In this chapter, to overcome the two shortcomings of existing mechanisms, we propose a privacy-preserving auction mechanism in large online social networks based on differential privacy.

4.3 Preliminary

4.3.1 Social network model

A social network is often represented using a graph model by researchers of social networks. An online social network is modelled as a graph $G(V,E)$, where V is a set of nodes $V = \{v_1, v_2, ..., v_n\}$ which represents the users in the online social network, and E is a set of edges $E = \{e_{ij} = (v_i, v_j) | v_i, v_j \in V, i \neq j\}$ which represents the connection between different users in the network. Two nodes are called adjacent if they share a common edge. In this chapter, this means that the one node is able to directly deliver auction information belonging to themselves or their neighbours. All information in the auction is delivered through edges in the graph model. In our mechanism, we assume that the graph is undirected which means that if two nodes share a common edge, they can both deliver information to each other. In the rest of the paper, we use agents, users and nodes interchangeably.

4.4 Weakness of current method

The current research of auctions in social networks has focused on how to attract more agents to participate in auctions. In 2017, Li et. al proposed the very first model for selling one item in a social network, in this thesis, Bin proposed a mechanism named Information Diffusion Mechanism (IDM) in social network auctions. This method expands the diffusion critical node into multiple diffusion critical nodes, and further into a diffusion critical sequence. The IDM mechanism allocates the item to the first buyer i in the diffusion critical sequence whose bid is the highest bid when the diffusion critical node i+1 does not participate in the auction. The agent who wins the auction pays the next highest bid amount, i.e. the highest bid excluding their own. Li *et. al*'s work has opened a new area on how to operate an auction in social networks. However, several problems exist in this mechanism:

- The key point of IDM is the use of the diffusion critical sequence for this mechanism. However, very few nodes in social network have diffusion critical sequence as the average node degree is too high in most real networks. One of our networks which is constructed from Twitter users has an average degree of 53. This mechanism is not applicable for most real networks.
- This mechanism fails in certain cases in that the highest bid sometimes is not selected as the winner. This result is caused by unreasonable payment in the network. Solving this problem is a significant hurdle in this area.
- The IDM is only a mechanism which gives rational payments for the agents who participate in an auction. It does not concern the specific process of an auction. The authors do not consider the use of their mechanism in a real social network.

To solve the above problems, we will use real data from a social network to build a new auction model. We will elaborate the entire auction process in social networks and demonstrate it in the environment of a real social network. In the following section, we will first present our social network data model, then introduce our auction process for it.

4.5 Auction model based on IDM mechanism

In this section, we try to use a social network model to modify the auction process in social networks with some principles. Considering a social network with N nodes, all the agents in this social network can only deliver information to their neighbouring agents. One of the nodes in this network wants to sell items in this social network, *i.e.* the seller. The auction process can be described as the following: the seller first delivers their auction information to neighbouring agents in the social network. Some of these neighbouring agents will then give their valuation information to the seller or deliver the auction information to their neighbourhood agents. The auction will be closed if the seller receives a satisfactory price.

In the following subsections, we will discuss our auction model in detail. We will first introduce a very important factor, the time delay in our model. We will then give the basic principles of our model's auction process. Finally, we will analyse the privacy leakage in our auction model and use differential privacy to fix this.

4.5.1 Reaction time

In social networks, users often need time to react to any occurrences. For instance, if a user comments on a tweet in Twitter, the user who published this tweet will not necessarily reply to the commenter immediately. It often takes time for the original author to reply, if they do reply at all. This phenomenon also exists in auctions on social networks. When an agent receives information, it often takes time for them to react to it. If they are not interested in the offer, they are unlikely to participate in the auction. In the following parts of this article, we express these time periods as 'react time'.

In order to solve this problem and improve our model's accuracy, we will introduce an important factor in our auction process: reaction time. It exists in two situations: 1) during information transfer between two agents, and 2) in user reaction time. In this thesis, we assume that all information transfers between two adjacent agents use one time interval. User reaction time depends on the experiment, as we will demonstrate in the next section.

To apply this factor in an online social network auction, we conduct our auction process over several time intervals. In each interval, agents have two possible statuses: 'reacted' or 'none'. According to Li *et al.* 's work, agents in the 'reacted' status are those who have already received the auction details and thus perform two actions: they their valuation to the agents from which they received the auction details, and they spread the auction's details to other agents in social network.The 'none' status covers two categories: agents who have not received auction details and agents who have received details but do not want to participate in the auction. Agents in this status do not perform any actions. In the next section, we propose a novel model based on time interval which will is suitable for real social networks.

4.5.2 Auction process

In this subsection, in order to address the issue of reaction time in social network auctions, we introduce our auction model. Based on the above notions, we first give some principles of our auction model.

- Two nodes i and j, where we assume that the auction details are transferring from i to j. The information transferred between them is divided into two parts:

1) the auction details from i to j where the starting point is the seller, and 2) the valuation information from node j to i.

- These two kinds of information transferred between nodes have a time interval, i.e. the information transfer from i to j needs takes a certain amount of time. In our model, we assume that information transfers between two adjacent nodes take one time interval.

- In each time interval, the status of agents who do not receive the auction details are always in 'none' status. For agents who have received the information, their status can either still be 'none' or they may report their valuation and spread the information. In our model, reporting valuation and spreading auction information are two independent actions. They are decided by two independent probabilities p and q. Agents who report their valuation and spread the auction information are in the 'reacted' status.

- Agents in the social network may be in the 'none' status for a long period of time. Therefore, each agent in the network can deliver their valuation information several times until he received the reject information.

- When the seller thinks there are enough bids, the auction is shut down. The seller chooses the agent with the highest valuation in the network and sends a reject message to other agents in the network. Notice that the reject message is also delayed by the time intervals regarding information transfer.

Based on the above principles, our auction model for social networks can be described as the following:

1. During the first time interval, a seller wants open a new auction in the network. He sends the auction information to all of his neighbourhood agents.

2. In the second time-interval, two probabilities are generated for every neighbourhood agent to determine their actions. Each agent who is in the 'reacted' status will report his valuation to the seller or deliver the auction information to other neighbourhood agents.

3. In the third time-interval, like in the second time-interval, two probabilities are used to judge the status of the agents who have received the auction information. If he/she is in the react status, he/she will report his valuation to the agents from

which he received the auction information or deliver the auction information to other neighbourhood agents.

4. The following time-intervals are the same. We first determine each agent's status. Then each react agents will report his valuation to the agents where he received the auction information or deliver the auction information to other neighbourhood agents.

5. After an enough time-intervals, when seller received a satisfactory price, he will close the auction and choose the agent who gives the highest valuation in the network. In the following part of this thesis, we use the term 'winner' to designate the agent who wins the item at the auction. If the seller does not receive a satisfactory price, he will also close the auction and open a new auction in the network.

Figure 4.1 is a simple example of the auction model. The node S is the seller in this network. At t = 0, this seller sends the auction information to all neighbourhood agents in the network. At t = 1, nodes A, B and C receive the auction information, however A is still has a status of 'none' while B and C have a status of 'reacted'. During this time interval, nodes B and C send the auction information to their neighbourhood agents in the network. At t = 2, nodes E, F and G receive the auction information. During this time interval, nodes A, E and G change to the 'reacted' status and they will reply with the valuation information to the agents who sent them the auction information and send the auction information to their neighbourhood nodes. At t = 4, nodes D and K receive the auction information and they are both in the 'react' status. They will reply their valuation information to nodes A and G. Node D will also send the auction information to nodes H and I. The valuation information from nodes E and G are sent to seller S from nodes B and C. This process will continue until the seller shuts down the auction process.

According to the theory in Li et al.'s work, to encourage more agents who are not known to the seller to participate and spread the auction information, the auction model should incentivize individuals to spread the auction information in the social network by giving the agents who do this some form of payment.

In this section, to address the revenue issue of our auction process in social networks, we design a mechanism based on the agents who spread the information to the winner of an auction. We first give another definition of spreading sequence in our auction model.

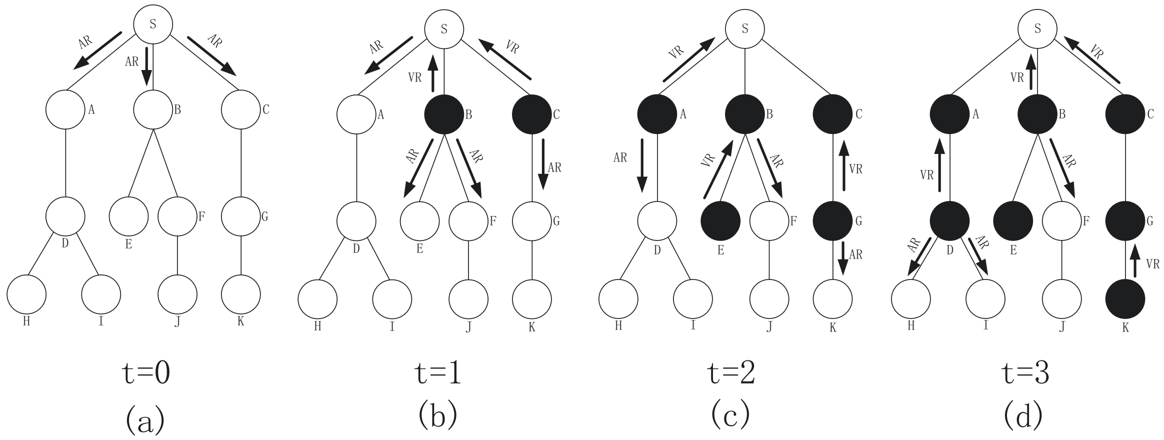


Figure 4.1: This figure is a simple example of our auction process.

Definition *spreading sequence*: Given an action process a, for each agent i $\in$N$_{-s}$, we define the nodes through auction information has passed by C_i = {x_1,x_2...x_k,i} the spreading sequence of agent i.

Assume that under the auction policy, agent w wins the item. The seller should pay the agents who are in the spreading sequence a part of the revenue as a reward. The payment rate depends on the seller themselves.

4.5.3 Privacy preserving in auction process

In our social network model, the nodes represent the individual agents and the edges represent the relationship between two individual agents in the social network. During the auction process, the information in social network is travels between different agents in the social network, so the auction process exposes users to the risk of leaking private information. In this subsection, we will describe two kinds of information leakage during the auction process and introduce the differential privacy mechanism in our model to solve this problem.

The first kind of privacy information which should be preserved is the valuation information of each agent. In our auction process, each agent in the social network should send their valuation information to the agent where the auction information comes from. This action may cause privacy leakage. We use the previous figure at time t = 2 as an example. When nodes E and G send their bids back to the seller, the information must pass nodes B and C, so their valuation will leak to these agents. The number of valuations the seller received is another type of private information. For example, if agents know the seller only received a few valuations in the first round of an auction, they may collude with each other and depress the valuation collectively.

Our mechanism adopts the differential privacy technique to preserve agents' valuation information and the number of valuation seller received by adding Laplace noise to them. Before introducing our mechanism, we first depict differential privacy in social networks. To hide this information, Laplace noise should be added to each agent's valuation information to confuse other agents. The valuation in an agent *i*'s bid is calculated as follows:

$$V_i^L = V_i + Lap(b) \tag{4.1}$$

Where V_i is the real valuation of agent i, V_i^L is the valuation with Laplace noise which is sent to the seller. The noise added to the valuation is chosen by the Laplace mechanism. The scaling parameter b is $\Delta f/\epsilon$. The value of Δf is determined by a random function ranging from 0 to 10, we use ϵ to represent the budget with ϵ ranging from 0.1 to 1.

For the number of valuations, we try to use differentially private counter to hide the number of valuations the seller received. The differentially private counter was proposed by chan *et al.* and Dwork *et al.*. It is a differential privacy methodology for stream data. Given a bit stream $\sigma = (\sigma_1, \sigma_2, ...\sigma_T) \in \{0,1\}^T$ streaming counter $M(\sigma)$ releases an approximation to $c_\sigma(t) = \sum_{i=1}^{t} \sigma_i$ at every time step t. The counters release accurate approximations to the running count at every time step. The definition of the differentially private counter is shown in the following:

A streaming counter M is (α, β) useful if with probability at least 1 - β for each time t $\in \{T\}$

$$\|M(\sigma)(t) - c_\sigma(t)\| \leq \alpha \tag{4.2}$$

In our model, the number of valuations received by the seller can be considered as a stream data during the time intervals. We divide the total time intervals into several periods. Then we add noise to each period's valuation count by the following perturbation:

$$N_i^L = N_i + Lap(b) \tag{4.3}$$

where N is the number of valuations in this time intervals. N_i^L is the number valuation with noise. The noise added to the valuation is chosen by the Laplace mechanism. The scaling parameter b is $\Delta f/\epsilon$. The value of Δf is 1.

4.6 Theorem Analysis

In this section, we theoretically analyse the performance of the proposed auction framework. We first introduce basic definitions and lemmas about differential privacy and game theory, and then give detailed analyses in terms of individual rationality, revenue and privacy.

[Individual rationality] An auction framework is individually rational if and only if no agents receive negative utility, namely, we have $u_i \geq 0$ for each agent i. Individual rationality is significant to incentivize agents participating in the online auction. Our proposed auction framework is individual rational. *Proof.* For our proposed auction framework, we need to prove the individual rationality for both buyers and sellers. We present the proof as follows.

(1) For the seller in the auction, if she/he starts an auction in the social network, and received satisfying bids before the auction is closed, then the seller has utility $u_s = (1-\rho)(p - v_s)$, where ρ is the ratio of money for rewarding agents on the spreading sequence, p_s is the final amount received by the seller, v_s is the seller's valuation of the good. It is obvious that $u_i \geq 0$ because otherwise the seller will not close the auction. On the other hand, if the seller does not receive a satisfactory bid, then the item cannot be sold, the utility of seller is 0 in this case. Therefore, the seller's utility is always non-negative.

(2) For agents (other than the seller) in the auction.

- If the agent decide to participate in the auction and submit her/his bid, then the agent is a potential buyer:
  - If the finally wins the auction (the winner), his/her utility will be $u_b = v_b - p$, where v_b is the agent's valuation, p is corresponding payment. As discussed in the payment policy section, no matter $p_{noise} \geq p_{true}$ or $p_{noise} < p_{true}$, the final payment is guaranteed to be p_{true} by a trustworthy third party. In other words, p_{true} is an acceptable price to the winner, therefore, $u_b \geq 0$.
  - If the agent does not win the auction but is in the spreading sequence, she/he will receive reward $\rho(p - v_s)$ from the seller, therefore, $u_b = \rho(p - v_s) \geq 0$.
  - if the agent does not win the auction and is not in the spreading sequence, then $u_b = \rho(p - v_s) \geq 0$.
- If the agent decides not to participate in the auction, then his/her utility will always be 0.

In summary, the utility of agents in our proposed auction framework is always non-negative.

Next, we analyse the amount of money the seller could receive and investigate how the Laplace mechanism influences it. [Accuracy bound of the Laplace mechanism] Let $f \to \mathbb{N}^{|X|} \to \mathbb{R}^k$, and $y = \mathcal{M}_L(x, f(\cdot), \epsilon)$ be the Laplace mechanism for $\forall \delta \in (0,1]$, then

$$\Pr[||f(x) - y||_\infty \geq \ln(\frac{k}{\delta})(\frac{\Delta f}{\epsilon})] \leq \delta. \tag{4.4}$$

Suppose all the N_p true prices of active agents (buyers) fall in the interval $[p_{min}, p_{max}]$, and denote $\Delta p = p_{max} - p_{min}$. In the condition where privacy is not considered, the seller could choose the agent whose bid is p_{max} as winner, and receive money p_{max}. However, when taking privacy into consideration, the agent who bids the highest price may not be the one who wishes to provide the most amount of money. Note that potentially, the maximum amount of money that may be received by the seller is p_{max}.

With probability less than $\Pr = 1-(\frac{1}{2})^{N_p}(1-e^{-\epsilon\frac{\Delta p}{\Delta f}})$, the seller obtains revenue less than p_{max}, where p_{max} is the maximum price of some agents without adding perturbation. *proof.* According to Lemma.1, set $\ln(\frac{1}{\delta})(\frac{\Delta f}{\epsilon}) = \Delta p$, we have that with probability less than $\delta_0 = \frac{1}{2}e^{-\epsilon\frac{\Delta p}{\Delta f}}$, the noisy bid p_{min} will exceed p_{max}, and the agent who bids p_{min} might become winner. Therefore, taking all the N_p bid together, we have the probability that the seller cannot receive money p_{max} is less than $1 - \frac{1}{2}[(\frac{1-\delta_0}{2})^{N_p-1}] = 1 - (\frac{1}{2})^{N_p}(1 - e^{-\epsilon\frac{\Delta p}{\Delta f}})$ □.

Finally, based the following lemma, we analyse the privacy promised by the proposed auction framework. [Parallel composition [23]] Given a set of privacy steps $\mathcal{M} = \{\mathcal{M}_1, ..., \mathcal{M}_m\}$, if $\mathcal{M}_i$ provides ϵ_i privacy guarantee on a disjointed subset of the entire dataset, the parallel of $\mathcal{M}$ will provide $\max\{\epsilon_1, ..., \epsilon_m\}$-differential privacy.

Our proposed auction framework is ϵ-differential private. *proof.* As discussed in section 4.3, the proposed auction framework allows each agent to perturb her/his price locally before submitting it to the seller. The Laplace mechanism parameterized by ϵ is employed to add noise. Therefore, for each agent, her/his data is protected in a ϵ-differentially private manner. As the price of each agent is disjoint with others' prices, according to Lemma.2, the proposed auction framework is ϵ-differential private. □

4.7 Experiment

The current auction model in graph model does not use a real social network as their datasets. This weakness makes them not enough persuasive. Therefore, in this thesis, we use a real online social network to test our auction model. This network contains 4158 nodes and 13428 edges, with an average degree of around 4. In the following parts of this section, we first give a result of the whole auction process, then analyse the result of the auction process with the differential privacy mechanism included.

4.7.1 Auction model and simulation

The first test was the whole process the auction model and the simulation in a real social networks. To evaluate how many agents will receive the auction information in the model, we used the following settings: the reporting probability p and deliver probability q is the same value 0.2 and the whole process will run for 100 time intervals. Figure 4.2 has four sub-figures which shows the result of the auction process at different time intervals. At t = 10, only 2.43 percentage of agents in this network are in react status. Then, the number of nodes grows very fast in the following time intervals. At t = 40, nearly 60 percent of nodes are in react status and 79 percent get into react status at t = 70. However, the propagation speed slows down towards the end. At t = 100, 82 percent of nodes are in the react status. Figure.3 is the number of react nodes in each time interval. At the beginning, the information will spread very quickly, after which most nodes in the social network will be in react status. The spreading speed will then slow down and it is difficult to make every single agent in a social network switch to react status. This result justifies the information spreading phenomenon in social network. Figure 4.3 shows the users affected through the auction process. From t = 0 time interval to t = 40 time interval, the valuation information spread very fast and 60 percentage of agents in social network join in the auction. After the t = 40 time interval, the spreading speed slow down and at t =70 time interval, 79 percent of nodes are in react status. Then the speed almost shut down and at t =70 time interval, 82 percent of nodes are in react status.

4.7.2 Auction process with differential privacy mechanism

Our evaluation of the auction process with the differential privacy mechanism performed over the entire auction process and add noise to every agent's valuation. This evaluation is divided into the following steps:

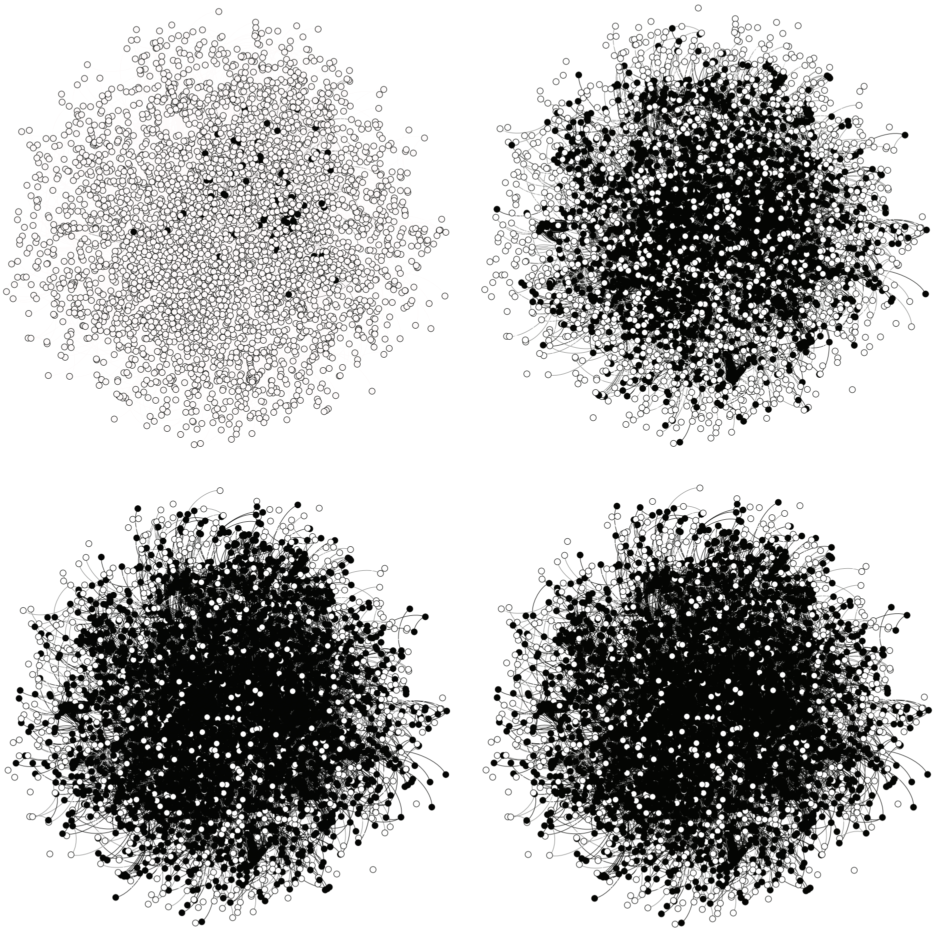


Figure 4.2: The process of our action model at the time interval of 10, 40, 70, 100

1) Confirm the sellers: for each auction in the experiment, we randomly choose a node from all the nodes in the network to act as seller.

2) Agents bid: every agent with a status of 'react' will send a bid for the item. We assume that every agent's valuation obeys a normal distribution per the following.

$$V_i = N(\mu, \sigma^2) \tag{4.5}$$

In our experiment, the each agents' bid is chosen from a normal distribution, the value of μ and σ we will discuss in the next subsection.

3) Add noise Use the differential privacy mechanism to add noise for every agent and the number of bids the seller receives over the course of the auction.

4) Find the auction winner The above steps cover a single auction process. To test the effectiveness of our mechanism, we need to repeat these steps multiple times to get a

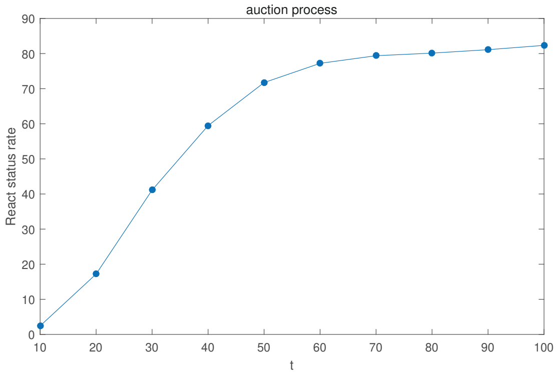


Figure 4.3: The percent of users affected in our auction process.

generic result. In this section, we test the auction process 1000 times.

4.7.2.1 Metrics

To compare the difference, we evaluate the auction process on those three networks. The effectiveness of the proposed method was evaluated by comparing the difference of the result sets with the process which did not add noise. We evaluated the accuracy of our method in terms of three parameters:

Wrong-buyer rate:Let $V = \{v_1, v_2, ...v_n\}$ be the winner of the auction by the IDM mechanism, where n is the number of experiments. Let $V^{'} = \{v_1^{'}, v_2^{'}, ...v_n\}$ be the winner of our mechanism. Wrong-buyer rate measures the difference of these two sets in the experiment:

$$
\begin{aligned}
&wrong-buyersrate = \frac{Wrong-buyers}{n} \\
&wrong-buyers = Num|V \cap V^{'}|
\end{aligned}
\tag{4.6}
$$

Disappoint-price: Sometimes the final price does not meet the expectations of the seller. In our evaluation, we assume that the threshold is the average valuation is the disappoint-price $V_{dis} = \{v_i | v_i < \mu, v_i \in V^{'}\}$:

$$
\begin{aligned}
&Disappoint\text{-}price\ rate = \frac{disappoint-number}{n} \\
&disappoint-number = Num|V_{dis}|
\end{aligned}
\tag{4.7}
$$

Average-difference:This parameter is the difference between the average value of V and $V^{'}$:

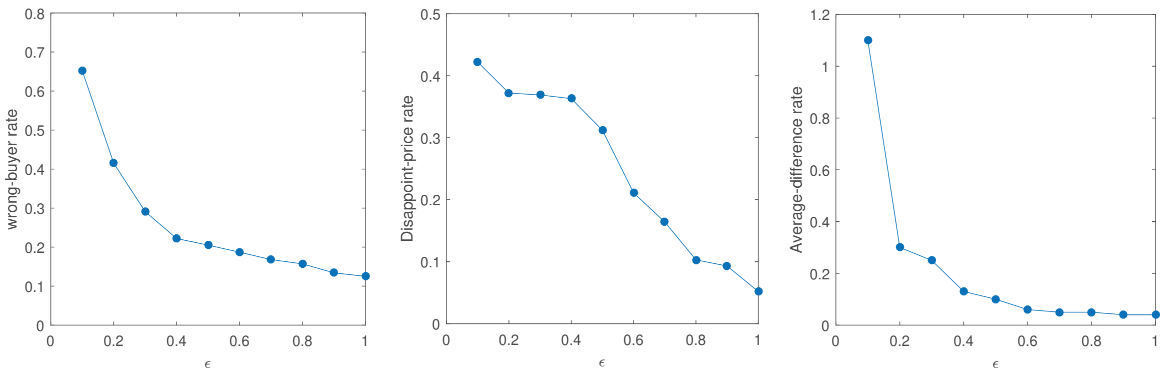


Figure 4.4: Comparison results of wrong buyer rate, disappoint-price and average difference.

$$\begin{aligned} Average-difference &= \frac{Difference}{n} \\ Difference &= Avg_v - Avg_{v'} \end{aligned} \tag{4.8}$$

4.7.2.2 Result of the auction

We examined the performance of the proposed method in relation to the networks which we have mentioned before. We first set the average valuation of agents $\mu = 1000$ and the variance $\sigma^2 = 100$ in the social network. The value of parameter ϵ ranged from 0.1 to 1. The threshold of disappoint-price is set to 1050.

The wrong-buyer rate corresponding to different values of ϵ for the social network is shown in Figure 4.4(a). As we have shown in this sub-figure, the wrong-buyer rate in this social network decreased with the increase of parameter ϵ: the wrong-buyer rate is 0.657 when ϵ is 0.1 while it increases to 0.125 when ϵ is 1. This result shows the metric of wrong-buyer is decreasing with the increasing of ϵ and a similar situation also happened in other two metrics.

The Disappoint-price rate corresponds to different values of ϵ on the social network as shown in Figure 4.4(b). The result is similar to the result of metric wrong-buyer rate which decreases from 0.422 to 0.052 as ϵ changes from 0.1 to 1.

The change of Average-difference corresponding to different values of ϵ in the social network is shown Figure 4.4(c). In this real network, the Average-difference decreases from 1.2% to 0.05% with the increase of ϵ from 0.1 to 1.

The above result shows the effectiveness of the proposed mechanism which preserves agents' privacy in online social network auction environments. However, the experiment

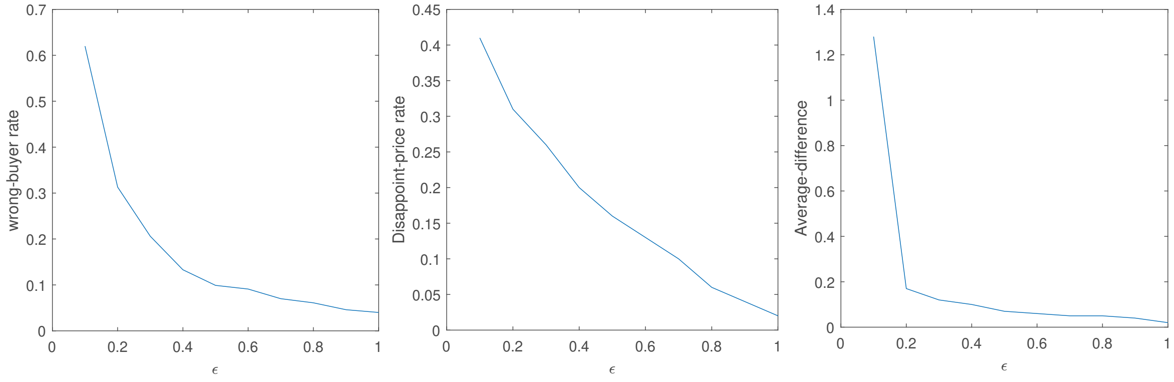


Figure 4.5: Comparison results of three metrics with the increase of privacy budget.

will run 1000 times for each value of ϵ. In the next section, we will set a termination-price to show that the valuation of agents in an online social network auction affects the metrics of average-difference, wrong-buyer rate and disappoint-price rate.

4.7.2.3 Influence of seller in the auction

In this section, we further examine the performance of the proposed method in relation to the agent's bid. We set a termination-price in our experiment. This parameter means that the seller will close the auction when they receive a satisfactory price. In this section we set the termination-price to 1050. The results of the three metrics are shown in Figure 4.5.

The wrong-buyer rate corresponding to different values of q for the social network is shown in Figure 4.5(a). As we have shown in this sub-figure, the wrong- buyer rate in this social network decreases with the increasing of parameter q: the wrong-buyer rate is 0.620 when ϵ is 0.1 and it decreases to 0.04 when ϵ is at the value of 1. The Disappoint-price rate corresponding to different values of ϵ on the social network is shown in Figure 4.5(b). The result shows that it decreases from 0.413 to 0.040 while ϵ ranges from 0.1 to 1. Change of Average-difference corresponding to different values of ϵ in the social network is shown Figure 4.5(c). It decreases from 1.28 percent to 0.02percent while ϵ changes from 0.1 to 1.

The result of this experiment shows that all three metrics are lower than the result where we did not set a termination-price. This phenomenon shows that setting a termination-price allows the seller to more accurately find the agent who gives the highest valuation price and decrease the payment to agents spreading the auction information. However, if the seller sets the termination-price at a very high value, it has little effect. How to choose a reasonable termination-price is a tested work for sellers.

4.8 Summary

In this section, we address a common limitation of existing auction mechanisms in online social networks by proposing a privacy-preserving auction mechanism and designing an entire auction process. Our mechanism utilizes the propagation of information in online social networks and the technique of differential privacy. By incorporating these elements, we can develop a new auction model that is suitable for online social networks, while also preserving the users' neighbourhood information and bid values, without compromising fairness or performance.

Moving forward, we plan to enhance our mechanism in two ways. Firstly, while our mechanism is based on the theory of information propagation in social networks, it currently only incorporates a simple model that does not consider the complex behavior of individuals in online social network environments. By accounting for these nuances, we can improve the model's accuracy and relevance. Secondly, our experiments were not conducted on the basis of auction theory, and in the future, we aim to develop an auction model based on auction theory to further refine and optimize our mechanism.

Part II

Privacy Diffsuion Analysis in Social Media

CHAPTER 5

RECONSTRUCTING THE PRIVACY INFORMATION DIFFUSION PATHS AND COMPARISON

In Part I, we design two auction mechanism for social media users while preserve users' privacy information in online social media. The diffusion process of auction privacy is a illustration of privacy information diffusion in online social network. Therefore, In Part II, we expand the auction privacy diffusion to all the privacy information in social media to analyse how the privacy information diffuse through the social media. In the following parts of this thesis, we will focus our attention on the following issues: 1) How the privacy information diffuse in social media, 2) How to model the privacy information diffusion in social media, 3) How to block the privacy information diffusion.

5.1 Introduction

Social media has brought us cheap, convenient communication and its enormous growth reflects the impact it has had on our daily lives [13]. However, as users share their personal information, privacy information can also be revealed, even if unintentionally, which can then also spread through the network [143]. For example, forgetting to turn off location sharing during a visit to a medical specialist may reveal privacy information about one's health that could be published or sold. In fact, privacy information does not need to be sensitive; it can be any information about an individual that has been published without the person's permission or awareness. Importantly, such leaks present great risks to users of social media and, for this reason, every one from the individual concerned to social media managers to data collectors and curators has an interest in taking measures to stop the propagation of privacy information in social media.

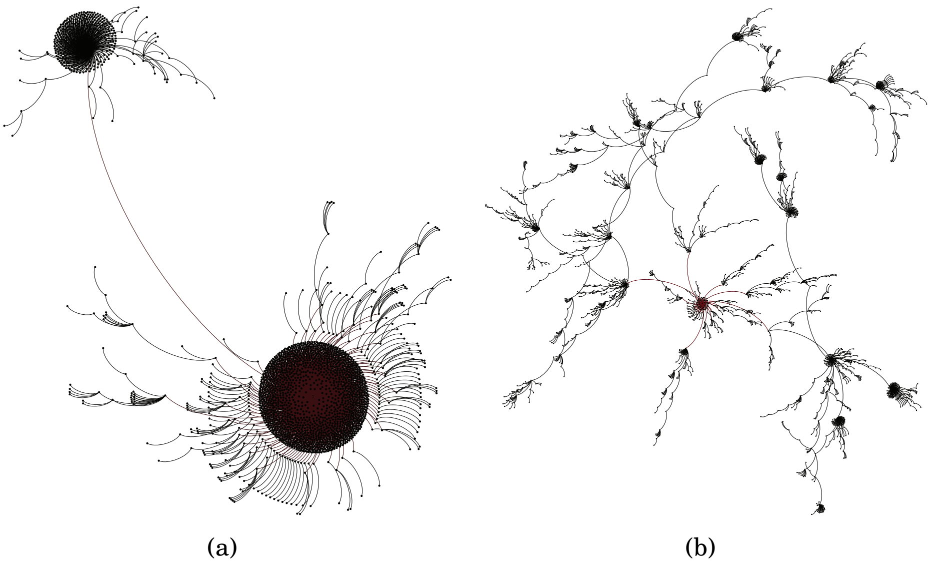


Figure 5.1: An example of diffusion paths on normal news and privacy information.

However, to preserve privacy efficiently, we first need to identify the features of privacy propagation. The closest studies on this subject fall into one of two categories: sensitive information diffusion and propagation modeling. Studies on how sensitive information spreads typically focus on rumormongering. And, most often, the solution to stopping their spread is to counter the rumor with truth [39] [143]. However, while this might work to constrain hearsay, it is not a suitable remedy for privacy breaches. Privacy information may differ significant from news or rumors. First, the paths by which privacy information propagates might have a unique structure. For example, normal news is often published by traditional media outlets, which have many followers, while private data is likely to stem from users who do not exert a strong impact on the platform in isolation. Second, not everyone is interested in a given individual's privacy information, so, unlike news or a celebrity rumor, it is probably wrong to assume that the information will spread throughout the entire network. Exactly how privacy information does spread through social media needs to be studied empirically and compared with what we know about the diffusion of other types of information. To illustrate these difference more clearly, we have selected two examples which we have shown in Figure 5.1. One is normal news and another is privacy information. Each of them shows the typical propagation features scenario in online social network.

Example 1: Sub-figure 5.1(a) is an example of normal news which was given by a news media. It gives a list of dangerous good in Amazon. Intuitively, we can find that most propagation process centered at a few nodes in the graph. This means that most users get this information from only a few users on social media and the propagation path present the dandelion structure.

Example 2: Sub-figure 5.1(b) shows a propagation path of privacy information in online social network. This privacy information exposed that one of the American chief justice, Brett Kavanaugh, is a drunk in leisure time. Compared with normal news in Sub-figure 1(a), the propagation path present a more divergent structure. Users got this information from various sources.

These two sub-figures illustrate that there are big differences between normal news and privacy information in the propagation process. we began by collecting dataset from Twitter API. We collected several spreading information in online social media and used them to construct the spreading process of the information in social media into the graph structures. These graph structures show the propagation path of different information in social media. We then used the resulting dataset to build a graph convolutional network (GCN) that captured the propagation paths of different types of information in social media. What we found was two key differences in the features and flows of normal news and private data. First, normal information often diffuses fastest at the first hop of the propagation path, while privacy information. diffuses fastest at the second or third hop. Second, normal news emanates from highly impactful nodes; privacy information does not have an obvious center of origin. To address the last challenge, we considered these findings in light of some of our existing preserving mechanisms.

These two sub-figures illustrate that there are big differences between normal news and privacy information in the propagation process. In this chapter, we take a first look at this problem and, in doing so, identify three major challenges with studying real-world cases. 1) How does one delineate privacy information from the clutter of other information on social media? Topological features may be a good starting point for this. 2) How can one identify the propagation paths of privacy information? With no previous research to rely on, this is greenfields exploration. However, with the privacy information identified, we can trace its flows and see whether there are unique patterns to its diffusion. 3) Once identified, what are the features of the information or the information's diffusion process that will help to inform adequate protection mechanisms? This last question is the ultimate goal in the pursuit for privacy preservation.

To tackle these challenges, we began by collecting dataset from Twitter API. We collected several spreading information in online social media and used them to construct the spreading process of the information in social media into the graph structures. These graph structures show the propagation path of different information in social media. We then used the resulting dataset to build a graph convolutional network (GCN) that captured the propagation paths of different types of information in social media. After observing the real world data, we claim that the spread of privacy information is different from the normal news in online social network. The key difference is that the normal news are often published by the traditional media which have large amount of followers; the privacy data spread through the users which do not have high impact on social media. Therefore, the normal news and privacy information on social media are different in propagation structures and features. What we found was two key differences in the features and flows of normal news and private data. First, normal information often diffuses fastest at the first hop of the propagation path, while privacy information. diffuses fastest at the second or third hop. Second, normal news emanates from highly impactful nodes; privacy information does not have an obvious center of origin. To address the last challenge, we considered these findings in light of some of our existing preserving mechanisms, discussing the insights in Section V. The main contributions of this case study therefore include:

1) An investigation of the differences between how privacy information propagates across social networks versus conventional news. We compare the two in terms of both topological features and diffusion paths.

2) A new method of network reconstruction that traces information flows. We proposed a framework to recurrent how a message diffuses through the online social media by using the follower-followee relationship and the time which the message has been published

3) An extended variant of the GCN classification algorithm for distinguishing between news and privacy information. We apply the Graph Convolutional Network (GCN) into our framework to recognise the normal news information diffusion cascade and the privacy information diffusion cascade.

4) A privacy preservation method for social networks that blocks a number of users to stop the propagation of privacy information. We try to block the privacy information diffusion by limit the influenctial users in social media.

5.2 Related work

The study of privacy preservation in online social network has attracted much attention. However, the study of privacy propagation in social media continue to have gap. Much of this due to the lack of privacy propagation data in social media. The following section describes three research direction which related to this area.

With the development of social media, information diffusion in online social network has attracted a lot of attentions in the past decade. Studying the diffusion of information could mainly be divided into two aspects. The first aspects is to build a propagation model to simulate how the information propagates through social media [39, 57]. In such aspect, some researchers built a diffusion model which has similar topological structures with the real information diffusion in social media [72]. Some researchers focused on modeling different information diffusion in online social network [57, 60, 72, 92]. In 2019, X.Wu *et.al* proposed an adaptive diffusion of sensitive information model. In this chapter, they minimized the sensitive information diffusion while preserving the diffusion of non-sensitive information by modeling the sensitive information diffusion size as the reward of a bandit [30][21]. The second aspect in this area is to discover the real diffusion path of information in social media and modeling them into a complex network, then find the topological structures difference from other kinds of complex networks. In this chapter, we followed the second aspect and improved the methods which rebuilt the propagation path of information in social media. Then we compared the difference in two categories of propagation path and classified them by using the Graph Convolutional Network (GCN).

In the current study, researchers have shown that the sensitive information are different from the normal news in social media [130]. In 2018, S.Vosoughi *et al.* investigated the differential diffusion of verified true and false news in social media and proved that the false news in social media are often more novel than the true news in social media [103]. However, the study of the difference between the privacy information and normal news is still a gap in this area.

Analysing the sensitive information in online social network is not new, however the study of privacy information propagation in this area is very limit. T.Zhu [143] first proposed the dynamic privacy propagation model in social media. In this chapter, Dr. Zhu assumed that the privacy information is different from other information, e.g. the normal news or the rumors in social media for the reason that the privacy information is accumulated. Therefore Dr.Zhu creatively adopted a accumulation mechanism into the propagation model. In her work, the privacy information of one user in social media has a certain value and part of this user's privacy information will be spread to other users in social media at one time with a certain probability. Therefore, we may take a more deep analysis of the privacy propagation in social media with a certain definition of privacy information and the significant dataset. In addition, she proposed two categories of privacy preservation mechanism, i.e. the centralized mechanisms and the decentralized mechanisms to prevent propagation in social media. In this chapter, we take the decentralized mechanisms in stopping the privacy information propagation in online social network.

5.3 Methodology

The collected data from Twitter is a fictitious diffusion paths for all the diffusion paths points to the original users who publish the tweets in social media. Therefore, this chapter gives a methodology to reconstruct the diffusion paths of information through the online social media. There are several methods for tracing the diffusion of information on social media ‚Äì Goel *et al.*'s being one method for reconstructing the path of tweets for Twitter. While good, this method could be further improved. For example, the method relies on follower-followee relations but does not consider that messages might be retweeted by a non-follower. In other words, a user could have simply stumbled across a message they were interested in and shared it. Also, users form communities and one user, say User A, may exist in the follower lists of several other users, some or all of whom might have retweeted the message User A saw and then retweeted themselves. So, how does one choose which message user A ultimately shared? 3) The method does not account for quoted text which is retweet and comment on one message in Twitter. It is a definitive indication of a forwarded message from one user to others.

When people collect both normal information propagation and the privacy information propagation data from social media data, the collected data do not contain any propagation path. The first step is to reconstruct the propagation path, so that the features of the propagation will be disclosed. There are several ways to achieve this, one of the most popular methods has been proposed by Goel *et.al* [120], who used the follower-followee relationships between different users in social media to recover the propagation path in online social networks. However, this method exhibits some weaknesses when rebuilding the privacy propagation path: 1) some users may not be in any of the user's follower list in the previous propagation path. They just discovered the messages in which they were interested and shared them. 2) some users may be in more than one user's follower list in the previous propagation path. How do they choose the message user they share? 3) the method did not consider the quotes in social media.

To overcome these weaknesses, we proposed an improved method to make the propagation path more reliable. In general, this method of reconstructing the propagation path is accomplished in the following two steps:

1) Set up a sharing list: this step is to sort and number all the users who have shared an identical message in social media in chronological order. The user who published this message will be numbered as number 1. The first user who shared this message will be numbered as number 2. Then, the following users' number is 3,4 and so on. It is the sharing list of one message in Twitter. Obviously, one user shared this message from the users whose number is smaller than him/her.

2) Construct a propagation graph for [one message only / each message]: In order of the sharing list, Node 1 is the original publisher (#1) with an edge to the first user to share the message (#2). From user #2 onwards, i.e., #3 and greater, edges are drawn according to the following rules:

- If a User X_n is in one and only one follower list from X_{n-x}, i.e., User Y , we deem that user #X retweeted the tweet from user #Y.
- If User X_n is in multiple follower lists from X_{n-x}, we assume that #X retweeted the user who was the most immediate predecessor on the list that they follow. This is in keeping with the way Twitter orders activity and would be a natural action that did not require scrolling.
- If User X_n is not in any follower lists from X_{n-x}, we assume they simply saw the message somewhere and shared it. By Twitter's own recommendation, a message with a high number of sharing, we assume that users X are more likely to share the message with a high number of sharing. From this assumption, we infer that user X in this situation will share this message from the previous sharing list of users with a certain probability. This probability is determined by the number of sharing by each user.
- If User X quotes text in their message, a directed edge is constructed from the quote to the quoter.

Taking Twitter data as an example, Figure 5.2 shows the example of our reconstructive methods. Sub-figure 5.2(a) is the sequence of 6 retweet users in chronological order.

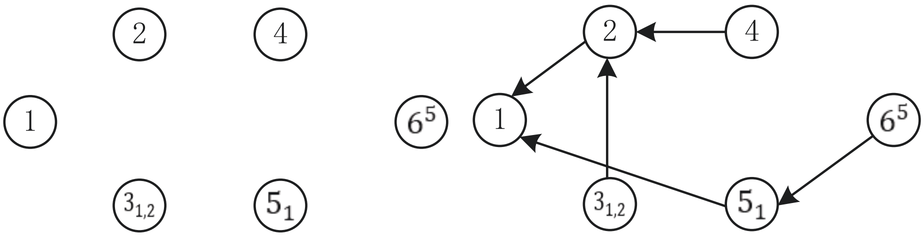


Figure 5.2: An example of our reconstructing propagation path in Twitter data.

This tweet and the 6 retweet relationships are all collected from the Twitter API. The content in each node is constructed by three parts: 1) the main part with a number shows the order of retweet in Twitter. 2) the subscript part with several number refers to the follower-followee relationship among these nodes. 3) the superscript with a number means the quote relationship in Twitter. 4) if a number in nodes has neither the subscript nor the superscript, it means that this user is not the follower of any other users. He just saw this tweet and retweeted it. We use $3_{1,2}$ as an example: '3' represent that this user is the third one to retweet the tweets in Twitter and the subscript '1,2' refers that this user is a follower of No.1 user and No.2 user.

Sub-figure 5.2(b) is the propagation path we rebuilt by the above four rules. For the No.2 user, the first user who retweet this tweet, he can only retweet this tweet from the No.1 user, i.e., the original user who publish this tweet. The No.3 user has two followees: No.1 and No.2. He/she should retweet this tweet from No.2 user according to the second rules. No.4 user is not the follower of the previous three users. Therefore, the probability of retweeting the No.1 user and the No.2 user is equal to 1/2. The No.5 user is one of No.1 user's followers. He retweets this tweet from the No.1 user. No.6 quoted this tweet from the No.5 users so the retweeted edge points to No.5 user.

5.4 Reconstruction evaluation

5.4.1 Dataset

There are no existing dataset about the privacy propagation through social media. In such situations, we try to collect data from the Twitter to achieve our research purpose. We used Twitter historical API to collect over 1 million tweets in English languages from October 20th 2020 to November 30th 2020 as our main dataset. we then collected the related information of these tweets. The related information mainly contains the following three kinds of information in Twitter: 1) the original tweets and the users who published these tweets, 2) the users who retweet these tweets and 3) the users who reply to these tweets and the reply content.

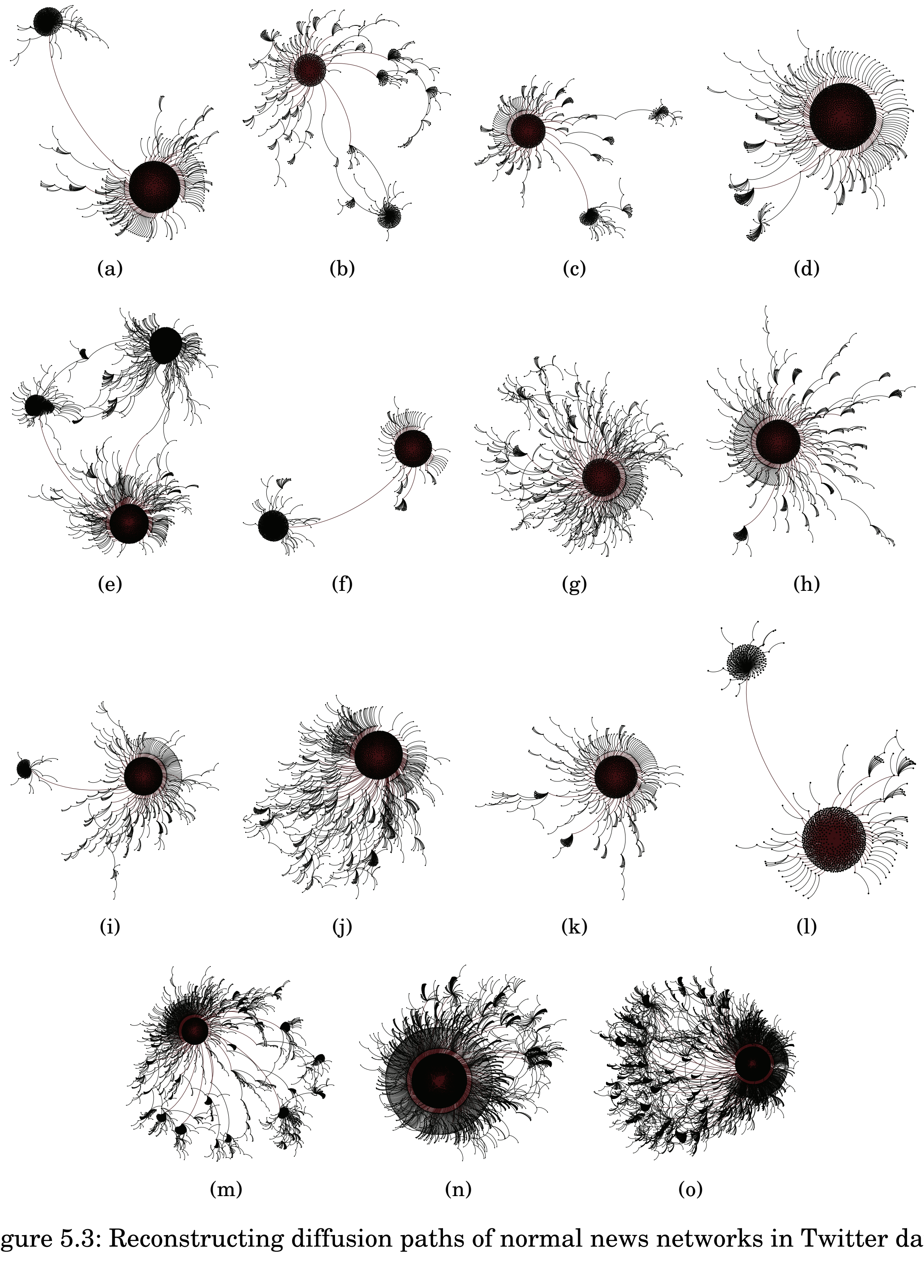
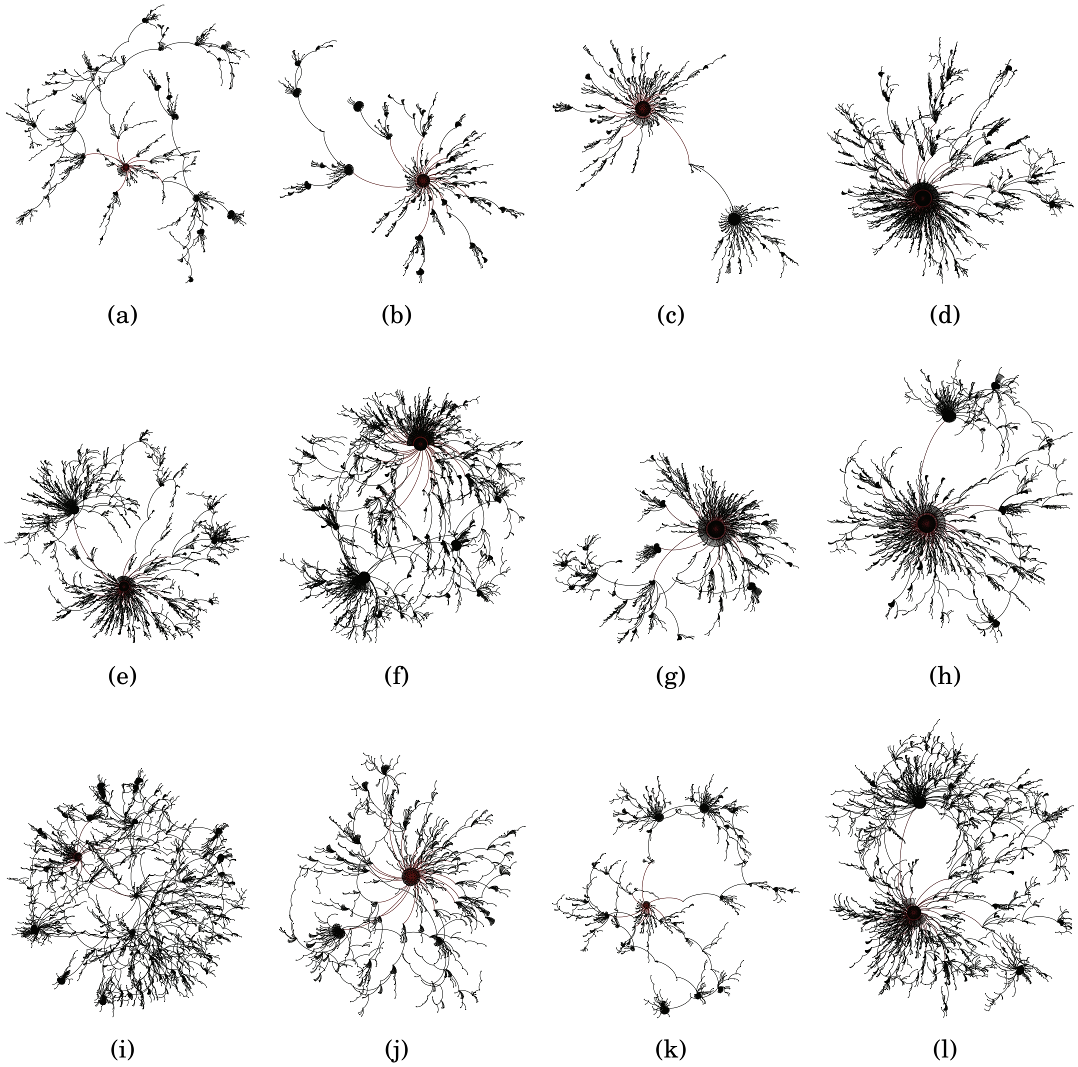


Figure 5.3: Reconstructing diffusion paths of normal news networks in Twitter data.

5.4.2 Selecting of normal information and privacy information

The concept of privacy was proposed by Louis Brandeis and Samuel Warren in 1989. In their book 'The Right to Privacy', they claimed the privacy is the right to be let alone'. However, because the elements of individual privacy have constantly changed in the past century, it is difficult for researchers to give a strict definition of privacy in the academic area given all that highlight one aspect of privacy.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 5.4: Reconstructing propagation path of privacy information diffusion paths in Twitter data.

In social media, privacy has multiple meanings in different situations and data format. Hai Liang *et al.* believed that the privacy in online social network is the individual information of users with negative consequences through leakage. Therefore, in this thesis, we consider privacy information as the sensitive information containing the individual information of specific users which is published on social media without the permission of these users. However, as we mentioned before, the elements of individual privacy are constantly changing. It is difficult to automatically detect the privacy information in social media. As a result, all the privacy information and normal information in the following of this thesis were manually screened by us. Then we used the Twitter API to collect their related information.

5.4.3 Reconstructing propagation path

For the purpose of finding the difference between the normal news propagation and the privacy information propagation, we try to reconstruct these two categories of propagation path in social media. To reconstruct the propagation path of information in social media, the first work is to find the privacy information and the normal information in social media. In the above section, we have defined that privacy is the individual information which is exposed to the public without their permission. We filter all the original tweets in the dataset, then selected the privacy information and the normal news from these tweets artificially. As we have discussed before that it is difficult to automatically detect the privacy information in social media. We manually screened 15 normal news and 12 privacy information from these original tweets. To evaluate these two categories of propagation path objectively, we set the number of users each tweet propagated at a similar standard from 1000 to 7000. We then use the reconstructing method, which we discussed in section IV, to reconstruct the propagation path graph. The structures of these reconstructing path are provided in Figure 5.3 and Figure 5.4.

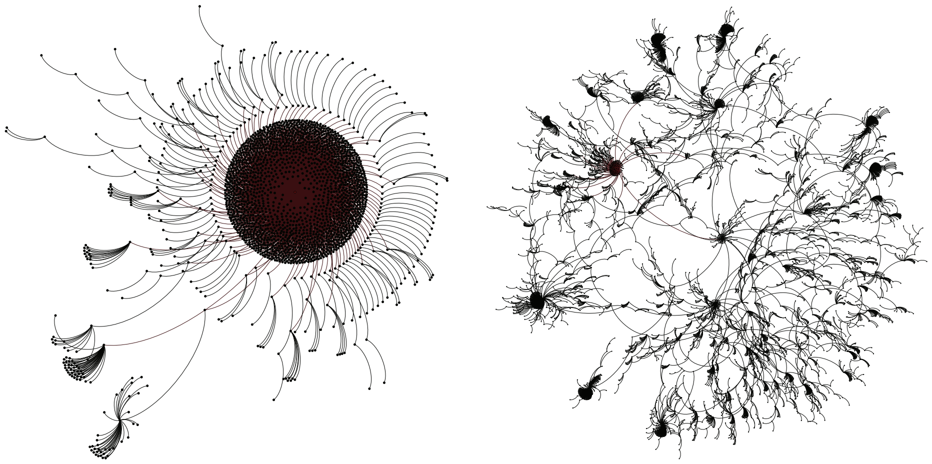
Figure 5.3 shows the propagation structure of our 15 normal information included in our dataset in online social media. The normal news propagation paths have two features: 1) the normal news often diffuses only a short distance in social media, all of the 15 privacy information propagation path diffused at most 7 hops in social media 2) most of the users who received the message from only one or two users in the social media. This means that the users who are affected by this message are often one or two hops from the original users who publish this message The propagation structure of the 12 components of privacy information are shown in Figure 5.4. Compared with the normal news, we can intuitively find that the privacy information paths have more complex structures in the following two aspects: 1) the privacy information often diffused deeper than the normal news in social media, all of the 12 privacy information propagation path diffused at least 8 hops in soical media 2) All the network structures show that the privacy information cascades are more divergent for the users who received this message did not center at one or two nodes in the propagation paths. The above sub-figures have shown that these two categories of propagation paths have big differences in topological structure. In the following, we will discuss this difference in several metrics.

5.5 Comparison of normal news and privacy information diffusion

The next challenge is to analyze the difference between the private information propagation and the normal information propagation. We first compare the two categories of propagation graph in several parameter of complex networks, then propose a machine learning method to classify the private information propagation from other kinds of information propagation.

The privacy information diffusion differs greatly from the normal news diffusion in propagation graph structures. Figure 5.5 is a case example of these two categories of network graph structures we collected from the Twitter data. Among these two figures, sub-figure 5.5(a) is the normal news diffusion which describe the Covid-19 spread in American and sub-figure 5.5(b) is the privacy information diffusion which gives an interesting things that Trump's SAT score in his high school. These two diffusion paths are different even seen by the naked eye. We can intuitively conclude from these two sub-figures that the privacy propagation path has more *star structure* in the network while the normal information propagation path has fewer. However, the star structures in privacy diffusion path are smaller than those in the normal information diffusion. Meanwhile, the users who are infected by the normal news are surrounded around the users who publish this message in the topological graph of propagation paths while the topological graph of propagation paths for privacy information has a decentralised distribution. There are several central nodes in the topological graph. This different topology is aligned with our claim that the normal information often diffuses fastest at the first hop of the propagation path while the privacy information diffuses fastest at the second or third hop. These key differences on the topology make it possible to classify normal news with privacy information. In the next section, we will use graph classification to separate the privacy information from the normal information.

However, the above conclusion is just a preliminary conclusion in this topic of conversation. We will provide a deep analysis of these two categories of network in this character.



(a) (b)

Figure 5.5: Examples of diffusion path.

5.5.1 Comparison of topological properties

The complex network provides a method to represent different complex system. The topological properties show the different connection relationships and the basic structure of a network. In this thesis, we will use three topological property metrics to compare these two categories of network in the following :

- **Average_degree:** degree is the basic features of nodes in complex network. The degree of one single node i in complex network is the number of nodes which connected directly to node i. Average_degree is the average value of all nodes in complex network:

$$Average\_degree = \frac{\sum_{i=1}^{n} D_i}{n} \tag{5.1}$$

  where D_i is the degree of each node in the graph, n is the number of nodes in the graph.

- **Density:** This metric is the ratio value of existing number of edges and the possible maximum number of edges in the graph. It shows the compactedness of edges between different nodes in complex network:

$$Density = \frac{2M}{N(N-1)} \tag{5.2}$$

where M is the number of edges in the graph, n is the number of nodes in the graph.

- **Degree_distribution:** set p_i is the probability of i-th degree nodes in complex network. The degree distribution is a probability distribution of all degree in complex network

$$p_k = \frac{N_{D_k}}{N} \tag{5.3}$$

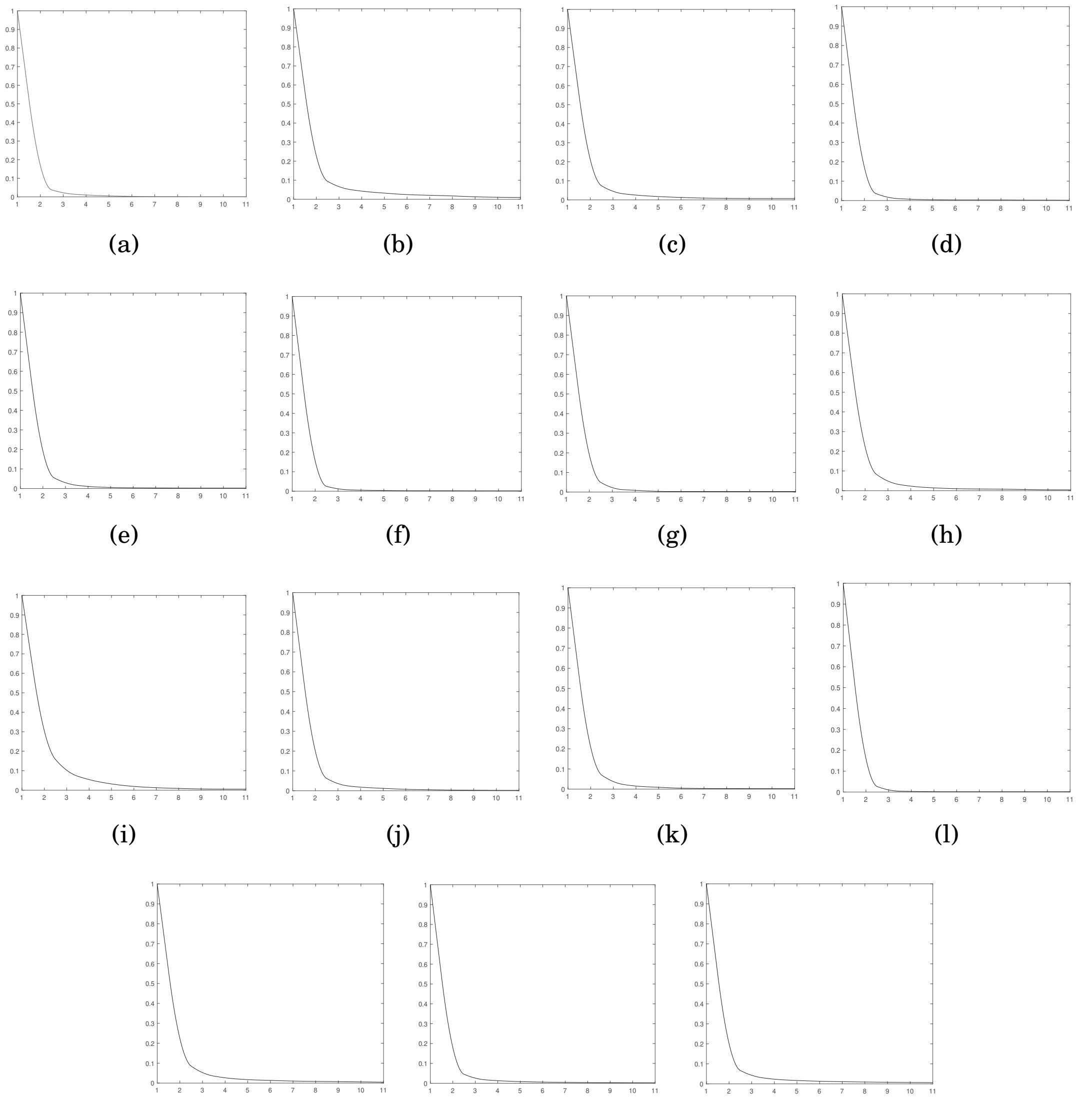
$$Degree\_distribution = \{p_1, p_2 ..., p_m\} \tag{5.4}$$

where N_{D_k} is the number of nodes whose degree is k in the graph, n is the total number of nodes in the graph. m is the biggest degree in the network

Table 5.1: Attributes of complex network

Category	Average degree	Density	Power exponent
Privacy	0.945	0.001	3.52
News	0.996	0.003	2.55

Table 5.1 is the result of the average degree and density for our two categories of networks. The average of the privacy propagation path is 0.945 and the normal news propagation path is 0.996. The density of these two categories of networks are 0.001 and 0.003. Figure 5.6 and Figure 5.7 is our 27 samples of information diffusion path in degree distribution by using the CCDF (Complementary Cumulative Distribution Function). Half of them are the normal news and the other half is the privacy information. In these sub-figures, the X-axis is the degree of nodes in each graph and the Y-axis is the probability of nodes in different degrees in the graph. We can conclude from these 27 sub-figures that both of these two categories of networks a obey the power-law distribution (or long tail distribution) [5] in that most nodes in the graph have few neighborhood while a few nodes have a lot. Therefore, we fit the degree distribution of each network into the power-law curve and calculate their power exponent. The power exponent of privacy information propagation path is 2.53 and 3.52 is the normal news propagation path. This means that the nodes in normal news propagation path are more likely to connect to the nodes with high degree. It is consistent with our observations that the star



(a) (b) (c) (d)

(e) (f) (g) (h)

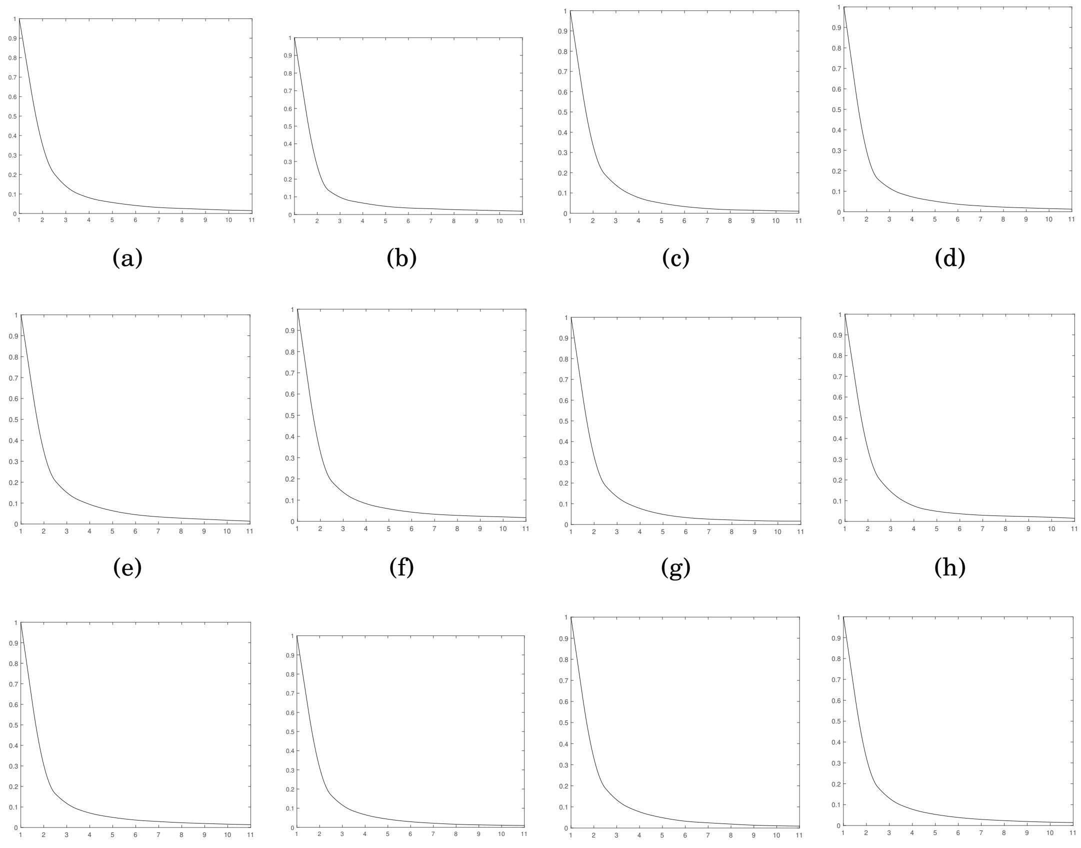
(i) (j) (k) (l)

(m) (n) (o)

Figure 5.6: Degree distribution character of normal information propagation networks.

structures in normal news propagation path is bigger than in the privacy information propagation path.

However, these three metrics of complex network show that these two categories of networks have similar results in the basic attributes of traditional topological structure features and the results could not show the difference between these two categories of propagation path.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

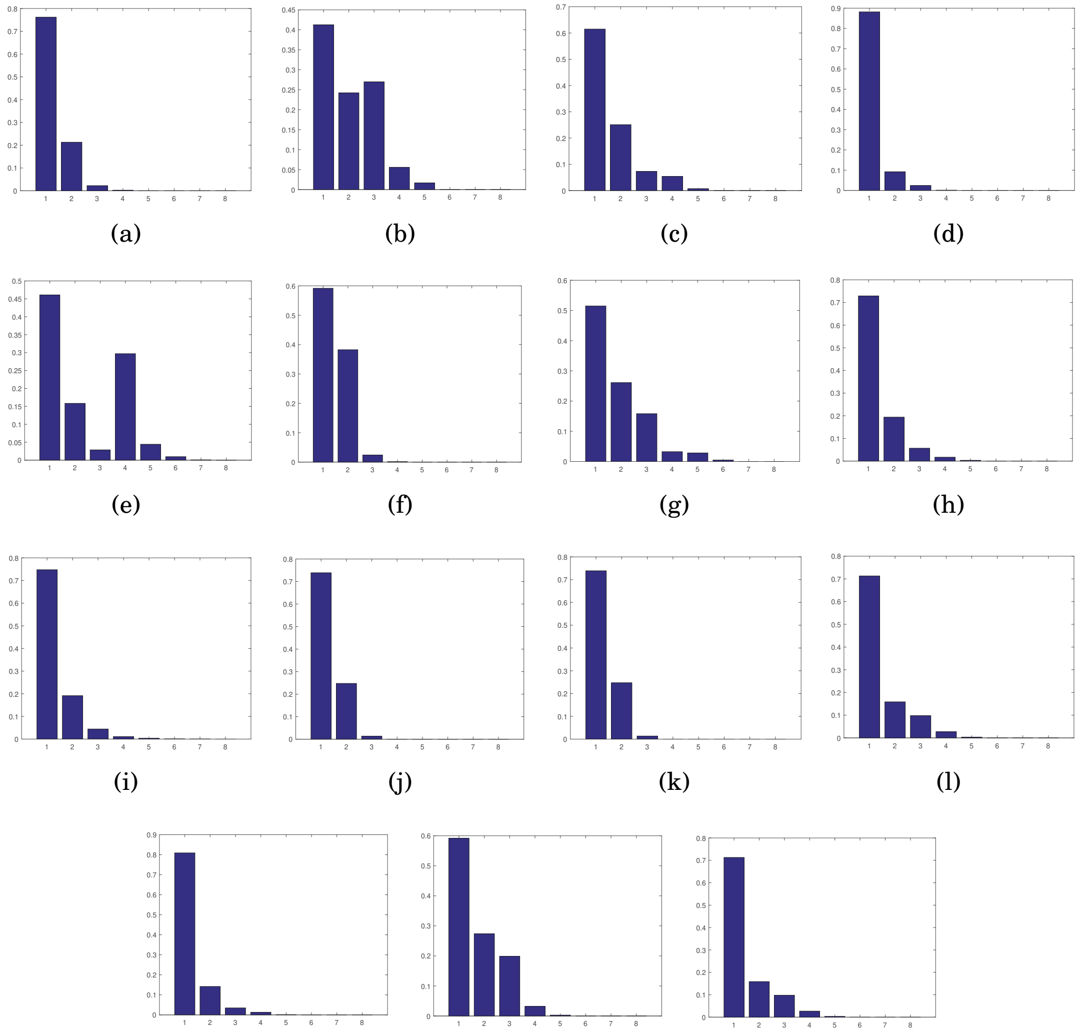
Figure 5.7: Degree distribution character of privacy information propagation networks.

5.5.2 Comparison of Propagation property

To have a deeper analysis of these two categories networks, we applied the second aspects of metrics to show their difference. These metrics are the propagation features of information in a social network. They show how information propagated and the propagation difference among the social network. In this thesis, we will compare these two categories of network in the following three metrics:

- **Average_pathlength:** consider two nodes i and j in the graph, the distance between these two nodes is the shortest path of node i and j. Average-pathlength is the average distance of every two nodes in the graph.

$$Average\_pathlength = \frac{2\sum_{i=1}^{m} D_i}{N(N-1)} \tag{5.5}$$



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o)

Figure 5.8: The propagation maximum branch of 15 normal news propagation networks.

where D_i is the distance of every two nodes in the graph. N(N-1)/2 is the number of nodes pair in the graph.

- **Propagation_depth:** the number of hops from the origin nodes over to the furthest distance nodes in the propagation path graph. It shows how further one information is propagated in online social media.

- **Propagation_maximum_branch:** in the propagation depth parameter, information may propagate multi hops during the propagation process and the number of nodes in each hop is different. This metric shows the percentage of users involved in each single hop of the graph among all the nodes in the graph.

Table 5.2: Propagation features

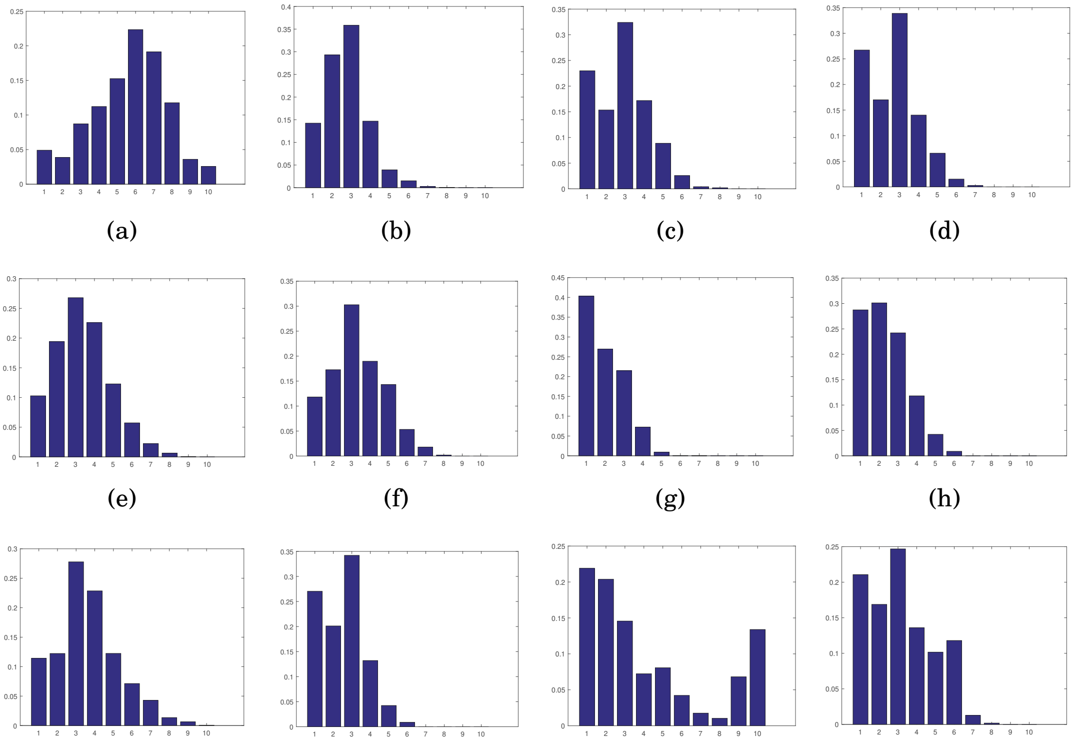
Category	Average pathlength	Propagation depth
Privacy	5.23	10.42
News	3.45	6.57

Table 5.2 is the result of the average path length and the t propagation depth of the two categories networks. The average path length of the privacy propagation path is 5.23 while the average path length of the normal information propagation path is 3.45. Obviously, the average path length of privacy propagation networks is longer than the normal information networks. The same situation happens in the propagation depth. The propagation depth of the privacy information networks is deeper than the normal information networks. Figure 5.8 and Figure 5.9 show 27 samples of both the propagation maximum branch in the privacy propagation and normal information propagation in social networks. In these two figures, the X-axis is the hops of each piece of information propagated in the graph and the Y-axis is the percentage of nodes in different hops through the graphs. We can see from the Figure 9 that most users often obtained this information from original users who published it and the number of users this information propagated is decreased as the hops increased. However, this situation is different in the privacy information propagation. In Figure 5.10, the peak value of the number of users who propagated privacy information often appeared in the second or third hop of the propagation path. In our experiment, the users who obtained the normal information from the first hop of the graph occupied over 60 percent while in the privacy information it occupied 60 percent of the second and third hops.

5.5.3 Star structure comparison

We can intuitively observed from figure 5.8 and figure 5.9 that the star structures are different in these two categories of network. In this part, to deeply analyze the difference of star structures in normal news diffusion and privacy diffusion, we compare the size of the biggest star structure and its proportion in complex network to show their difference. We first give definition of star structures in the graph data, then give the comparison.

Star structure analysis: Figure 6.4, as we discussed before, shows the examples of normal news and privacy information propagation path. Obviously, the nodes of privacy information cascade not only have a high proportion in the high-hops from the original information publisher but also show more small star clusters far from the information publisher while there are only a few large star clusters in the news information cascade.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 5.9: The propagation maximum branch of our 12 privacy information propagation networks.

In this section, we will study the star structure in the cascade in order to more accurately describe the characteristics of the privacy information cascade. We define the star structures in complex network as the following:

Definition 1 Out-degree centrality and central nodes: For a directed graph, the out-degree centrality of a node in graph is the ratio of its out-degree and the possible maximum out-degree:

$$ODC_i = \frac{k_i^{out}}{N-1} \tag{5.6}$$

where k_i^{out} is the out-degree of node i in graph. N is the number of nodes in the graph and N - 1 is the possible maximum out-degree of a node in the graph. The central nodes are the center node of the star structure. In a information cascade G with n nodes, the node is defined as the central node of a star structure when:

$$c \in V, ODC_C \geq \theta \tag{5.7}$$

where V is a set of nodes in graph G. θ is a threshold which could reflect the size of the star structure around the central node.

Table 5.3: Biggest star structure in the path

News	Size	Proportion	Privacy	size	Proportion
AmazonDangerous	1876	76.17%	BrettKavanaugh	90	4.90%
AmyNominationApproved	400	46.45%	CharactersofCOVID-19	224	14.25%
ApplePullFortnite	674	61.44%	HunterFailureRegister	337	22.97%
BreonnaTaylorIsKnownMore	2043	88.14%	HunterFBIinves	1601	26.71%
CoronavirusVaccineafew	2238	46.07%	IvankaTrumpSkimFund	456	10.28%
CovsAffectUk	767	59.18%	KimberlyInfected	650	11.82%
IsraelArabReachDeal	1106	51.49%	PemexEmployeesDied	420	40.30%
JoeBidenRespondToTrump	302	89.88%	TomParkerbraintumour	3862	30.39%
aManKilledByCovs	745	51.45%	TrumpSAT	779	11.42%
RobertTrumpDied	1323	72.85%	TrumpStroken	828	27.54%
TedCruzblastsmedia	1564	74.72%	TrumpTheories	706	21.07%
Trumpliedmore	2017	71.27%	TwistedLydaKrewson	4223	21.10%
UsPostalCurtailed	3599	87.70%			
WalesSharpLockdown	8097	59.21%			
WoodlandHitsMilestone	707	94.77%			

Definition 2 Surrounding nodes: Refers to the surrounding node set of star structure. The surrounding nodes of the star structure are centered around the central node, which is composed of special nodes distributed from the central node. The surrounding nodes set with central node C are defined as the following:

$$A_c = \{v \| v \in V, c \to v, k_c^{out} = 0,1\} \tag{5.8}$$

where V is a nodes set of G, $c \to v$ represent the nodes which could be connected to node c. k_c^{out} is the out-degree of node c.

Definition 3 Star structure: A star structure S_c is composed of the central node C and its surrounding node set A_c. The size of the star structure S_c is defined as the number of all nodes constituted the star structure.

Based on the definition of star structure, this chapter analyzes the star structure in normal news cascade and privacy information cascade respectively. We compare the size of star structure in two categories of network in Table 5.3. From this Table, we can conclude that most the biggest star structure in normal news propagation has occupied over 50% while none of the privacy propagation has occupied 50% in the dataset.

From the above three aspects of metrics, we can conclude that the privacy propagation graph and the normal information propagation graph are similar in the topological characters but differ greatly form the propagation features. This phenomenon is due to the following: 1) compared with the people who are interested in privacy information, the number of people who are concerned about the normal news is bigger. Therefore, in the first hop, the normal information will propagate to more users in social media, 2) the normal news in social media is often published by the users who a have high degree in online social network while the privacy information are published by the users who are not very influential in social media. This means that the privacy information will not diffuse to many users in social media at the beginning, but may propagate to more users in the second or third hop and may diffuse deeper than the normal news.

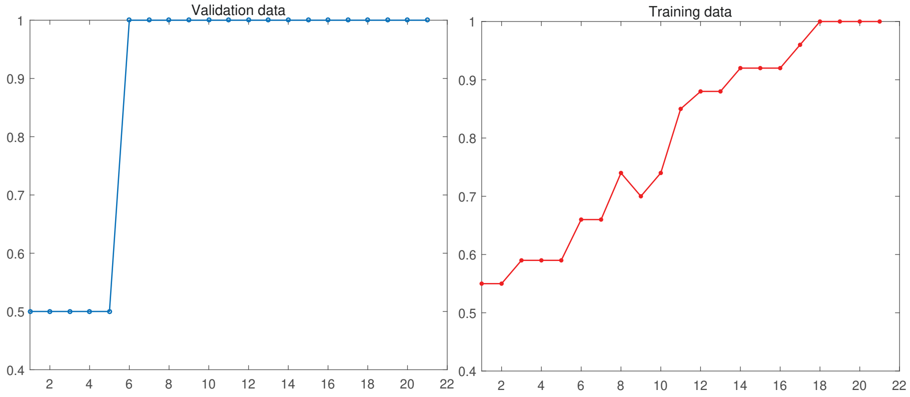


Figure 5.10: The result of our classification through two categories of networks.

5.6 Classification of two categories network by GCN

In this section, we proceed to classify these two network categories by using the famous Graph nerual network(GCN). In our classification experiment, we take half of the network as the training dataset and another as the test. The result of the classification is shown in figure 5.10. The X-axis is the iteration times running in the experiment and the Y-axis is the classification accuracy in percent. The read polyline in the right-subfigure is the result of classification accuracy in training data and the blue polyline in the left-subfigure is the classification accuracy in the valid dataset (i.e. the dataset including the training data and the test data). The accuracy of both the validation dataset and training data is 100 percent after 18 times of iterations. This result shows that the privacy information propagation graph and the normal news propagation graph has a high discrimination in the GCN algorithm.

5.7 Summary

The above analysis has shown the differences between the privacy information diffusion and normal news diffusion in both the basic parameter and the high-order structures. The reasons for these differences are as follows: 1) the normal news are often published by the users who have high influences in social media while the users who release the privacy information are not very influential. As a result, compared with the privacy information, the normal news will spread very fast at the beginning and this phenomenon forms the huge star structures in the first hop of information cascade. Therefore, we can see that most nodes are distributed around the sources node in normal news information cascade and the nodes in privacy information cascade are decentralize in the whole network. 2) compared with the normal news, the social media users are more interested in the privacy information for the privacy information is often more innovative than the normal news [103]. Therefore, they are more inclined to diffuse the privacy information in social media. This phenomenon causes the polycentric sources of privacy information in information cascade and the privacy information diffuse deeper and wider than the normal news. Therefore, the information cascade of privacy information has a longer average shortest paths and propagation depth. Meanwhile, users in social media who got the privacy information are more disparate than the normal news which cause the star structures in privacy information diffusion is different with the normal news diffusion.

CHAPTER 6

A new model of simulating the privacy information in social media and the blocking mechanism

6.1 Introduction

Information diffusion is an essential research area in online social network science [119][50]. Although many researchers have studied and modeled how the information diffuses through the online social network, previous research has mainly focused on the normal information [66][95][76][130] or sensitive information diffusion such as rumor distribution [136][69]. Few of them have discussed the propagation of privacy information through social media while privacy leakage has become an challenge problem for the social media providers [143].

6.1.1 The classical diffusion models and their shortage

The process of information diffusion through the network has always been one of the main research areas in the field of social network analysis [58][93]. The phenomenon of diffusion can be observed in many complex systems, such as the diffusion of innovation and information in social networks[93][73], the spread of epidemics in human and animal physical contact networks [101], cascading failures in infrastructure networks [105], the spread of viruses in computer networks, etc. Among the existing models, linear threshold model [36] (LTM) and independent cascade model [48] (ICM) are widely studied. In LTM model, each node has a threshold which obeys the uniform random selection in interval (0,1). Threshold represents the proportion of the node's neighbors that must be active for the node to be active. When the model begins to spread information, if the proportion of the active neighbors of a node is greater than its threshold, the node will become an active statue. On the contrary, in ICM model, the success of u activating node V only depends on the diffusion probability of information from node u to V, and each edge has its own propagation probability. Whether successful or not, the same node will no longer have the opportunity to activate the same neighbor. When no other nodes are activated, the diffusion process will be terminated.

In the past decade, based on the LTM and ICM model, studying the diffusion of information focused on the diffusion of one piece of information, and most studies were based on Sir or SIS models, which are typical representatives of information dissemination in social networks. For example, Su *et al.* studied the information dissemination in microblog by using the epidemic model, which considers the incomplete reading behavior of microblog users [88]. Zhao *et al.* proposed a model to study the dynamics of information diffusion in social media by using the lethe mechanism. A hydrodynamic model describing information diffusion in online social networks is proposed [133]. Some modeled the information diffusion by other mathematical methods. Hu *et al.* believed that the superposition effect of information diffusion originated from influential users in social networks, and proposed a hydrodynamic model to describe information diffusion in online social networks [42]. Firdaniza *et al.* believe that the forwarding behavior of social media users (such as users on twitter) can be regarded as a continuous time Markov model, because it satisfies the Markov property, that is, the forwarding of subsequent users only depends on the forwarding of the current user, so they use the continuous time Markov model to establish the information diffusion model in social media [29]. Some researchers focus on studying the sensitive information in the social media applied rumor propagation model to an online blog platform called livejournal, which proves that online social networks are more vulnerable to rumor compared with the normal information and the impact of information on livejournal may reach saturation when combined the forgetting mechanism with the rumor model [82][132]. In 2019, X.Wu *et al.* proposed an adaptive diffusion of sensitive information model. Wu has minimized the sensitive information diffusion while preserving the diffusion of non-sensitive information by modeling the sensitive information diffusion size as the reward of a bandit.

Although the previous works have focused on modeling the diffusion of normal information or sensitive information (like rumors), while it is hard to model the privacy diffusion in social media for the following reasons:

- Firstly, although previous studies have shown that the privacy information and normal news propagation path differ greatly from each other in topological structures. It is hard to deeply analyse the difference between them.
- Secondly, modelling the diffusion structures of different information in online social media is a foreseeable area. a novel information diffusion model should be discovered to estimate the extent of the new privacy diffusion.
- Thirdly, the third challenge is how to constrain the diffusion of privacy information. In this thesis, we formulate this challenge into a problem that constrained the diffusion size of privacy information while users have good user experience in social media.

The above items point the existing challenge in modeling the privacy diffusion paths through the online social network and prevent social media users from privacy leakage. Current information diffusion mechanisms, such as SI models, LT model and IC model, did not consider the diffusion features of privacy information in the online social network. Previous work on stopping the sensitive information focused on limiting the diffusion ability of key users, which can hardly be applied directly forward stopping the privacy information diffusion.

6.1.2 Contribution

Considering the difficulties and complexities in solving such problem, we first present an example to show how to model information diffuse through social media and discuss the different types of users in the diffusion process. In this chapter we will discess the following three research issues:

- First, the problem of modeling privacy information propagation is discussed, together with the propagation features and parameters of privacy information. This provides a clear problem definition and explains what kind of diffusion paths we need in modeling.
- Second, we illustrate this problem by proposing a new approach to model and analyze the privacy information diffused through the online social network. This

model considers three key parameters: the influence of neighborhood users in online social media, the interest of different users and the time of information publishing.

- Third, we propose a novel mechanism to stop the propagation of privacy information in online social media and prevent the privacy of social media users from leakage. Our mechanism can effectively restrict the privacy propagation in our diffusion model. Substantial experiments have been conducted on real data from both our collection and other public sources. The results of our experiment are evaluated in term of both the affection and the business performance of the users experience in our experiment. It shows that our mechanism can not only constrain the privacy propagation in social media but also provide with an improved experience.

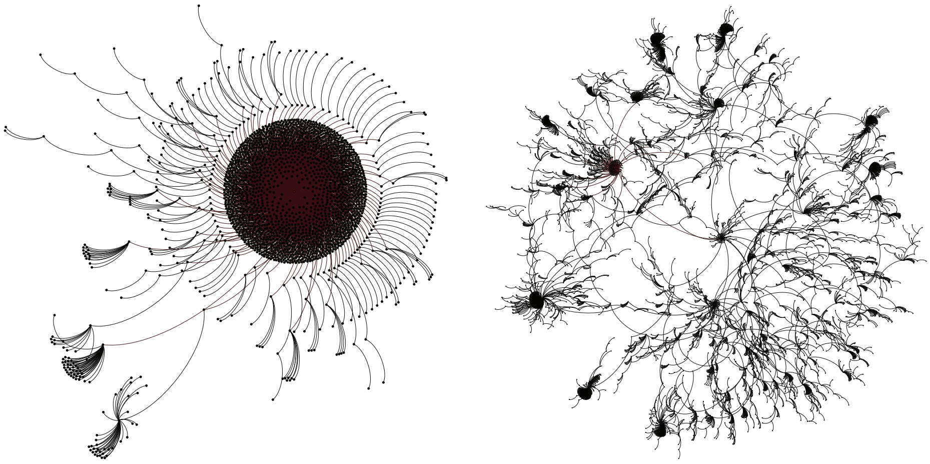
6.2 Related work

Privacy preserving of the Internet has been well-studied by researchers during the past several decades. However, the study of how the privacy information diffuses through social media still remains lacking. However, while the study of modeling the information diffusion in social media has attracted much attention, the modeling of privacy information is very limited. T.Zhu [143] first propose an accumulation model for modeling privacy propagation. She assumed that the privacy information could not be diffused to other people through only one propagation process. Therefore, she adopted an accumulation mechanism and the privacy information has a certain value and part of the it will be diffused to other users with a certain probability in one time. However, her work did not consider the special characteristic of the privacy diffusion in online social network and our work has made up for this point. As a supplement, Hu *et.al* studied the real information propagation path in social media [41]. Their work points out that current studies lack the analysis of the difference between the normal news and privacy information. As a response to this gap, they collected data from Twitter and designed a mechanism to reconstruct the propagation path of the information in Twitter. Their study found that the diffusion paths of normal news and privacy information are similar in topological features but different in propagation features. Normal news often propagates in the first and second hops of the interpersonal network, while privacy information will spread deeper.

6.3 Preliminary

6.3.1 Privacy diffusion analysis

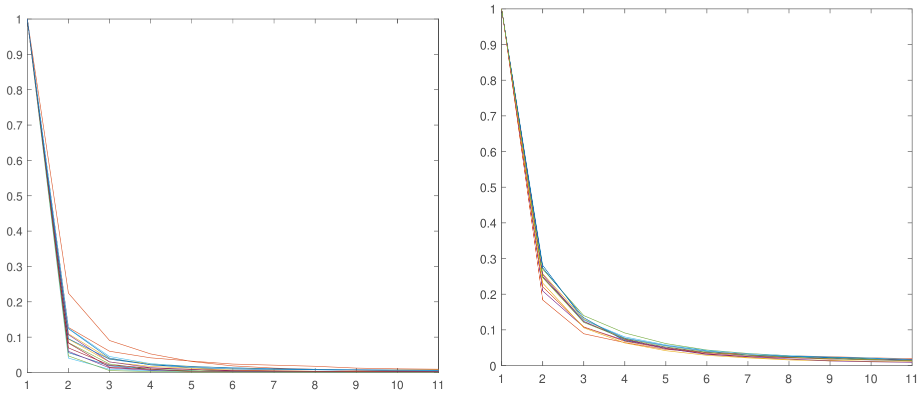
According to the previous study, the privacy information diffusion differs greatly from the normal news diffusion in propagation graph structures. Figure 6.1 is a case example of these two categories of network graph structures we collected from the Twitter data. Among these two figures, sub-figure 6.1(a) is the normal news diffusion which describe the Covid-19 spread in American and sub-figure 6.1(b) is the privacy information diffusion which gives an interesting things that Trump's SAT score in his high school. These two diffusion paths are different even seen by the naked eye. We can intuitively conclude from these two sub-figures that the privacy propagation path has more *star structure* in the network while the normal information propagation path has fewer. However, the star structures in privacy diffusion path are smaller than those in the normal information diffusion. Meanwhile, the users who are infected by the normal news are surrounded around the users who publish this message in the topological graph of propagation paths while the topological graph of propagation paths for privacy information has a decentralised distribution. There are several central nodes in the topological graph.



(a) (b)

Figure 6.1: Examples of diffusion path.

However, the above conclusion is just a preliminary conclusion in this topic of conversation. We will provide a deep analysis of these two categories of network in Section III.



(a) (b)

Figure 6.2: Degree distribution of normal news and privacy diffusion paths.

6.3.2 Problem definitions

Modeling the private information and normal news propagation through online social media is challenging because of its special cases and corresponding analytical tasks. In the following sections of this chapter, we will analyse the following research issues:

- Before we try to model the diffusion path of both the normal news and privacy information, we need to get the real diffusion paths in online social media. In this chapter, we use the methodology and dataset from [41] for further discussion.

- Unlike the normal news which is often released by the users who have high influence, the privacy information is often published by the users who have a certain influence but not with high influence in online social media. Therefore, these two categories of diffusion paths have difference in topological structure. A deep comparison of these two categories of network is needed before modeling their diffusion.

- The privacy propagation features are different compared with the normal news diffusion. The normal news propagation paths have fewer star structures compared with the privacy propagation network. However, the star structures in privacy propagation path are smaller than those in the normal information propagation. These star structures are formed when a node with high degree forwards the message in social media and the follower of this user view it. To construct the diffusion paths which have similar structures with it, an advanced propagation model to simulate the privacy propagation is necessary for us to understand the information diffusion.

- An additional difficult point in this area is about the privacy leakage in social media for the diffusion of privacy information lead to establishing the facts of privacy leakage. It is an essential challenge for both researchers and social media providers. This means that social media providers should propose a new privacy preserving mechanism in online social media. Therefore, how to stop the privacy information diffusion through the online social network is another challenge in this area.

6.4 Information diffusion analysis

This section presents an in-depth comparison of normal news and privacy information propagation. We compared 27 information diffusion paths which we collected from Twitter data in details. 15 of these information diffusion paths are the normal news diffusion and 12 are privacy information diffusion. The dataset we used are from [41]. This dataset are collected from Twitter API. Then we proposed a method to reconstruct the diffusion paths of these information in online social media by using several metrics like: tweets publish time, follower-followee relationship and users' behaviors. By using such mechanism, we have obtained 15 normal information diffusion paths and 12 privacy information diffusion paths. We will first analyse these two categories of network in basic topological structure parameters, then we use the high order structures to deeply describe their differences in high-order sub-structures.

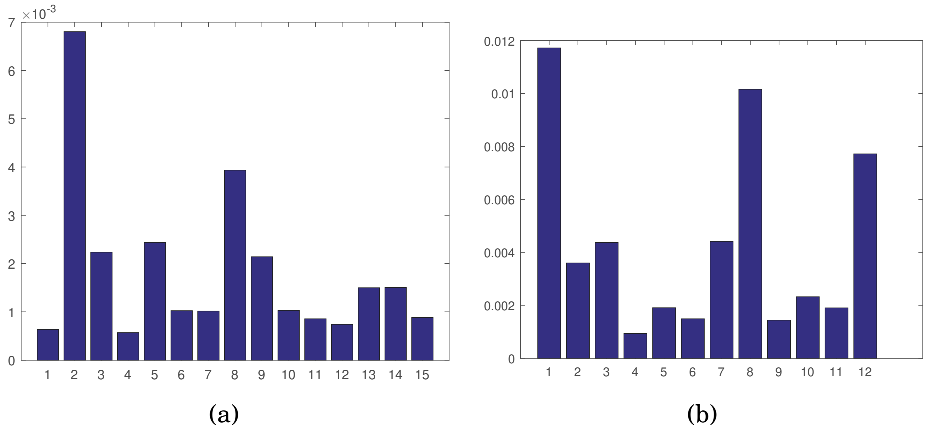
6.4.1 Basic parameters in complex network

As we referenced in Section II, the nodes and edges in complex networks represent different entities and social relationships in real world. Usually, researchers use some basic parameters to describe the topological structure of these networks. These parameters include: degree and degree distribution, Average shortest path and the depth/width of networks.

Degree and degree distribution: In most real network, the degree distribution obeys the power-law distribution (or long tail distribution) as the following:

$$P(k) \propto k^{-r} \tag{6.1}$$

Figure 6.2 is the degree distribution of these two categories of propagation path. These two categories of networks are both scale-free networks which mean that most nodes in the network are of low degree value while a few nodes are of high degree value. Both the privacy information and normal news obey the power-law degree distribution. We can find a few differences from these two figures that the proportion of nodes with only one degree in normal news path is bigger than those nodes in privacy information propagation. This means that the nodes in normal news propagation path are more likely to connect to the nodes with high degree. It is consistent with our observations that the star structures in normal news propagation path are bigger than in the privacy information propagation path.



(a)

(b)

Figure 6.3: Average shortest path of propagation paths.

Average shortest path: Set d_{ij} is the shortest distance (*.i.e* the minimum hops) between two different nodes i and j in complex network. When two nodes are not connected, $d_{ij} = 0$. For a graph with N nodes, the Average shortest path is:

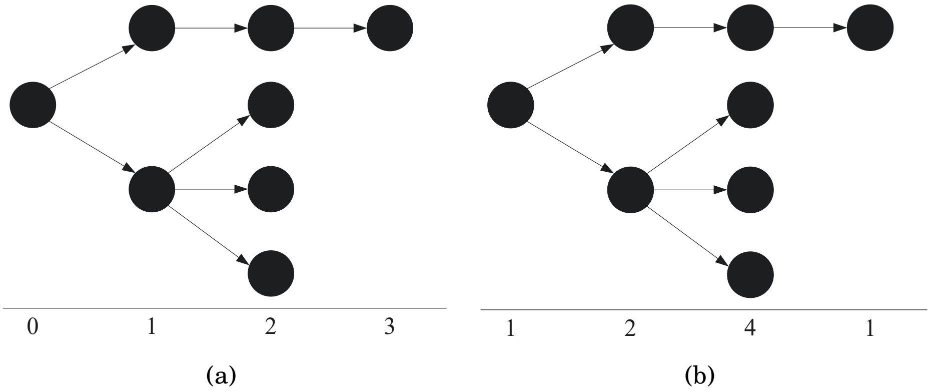
$$l = 1/N(N-1)\sum_{i \neq j} d_{ij} \tag{6.2}$$

Our diffusion path networks are the directed graph. This means that in these propagation networks, $d_{ij} \neq d_{ji}$. For the propagation path is essentially a tree graph, the value of shortest distance for many pairs of nodes $d_{ij} = 0$. Therefore, the average shortest path in our propagation path is less than 1. Figure 6.3 is the average shortest distance of these two categories of propagation path. Although their average shortest distance is less than 1, we can find that their values in privacy information are bigger than in normal news.

Propagation depth and width: for a propagation path graph with N nodes, its propagation depth is defined as the following:

$$D = Max(d_i), 0 \leq i \leq N \tag{6.3}$$

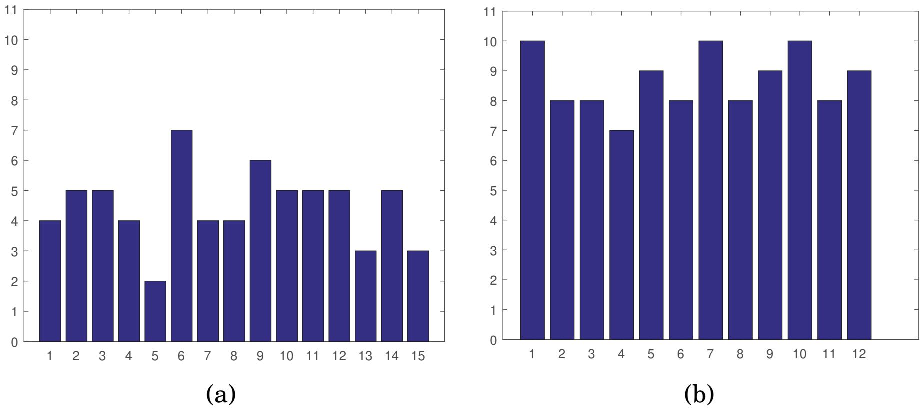
propagation width is the number of nodes which are in the same depth form a cascade layer in the network.



(a) (b)

Figure 6.4: Description of propagation depth and propagation width.

where d_i is the depth of nodes i from the original node in the graph. Figure 6.4 is an example of propagation depth and width. Sub-figure 6.4(a) is the sketch map of propagation depth. The depth number of each layer increases from 0. Sub-figure 6.4(b) is the sketch map of propagation width. It is the number of nodes in a certain propagation depth. Figure 6.5 and figure 6.6 show the propagation depth and width of both the normal news and privacy information. Among these figures, figure 6 shows the propagation depth of propagation path in our two categories of networks. Figure 6.6 shows the propagation width in the privacy propagation and normal news of our 27 information propagation paths. In these two sub-figures, the X-axis is the hops of information diffused in the graph and the Y-axis is the percentage of nodes in different hops through the graphs. We can conclude from these two figures that in normal news, the number of users this information propagated is decreased with the increasing of the hops from the original users. However, this situation is different in the privacy information propagation. The privacy information propagated most in the second or third hop of the propagation path. Meanwhile the propagation depth of normal news is smaller than the privacy information.



(a)

(b)

Figure 6.5: Propagation depth of two categories diffusion paths.

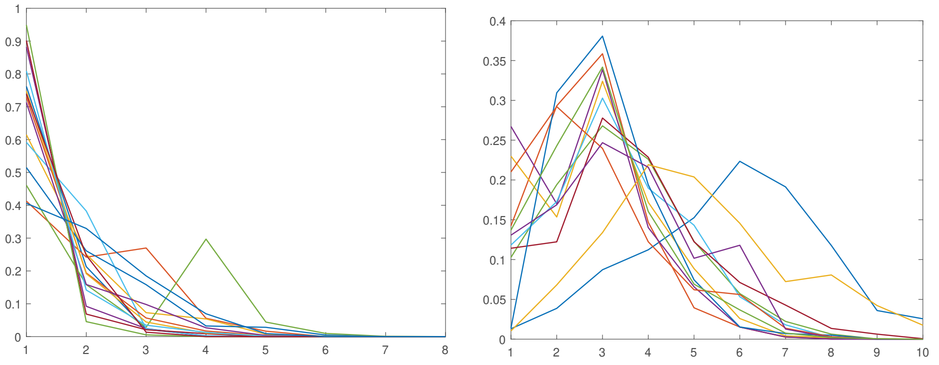
Discussion: The analysis of the basic structure in normal news and privacy information cascade shows that the normal news cascades have a shorter average path and propagation depth. Meanwhile, most of the nodes in normal news information cascade are distributed around the information source, while the nodes of privacy information cascade are distributed farther away from the source.

6.4.2 The high-order structures in complex network

Apart from the above global characteristics, there are also some unique basic structural elements in complex networks. These unique basic structural elements which consist the complex network are called high-order structures. In our information diffusion network, the most common high-order structures are the star structures which consisted by a central node with high degree and amount of surrounding nodes connected to the central node with low degree.

Figure 6.7 and 6.8 show the characteristic of star structures in both the normal news and privacy information. Among these figures, sub-figure 6.7(a) and 6.7(b) show the number of star structures in these 27 information diffusion paths. Sub-figure 6.8(a) and 6.8(b) is the proportion of the biggest star structures in the diffusion paths. We can intuitively observe that the number of star structures in privacy information diffusion paths is bigger than in normal news diffusion paths. However, the size of star structures in normal news diffusion paths is bigger than in privacy information diffusion paths. The biggest star structure in normal news propagation has occupied over 50% while none of the privacy propagation has occupied 50% in the dataset.

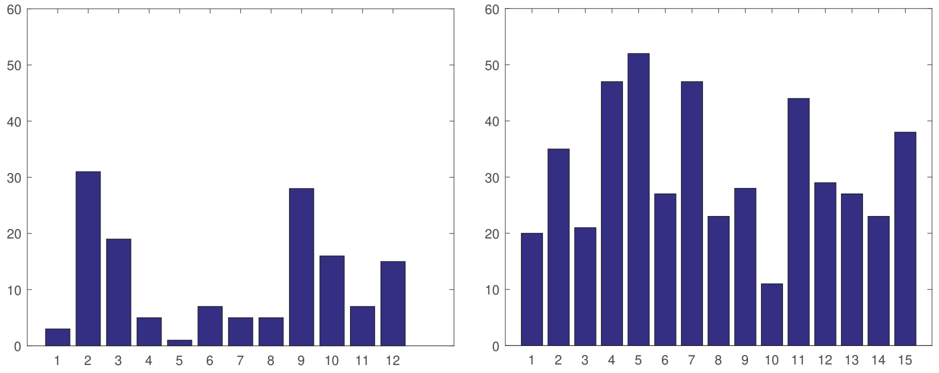
Discussion: The analysis of star structures in social media shows that the normal



(a) (b)

Figure 6.6: Propagation width of two categories diffusion paths.

news tends to be widely spread in social media, but it will not extend very deep in the propagation path, while privacy information will not spread very wide at first but spreads deeper and further in the following. The diffusion characteristics of normal news present a centralized diffusion modality. There are a few key nodes with great influence in the cascade, which play a key role in diffusion. The diffusion characteristics of privacy information present a decentralized diffusion modality. This category of information spread in a small range but in multiple places at the same time.



(a) (b)

Figure 6.7: numbers of star structures in diffusion path.

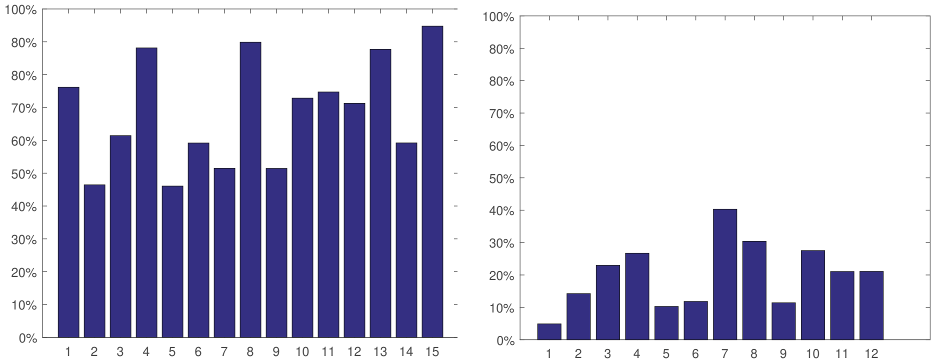
6.5 Privacy information diffusion model

A typical diffusion model is consisted of two parts: 1) the diffusion probability from the active nodes to the inactive nodes and 2) the rules obeyed during the diffusion. In this section, we will gives some principles of our modeling diffusion process, then describe the diffusion probability of different nodes in the network.

6.5.1 Assumptions and rules in privacy diffusion

According to the diffusion characteristics of privacy information which we discussed in Section 3, this thesis proposes the following assumptions for modeling the privacy information diffusion in social media.

- ***Assumption 1:*** Normal news publishers generally have many followers, while publishers of privacy information are usually ordinary users without many followers, which show low degree in the follower-followee network.



(a) (b)

Figure 6.8: proportion of biggest star structures in diffusion path.

- ***Assumption 2:*** The information could only diffuse from the active users to the inactive users. That is to say it can only diffuse from the users who have viewed the information to other users who have not viewed it.

- ***Assumption 3:*** The average shortest path and depth of normal news cascade and privacy information cascade show that privacy information spreads deeper and farther than news information. The ordinary users prefer to forward privacy information. While for the normal news, users prefer to view it rather than forward. Therefore, most users who forward the privacy information are ordinary users, and the users with large followers such as official account are the main users who forward the normal news.

- ***Assumption 4:*** Both the normal news and privacy information have strong timeliness. This means that the diffusion of the information will spread very quickly at the beginning but the diffusion probability will decrease with the passage of time.

Based on the above assumptions, we set five rules of our privacy information diffusion model in online social media as follows:

1. According to ***Assumption 1***, the initial nodes who publish the privacy information should start from the users whose followers are not many, i.e. the degree of the initial nodes are not very high. The normal news, on the contrary, start from the users who have multiple followers.
2. According to ***Assumption 2***, in social media, a user forwards a message when his/her followees forward this message. Therefore, the probability of a user forwarding a message should be related to whether the followees of this user have forwarded the message.
3. According to ***Assumption 3***, Different users have different attitudes towards different types of messages. As we have discussed before, users with large number of followers have a smaller forwarding probability for they are meticulous about their appearance in social media, while ordinary users should have a greater forwarding probability.
4. According to ***Assumption 4***, The probability of users forwarding this message will gradually decrease with the increase of time. Moreover, because news information is more time-sensitive, the probability of users forwarding news decreases with time at a higher rate than that of privacy information.
5. Different people have different levels of interest in different topics. Each user has different probabilities of forwarding different information.

6.5.2 Parameter set

Consider the following situation, in which a user views a message in social media, he/she is interested in this message and is ready to forward it. This process shows that the information diffusion in social media is decided by the following three parameters: 1) the probability of users to receive and view this message $View_i$, 2) the probability that users have tendency to forward this message $Forward_i$ and 3) the interest the users hold for this message $Interest_i$. Therefore, based on the above principles and situation the probability P_i of a user i to forward a message could defined as:

View: This metric represents the probability that user i views this message in social media. Users' attitudes to a message in social media are affected by its followers. The users with large number of followers have bigger influence than the users with small group of followers. That is to say when user i received message from his/her neighborhood users, the probability that user i diffuse this information is not only depend on the numbers of neighborhood users who forward this message but also decided by the importance of the neighborhood users. Therefore, parameter $View_i$ is defined as the following:

$$View_i = \sum_{u \in P_i} View_{i,u} \tag{6.4}$$

where P_i is the nodes set represent the users in user i's neighborhood users who have diffused this message. $View_{i,u}$ is the influence of user u to user i. Formula (8) shows that the users' tendency to diffuse this message is affected by all his/her neighborhood users who have diffused it. We consider the degree of the nodes as the importance of users in social media. Therefore, the influence of user u to user i is shown as the following:

$$View_{i,u} = \frac{d_u}{\sum_{w \in Ne_i} d_w} \tag{6.5}$$

where d_u is the degree of node u, Ne_i is the nodes set of user i's neighborhood users in social media. The influence of user u to user i is the ratio between the degree of node u and the sum degree of nodes set Ne_i.

Forward: As we have discussed before, the willingness of a user to forward certain types of messages is related to the number of users' followers. Therefore, the tendency of a certain user to forward a message $Forward_i$ is defined as the following:

$$Forward_i = cd_i^{\alpha} \tag{6.6}$$

where d_i is the in-degree of user i in follower-followee relationship network. α is a constant. If $\alpha \geq 0$, the $Forward_i$ is in direct proportion to the number of followers. On the contrary, it is inversely proportional to the number of followers. c is a constant in this formula.

Interest: Different people have different levels of interest in a different message. In the real world, the interests of people to different topics obey the normal distribution:

$$Interest_i \sim N(0,1) \tag{6.7}$$

$\triangle t^{-\alpha}$: In real cases, the probability of forwarding a message will change over time. This is because the information in the real world is time-sensitive. That means the communicability of information are decreasing with the increasing of time. Therefore, the probability of node i to forward a message is affected by how much time has passed when node i received this message. In our model, assuming that a message appears in user i's field of vision at time t, the influence of time changing in forwarding message probability at time $t+ \triangle t$ is:

$$\triangle t^{-\alpha} \tag{6.8}$$

where α is a constant.

6.6 Difference of our rebuilding propagation path and modeling different information

The normal news and privacy information, as we discussed before, is different in the diffusion paths. Therefore, our mechanism in modeling these two categories of diffusion paths has two differences: 1) the normal news often start from the nodes with a high degree while the privacy information are from the nodes with a middle degree or low degree and 2) the value of α in parameter $\triangle t^{-\alpha}$ for normal news is bigger than it in privacy information

6.7 Details of our modeling diffusion path

Figures 6.9 and 6.10 shows our modeling diffusion path of normal news and privacy information. The sub-figures in figure 6.9 are examples of the normal news and sub-figures in figure 6.10 are the examples of privacy information. The red nodes in these sub-figures are the source nodes of the diffusion paths and the red edges are the propagation direction of sources nodes in the model. We can conclude from these two figures that the modeling diffusion paths have the similar topological structures with the real diffusion paths for the following observations: 1) the information diffuses fewer hops than the privacy information, 2) the privacy propagation path has more *star structure* in the diffusion paths than the *star structure* in the normal information diffusion paths while the *star structure* in the privacy information diffusion paths is smaller than those in the normal news diffusion paths and 3) the source nodes diffuse to more users at the first hop in normal news than in the privacy information. All three observations are consistent with the real diffusion paths which we have discussed before.

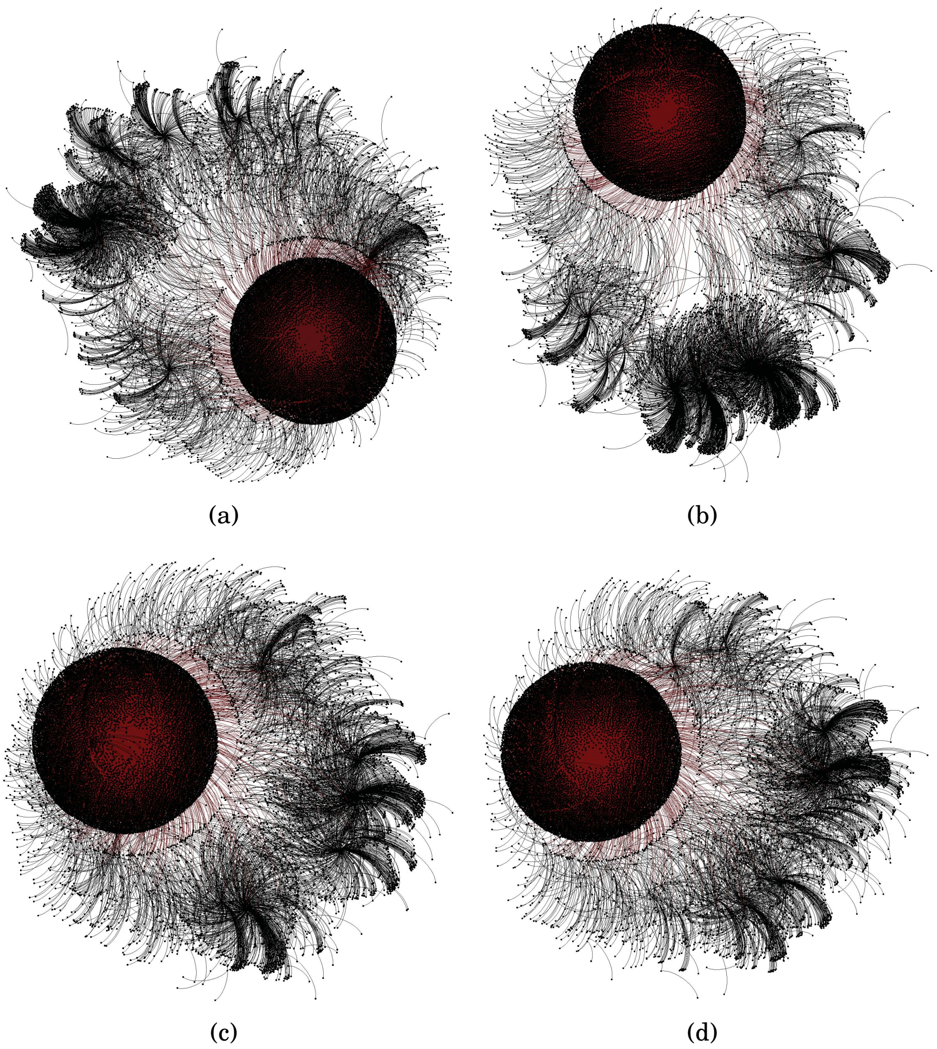
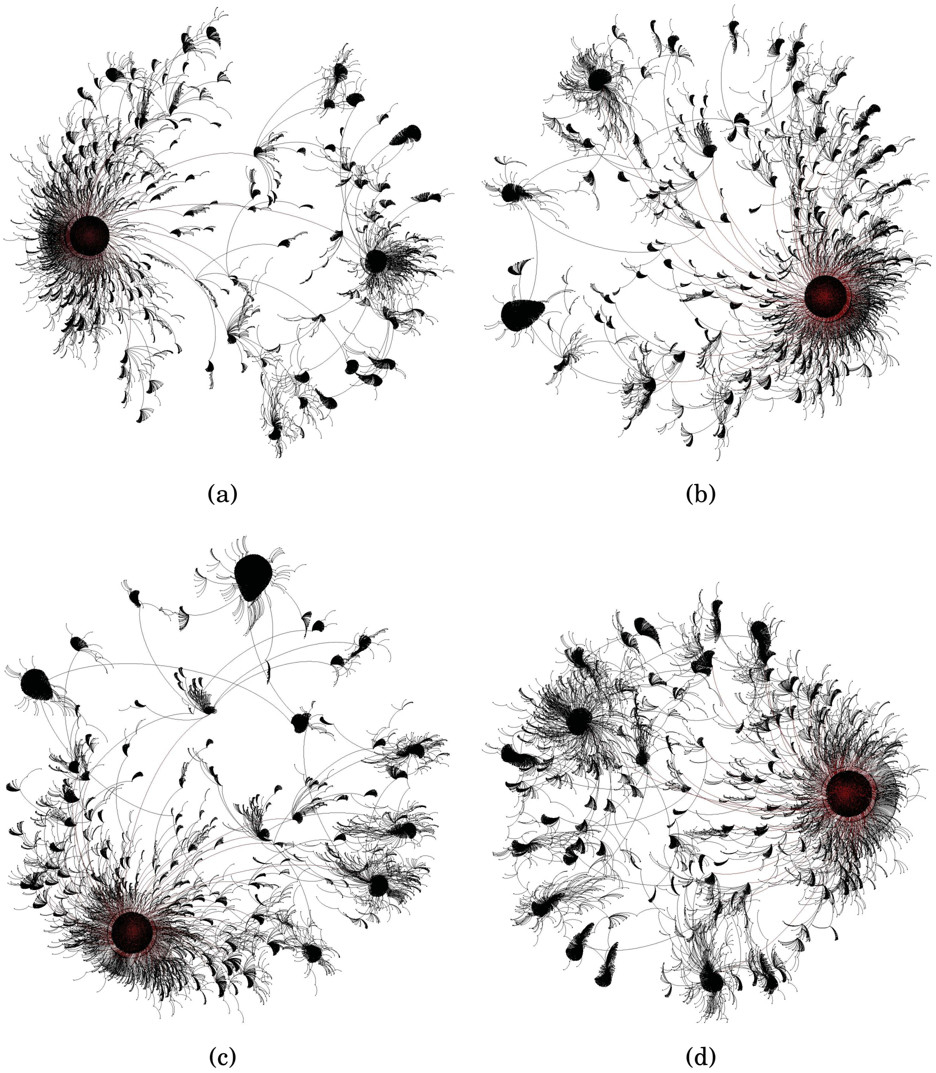


Figure 6.9: The examples of our simulating normal news diffusion paths .



(a)

(b)

(c)

(d)

Figure 6.10: The examples of simulating modeling privacy information diffusion paths.

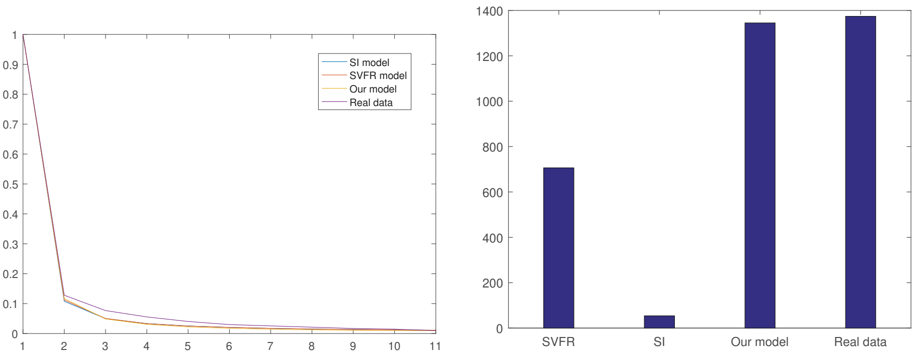
6.8 Comparison of our modeling mechanism and two other models

In this section, to evaluate the performance of our modeling mechanism, we compare the modeling diffusion paths with the real diffusion paths and two information diffusion models: the SI model and the SVFR models [61].

SI model: the SI model is a classical epidemic process. It can also be used in modeling the information diffusion in online social media. In such model, there are two-state situation: the susceptible (state S) and the infected (state I). The susceptible (state S) individual interacts with an infectious individual and becomes infected (state I). The classic understanding of epidemic dynamics is based on taking the continuous-time limit of difference equations for the evolution of the average number of individuals in each compartment. Usually, a differentiate equation is applied to describe the population over a certain time frame.

SVFR model: This model is used to describe the forwarding behavior of users on a given social network. Svfr model is based on classical virus propagation models such as Sir model, but it is more common and practical in the definition of user possible state. Each node in the svfr model may have four states: 1) Susceptible(S), the users who are possibly to read the messages, 2) View(V), the users who view the message, 3) 3. Forward(F), the users who forward the message and 4) Removed(R), the user who ignore the message.

This thesis modeling the information diffusion in the same real social network, and compares the gap between the information cascade generated by different models and the real data. The experimental results shown below are the average results of 20 experiments. In the following part of this section, we will compare these three models in the metrics we have discussed in section 3.

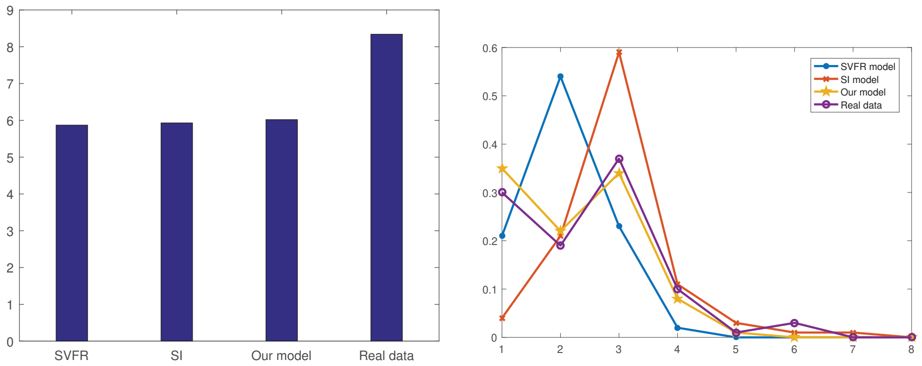


(a)

(b)

Figure 6.11: Degree distribution and variance of degree in four categories of diffusion paths.

Comparison of degree distribution: Sub-figure 6.11(a) is the degree distribution of the real diffusion paths, our modeling diffusion paths, the SI models and the SVFR model. We can observe from this sub-figure that these four categories of diffusion paths have similar degree distributions and obey the traditional power-law degree distribution in complex network. Sub-figure 6.11(b) is the variance of degree in these three categories of diffusion paths. The results show that our models of diffusion paths have similar variance of degree with the real diffusion paths while the SI model and the SVFR model has a large gap with the real cases. This shows that our modeling diffusion paths are closer to the real data.



(a)

(b)

Figure 6.12: Diffusion depth and width in three categories of diffusion paths.

Comparison of diffusion depth and width: Sub-figure 6.12(a) is the result of the diffusion depth in these four categories of diffusion paths. Although our model has some gaps with the actual result in diffusion depth. The results of our model are more closer to the real data compared with the SI model and SVFR model. Sub-figure 6.12(b) is for the results of diffusion width, compared with the SI model and SVFR model, our model gives a similar change tendency of diffusion width with the increase of diffusion hops while the SI model performs not very well with the real diffusion paths. We can also observe from sub-figure 6.12(b) that compared with the real diffusion paths, all of the three models have very few information which diffuse over 7 hops in the follower-followee relationship network. That is the reason why these three models perform not very well in the diffusion depth.

Comparison of star structures: Sub-figure 6.13(a) shows the number of star structures in the diffusion depth of these four categories of diffusion paths. Sub-figure 6.13(b) is the proportion of nodes in the biggest star structures for the whole network. These two sub-figures show that the SI model generates more small star structures compared with the real data. The SVFR perform better than the SI model but still exist some gaps with the real propagation paths. As a contrast, our model has similar star structures with real diffusion paths.

Discussion: Figure 6.14 are the examples of the three models we compared in this thesis. We can find that compared with our model and the SVFR model, the SI model are too separate to model the real information diffusion paths. Although the privacy information diffusion paths are decentralised compared with the normal news which

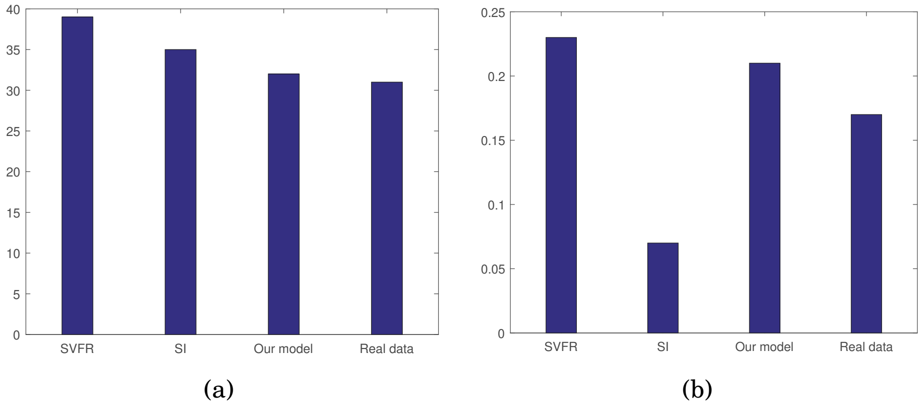


Figure 6.13: Star structures analysis in three categories of diffusion paths.

have one central node to diffuse the information. The privacy information still has several users which are regarded as the key nodes to diffuse the information in online social media. Therefore, the SI model is not suitable to model the information diffusion process in social media. The SVFR model has the similar intuitive topological structure with our diffusion model. However, based on the above comparative experimental results, our privacy information diffusion model in this thesis has better results in all of the parameters we mentioned in Section 3 compared with the SVFR model. And the privacy information cascade generated by the model is essentially close to the real data. The SVFR model is not suitable for modeling the diffusion of privacy information due to the lack of special research on privacy information. As a result, the SVFR model performed not very well in the parameter we list in this thesis. The effectiveness of our model in modeling privacy information diffusion is significantly better than other models.

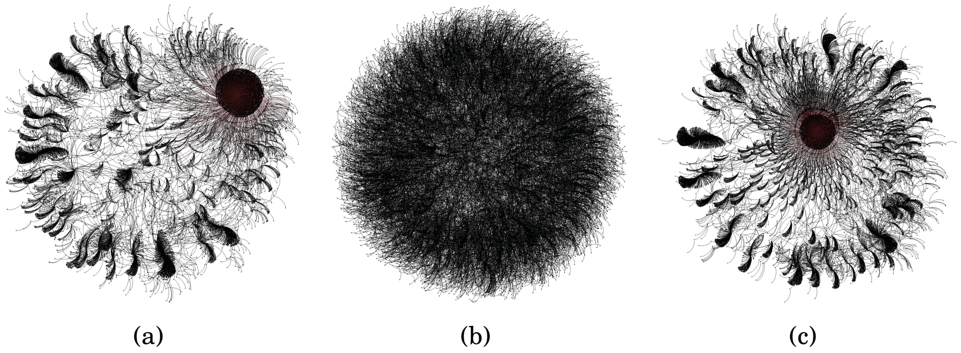


Figure 6.14: This figure is examples of the modeling privacy information diffusion paths.

Algorithm 1 Characteristic of Privacy Diffusion Based algorithm(CPDB)

```
Input: Initial active nodes set $A^0$
Initialization
The inactive nodes set $N^t$ = V - $A^0$,
The set of nodes which possibly diffuse amount of information $V_A^t$,
The set of nodes which have high diffusing probability $V_p^t$,
Congested nodes set $B^t$,
Process
for $t$ = 1 to T do:
  for $i$ = 1 to K do:
    Calculating the amount of information this node
    may spread in social media
    Obtaining the nodes set which may spread most
    amount of information $V_A^t$
  end for
  for $i$ = 1 to K do:
    Selecting the nodes set which have the highest
    probability to spread the information $V_P^t$
  end for
  If node $i$ in both $V_A^t$ and $V_p^t$
    Selecting all the nodes in $V_p^t$ which the diffusion
    probability are higher than node $i$ into $B^t$
  end If
  If $\|B^t\| \leq K$ do
    Selecting $n$ nodes in $V_p^t$ which may diffuse highest
    amount of information into $B^t$
    $n = K$ - $\|B^t\|$
  end If
Update $N^t$, $A^t$
end for
Output: $B^t$
```

6.9 Preserving mechanism

Following the modeling of information diffusion in social media, we propose a mechanism to block the diffusion of privacy information in the model. Our mechanism considers not only limiting the nodes with high ability to diffuse information but also the constraint of central nodes in small star structures in the propagation path. To achieve the goal of constraining the diffusion of privacy information in our privacy diffusion model, we select the nodes that need to be congested at a certain time interval dynamically to minimize the possible privacy diffusion. The goal of the algorithm is to measure the range of privacy information that may spread in online social network. Therefore, we define all the inactive nodes which are directly connected to the active nodes in the follower-followee relationship network are the target nodes. The directed edges which point from the active nodes to the inactive nodes are the target edges. We set the target nodes set at time interval t is N_t, $\beta_t(v)$ is the probability that the active nodes forward the privacy information. We set D^t as the privacy information the active nodes diffuse in social media, D^t could be written as:

$$D^t(v) = \sum_{e \in E(v)} w_e \tag{6.9}$$

where $E(v)$ is the edges set which the source node is node v, *.i.e* the out-degree of node v. Therefore, the privacy information that may spread at the time round t could be calculated as the following:

$$\beta_t(v) \cdot D^t(v) \tag{6.10}$$

In order to minimize the diffusion of privacy information with the observation time round t in the follower-followee relationship network, *.i.e*, to minimize the number of the nodes that transformed from inactive to active, we adopt the method of congestion some users in the information diffusion process. Assuming that the number of users which are congested in each round is K, it determines the budget for constraining privacy information diffusion. The greater K, the higher the cost is in this mechanism. Therefore, the goal of this mechanism is to find K users in each round of time which the spread of privacy information can be suppressed after congestion them in the follower-followee relationship network. Set N_t is the target nodes at time round t. d_i is the congestion of node i. $d_i = 0$ indicates that the user will be congested at this round. On the contrary $d_i = 1$ means the user will not be congested. The target function of this mechanism could be written as:

$$min(d^t \cdot \beta^t) \cdot D^t \tag{6.11}$$

$$s.t \sum_i (1 - d_i) = K \tag{6.12}$$

In the privacy information diffusion process, this mechanism dynamically and instantly selects important nodes at each time round, and congests these nodes from forwarding privacy information.

In real cases of online social media, the degree distribution of the network is uneven. The degree of some nodes in the network are very large while some are very small. The nodes with a large degree represent users having a large number of followers and a small number of followees. The probability of them to forward a message in social media is very small. Considering the above situation, we propose an algorithm called characteristic of privacy diffusion based (CPDB) which is based on the greedy algorithm to congest the privacy information diffusion. We have shown the pseudo code of this algorithm in ***Algorithm 1***.

Algorithm 2 Greedy algorithm

```
Input: Initial active nodes set $A^0$
Initialization
The inactive nodes set $N^t$ = V - $A^0$,
Congested nodes set $B^t$,
Process
for $t$ = 1 to T do:
  for $i$ = 1 to K do:
    $u = \underset{v \in N_t}{\arg\max} \beta_t(v) \times D_t(v)$
    $V_B^t = V_B^t \cup \{u\}$
  end for
  Update $N_t, A_t$
end for
Output: $B^t$
```

This algorithm congests the nodes which may spread the largest amount of information to the other nodes in the network or the nodes which have high forwarding probability. It selects K nodes which have the largest forwarding probability. If these nodes in the node set which have high forwarding probability also may diffuse a high amount of privacy information, then all nodes which have higher forwarding probability than this node will be selected. Finally, select nodes from the node set with high forwarding probability until K nodes are supplemented. This algorithm avoids selecting some users who have many followers but have a small propagation probability. Congesting these users not only has no obvious effect on congesting the diffusion of privacy information, but also greatly affects the user experience of these users. We will compare the CDPB mechanism with the greedy algorithm which congest the nodes which may possibly diffuse the largest amount of information.

Greedy algorithm: The greedy algorithm will first assess the nodes that may possibly spread the amount of privacy information in the network. Then it congests the nodes may diffuse the most amount of privacy information in each time rounds. In this way, it can block the diffusion of privacy information.

6.10 Evaluation

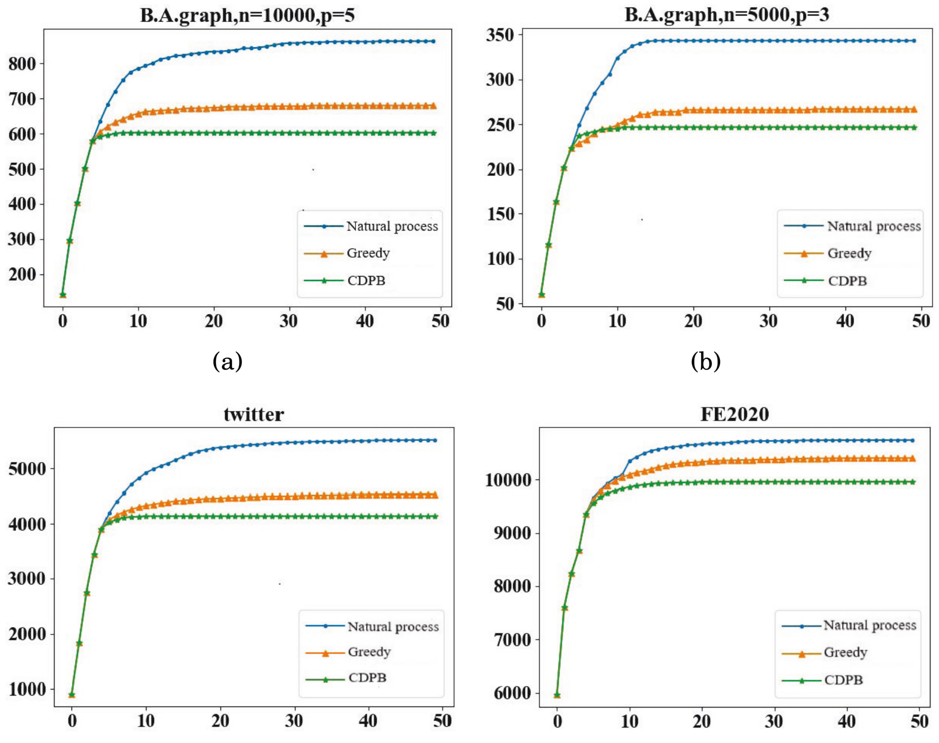
In order to simulate the real privacy preserving scenario, the privacy information diffusion model proposed in this thesis is first used to simulate the diffusion process of privacy information in the network. After several time rounds, the diffusion of privacy information has been detected and the congest algorithm is applied. In this section, we evaluate our algorithm by comparing it with the greedy algorithm.

Figure 6.15 shows the results of congesting effectiveness on the four datasets which we have described in Section 6.1. In these four sub-figures, the blue lines are the number of affected nodes in the natural process without the congesting mechanism in the diffusion process. The red lines are the results with the greedy mechanism in diffusion process. The green lines are the results with the ***CDPB*** mechanism we proposed in diffusion mechanism. We can observe from these four sub-figures that our proposed CDPB mechanism is better than the greedy mechanism in all of the datasets for the green line has diffused to fewer users than the red lines in the sub-figures.

Discussion: The reason why our mechanism performs better than the greedy algorithm is because our mechanism selects the nodes set dynamically in each time rounds and the method of finding the key nodes will select the nodes with low degree but may likely generate the small star structures in the diffusion paths. Therefore, we make a better performance than the greedy mechanism.

To evaluate the influence of our block mechanism on the users of social media, we introduce a metric called 'user comfort' to access the congest algorithm. This metric shows the users' experience during the block time in social media. It is defined as the following:

user comfort: Assuming that each user in social media has a maximum tolerable blocking time $T_th(u)$, we define the user comfort of the user as the following:



(a) (b)

(c) (d)

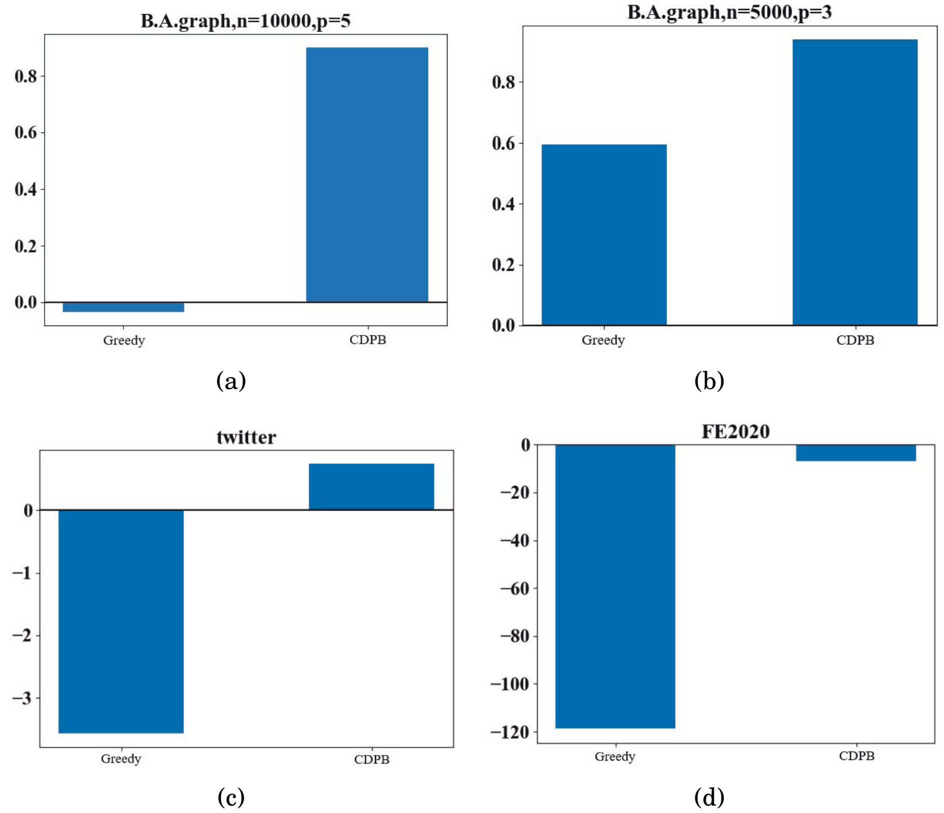
Figure 6.15: The effectiveness of the congesting mechanism during the diffusion process.

$$U = \frac{1}{N}\sum_{u=1}^{N}\frac{T_t h(u) - T_b(u)}{T_t h(u)} \tag{6.13}$$

where the $T_b(u)$ is the actual block time of a user in social media. From the actual situation, the blocking time thresholds that users can tolerate is different according to the followers they have in social media. Users with a large number of followers should have less tolerance time. Once these accounts are blocked, they will be more dissatisfied than ordinary users. Meanwhile, the followers of these users will be affected by this action during the block time. The users with few followers, on the other hand, will appear less concerned after being blocked. In other words, the ordinary users can tolerate longer blocking time. Therefore, we have:

$$T_t h(u) \propto k^{in}(u)^{-1} \tag{6.14}$$

Figure 6.16 is shows results of user comfort on our four datasets for both the greedy algorithm and our CDPB algorithm. We can observe from these four sub-figures that the



(a) (b)

(c) (d)

Figure 6.16: The user comfort of the congesting mechanism during the diffusion process.

user comfort of our CDPB algorithm is better than the greedy algorithm in all four of the datasets. This means that users in social media will feel more comfortable in our block mechanism than in the greedy algorithm.

Discussion: Our mechanism first blocks the users who may become the central node of small star structures participating in the privacy information diffusion process according to the special characteristics in the privacy information diffusion paths and we avoid blocking the users with large followers in the algorithm. This not only prevents the key nodes of privacy information from spreading information, but also better ensures the user experience. Based on all the experimental results, this algorithm is superior to the greedy algorithms in social media and ensuring user experience.

6.11 Summary

Privacy leakage is an important and troublesome issue for researchers and social media managers. However, the mechanism of the diffusion process for the privacy information in social media is still very limited due to the lack of studying the special features of privacy information in social media. In this chapter, we first compared the real diffusion paths of both the normal news and privacy information in the basic parameters and some high-order structures, i.e., the star structures in diffusion paths. We then build a new information diffusion model to simulate the diffusion process in social media. By comparing it with the traditional SI model, our model shows a closer topological structure with the real diffusion paths. Finally, we propose a mechanism to block the privacy diffusion in social media. The result shows that our mechanism has better effectiveness in congesting privacy diffusion and users' experience.

CHAPTER 7

A GRAPH EDGE BASED PRIVACY BLOCKING MECHANISM BASED ON SPARSE REPRESENTATION

7.1 Introduction

The fast development of social media has brought the cheap and convenient communication ways. However, as users share their personal information, which can then also spread through the online social network. The privacy information leakage refers to the information about an individual that has been published without the person's permission or awareness and such leaks present great risks to users of social media. Therefore, every one from the individual concerned to social media managers to data collectors and curators has an interest in taking measures to stop the propagation of privacy information in social media.

7.1.1 Information diffusion blocking mechanism

The structure of the social network is defined as a complex network, which the node represent the users and edges represent the connected relationships [50] As privacy diffusion has some similarity with the rumor propagation, we can start with analyzing current rumor blocking methods. Currently study of stop the illegitimate texts which include rumors, fake news and privacy information propagation in online social media has mainly focus the block of rumor information [136][14][38]. These methods often modeled the relationships of social media users into complex network and simulate the blocking mechanism through the network. They could be divided into the following three categories:

The first is limiting the users' behaviors. These behaviors include comment, share and retweet [14] [17]. The users limited in these methods are the users with high influences in the network. These high influence presented in the complex networks are the nodes with high degree, high intermediate centrality or high CI value which represent the influence of a node in complex networks [16][58]. The second method is to delete the users from social media and block the diffusion paths. The third method is diffusing the truth news against the rumor. In recent years, some researchers believed that limiting the users behaviors could not really stop the information diffusion in social media. Thus, they proposed several new methods to block the rumor information spreading. The key idea of these methods is to . By applying such method, the rumors will be collapse [98] [38].

However, these two types of methods are not suitable for stopping the privacy information propagation efficiently for the following reasons:

Firstly, the diffusion of privacy information is different from rumors. Users who diffuse the privacy information might be the normal users who do not have a high influence in online social media [41]. Therefore, limiting the high-influence users is less efficient in blocking the privacy information in social media.

Secondly, limiting the users behaviors may violates the spirit of ŕreedom of speechẃhich is the footstone of the modern online social network [98]. Meanwhile, limiting nodes in social media will drop a large number of unnecessary connections, which may decrease the utility of the social media.

Thirdly, unlike the rumors or fake news, the privacy information is the truth information but published without the person's permission or awareness. As the result, diffusing the truth information against the privacy information has little impact on the privacy diffusion.

Above challenges make the privacy data blocking a hard question in online social media. Blocking the privacy diffusion is different from the rumors as we can not accurately catch the privacy information from social media.

7.1.2 The advantage of blocking edges

Current research study directly on blocking the privacy diffusion mainly focused on limiting the important nodes [69]. This method will drop a large number of connection relationships when deleting the nodes from the network. For most dropped nodes in these methods are the nodes with high degree, the deleting action will drop the connections between these dropping nodes and their neighbourhood. These dropping connections are mostly unnecessary and cause a bad utility for users in social media.

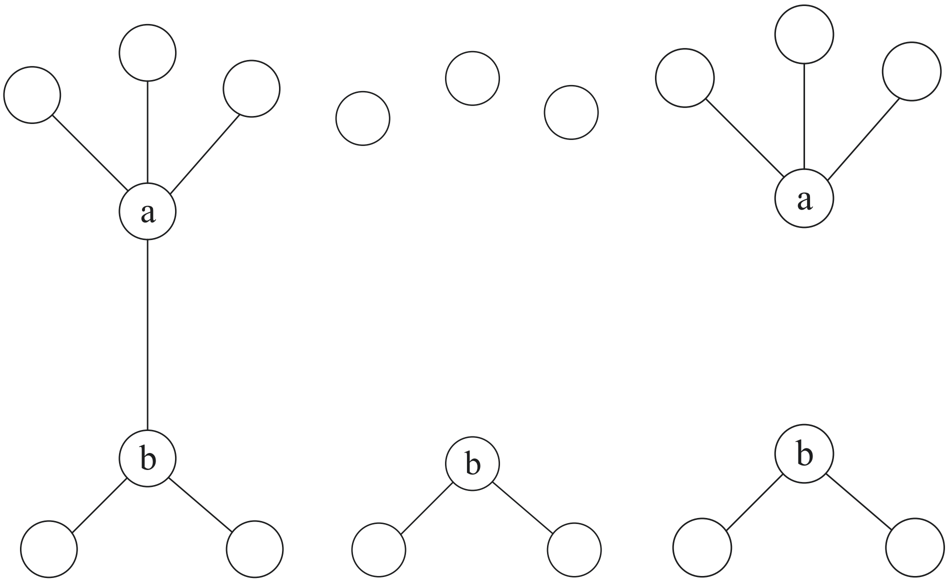


Figure 7.1: This figure show the advantage of blocking edges.

Therefore, In this chapter, we block the privacy data diffusion by simply limiting the connection between different users in online social network. Compared with limiting nodes, this method has the following advantages: 1) it will not limit users behavior in social media which provide a more comfortable experience, 2) it will limit only some important connection relationships in social media which means that we will not drop the unnecessary connections in the nodes limiting method. Figure 7.1 has shown the example of their difference. Sub-figure 7.1(a) is the original network. Sub-figure 7.1(b) is the node limiting methods which has delete 1 nodes and 4 edges from social media. Sub-figure 7.1(c) is the edges limiting method which only remove the edge E_{ab} in the network. However, current study has not given the parameter to evaluate the importance of edges in complex network. How to find these important nodes in the network is the key problem in this chapter. Considered the difficulties and complexities in solving such problem, in this chapter, we will first deeply analyse the topological structure in the complex network and discover the important edges in the network by using a newly proposed complex network analysis approach. This method will decompose the complex network into several atoms according to each node in the network, we will use these decomposed atoms to discover the important edges in social media. Our contributions could be summarised as the follows:

- We first analyse the topological structure of complex network which modeled from the social media based on a newly proposed complex network analysis approach, the ńetwork sparse representationt́o find the important relationships between different users in online social networks. By doing such work, we divide the edges in the network into three categories and describe why the third categories of edges are important in the complex network.
- We then propose a privacy diffusion blocking mechanism based on limiting these important relationships in social media.
- We compared our information diffusion blocking mechanism with those of using the traditional mechanism in online social media. Our mechanism can effectively restrict the velocity of privacy progress in the privacy information diffusion model.

7.2 Related work

In this section, we will discuss two research directions related to our research. We first introduce the privacy information diffusion area, then we come out with the high-order structures of complex network.

7.2.1 Network Sparse representation

The sparse representation is a signal decomposition mechanism to analyze the complex system. It is a mechanism used in image processing, signal processing, machine learning, and medical imaging. This method could also apply to face recognition, single image super-resolution, image sharpness assessment, and 3D shape estimation. Researchers have proposed several mechanisms for sparse representation, like locality-constrained linear coding, tree-structured dictionary learning, nonparametric Bayesian dictionary learning, etc. K-SVD is the most porpular method among these mechanisms. It seeks the dictionary that leads to the best representation under strict sparsity constraints. This algorithm could also be used in the complex network area to complete the network sparse representation [115].

The network sparse representation uses the k-svd mechanism which used in image sparse representation to demonstrate the notable contributions on network decomposition, dimensionality reduction, and reconstruction. Through the network sparse representation, researchers could decomposed several real-world networks into a set of atoms and its sparse coding. These atoms are finite and have a simple structure. For the dimentionality of atoms is smaller than the original network, the network sparse representation could be used for reconstruct the original network by using the linear decompositions of the atoms.

Meanwhile, the atoms generated by the network sparse representation could be used to analyse the basic structure complex network. Networks with similar structures have sets of atoms with similar statistical properties. It can be used to understand the details of different kinds of network and deeply analyse the problems related to the complex network such as network classification, recognition, and reconstruction. Beside, it can also be used in sparse representation can also be helpful in network synchronization, spreading, high-order organizations, and self-similarity.

7.2.2 High-order structures of complex network

Most complex networks in nature have global characteristics. Their degree distributions usually follows a long tail distribution, in which some nodes have more connections than the average. In addition, natural networks often share small-world characteristics in which they often show short paths and dense connections between nodes in the networks. Apart from these global characteristics, there are also some unique basic structural elements in the network. Motif is a kind of typical basic structural elements in complex network. It is defined as the network substructure which is more likely to appear in the actual network than the corresponding random network and they are different in different kind of network [46]. The simplest motif is consists of two nodes and an edges from one node to another. In most cases, the motifs researchers most used are the 3-order motif and 4-order motif. The information cascade in this chapter is the propagation path of tweets in social networks. Therefore, the whole cascade network is a tree network without many complex motifs [65][20].

The motifs could represent the complex network into several pieces. However, the presented structures are simple and is hard to model the complex network. Therefore, in recent years, researchers try to split the complex network into several pieces. The sparse representation is a good example. This method split the complex network by using the matrix decomposition and split the complex network into atoms [115].

7.3 Problem formulation

7.3.1 Online social network model

The relationships of online social network users are often modeled into a complex network $G(V,E)$. In such network, V is the set of nodes which represent the users in social media and E is the set of edges which represent the relationships between different users in social media. As this chapter describes blocking the privacy information diffusion, we give the definition of follower-followee relationship network and information diffusion network in social media, which are explained as follows:

Definition (Follower-followee relationship network). The nodes in the follower-followee network represent users. The edges are directed and represent one user's followership with the other users in social media [136].

Based on the follower-followee relationship network, the process of information diffusion happens through the directed edges between two nodes in the network. This diffusion process forms the information diffusion graph. The definition of information diffusion network are shown as the following:

Definition (Information diffusion network). The nodes in the information diffusion network represent users. The edges between node i and j are directed and represent the information are diffused from node V_i to V_j.

The adjacency matrix is a kind of square matrix which used to represent the finite network. Each element of the matrix indicate whether the pairs of nodes are adjacent or not in the complex network. For example, an element $A_{ij} = 0$ indicates that there is no edges between node i and j while it show there is an edges when $A_{ij} \neq 0$. The diagonal elements of the matrix are all zero, Since most research on complex network only consider the topological structure of complex network. The adjacency matrix are normally binary in research area, *.i.e* the non-zero element are 1 in adjacency matrix.

Degree is the basic features of nodes in complex network. The degree of one single node i in complex network is the number of nodes which connected directly to node i. Average_degree is the average value of all nodes in complex network:

$$Average\_degree = \frac{\sum_{i=1}^{n} D_i}{n} \tag{7.1}$$

where D_i is the degree of each node in the graph, n is the number of nodes in the graph.

For an edges connected 2 nodes in network, the sum of all the nodes' degree is 2 times the number of edges. Therefore, the product of average degree and number of nodes is also 2 times the number of edge.

7.3.2 Privacy diffusion model

In our previous work, we have modeled the diffusion of both the normal news and privacy information in online social media [40]. We assumed the privacy information and normal news diffusion obey the following rules:

- The information diffused through the connections between different users. As the follower-followee relationship are the most common sense, our models of information diffusion are based on this complex network
- the initial users who publish the privacy information should start from the nodes which do not have many followers while the users who publish the normal news are the nodes with high degree.
- the probability of an user diffuses a such information is determined by four parameters: 1) ***View:*** it indicates a user i views one message. Users' attitudes to a message are affected by its followers and followers' influence, 2) ***Forward:*** it shows the willing of an user to diffuse one message. In this model, the nodes with large degree have a smaller forwarding probability than other nodes for these users are meticulous about their appearance, while ordinary users should have a greater forwarding probability. 3) ***Interest:*** users have different interests through the social media and the interests of people to different topics obey the normal distribution in real world. 4) ***Time factor*** $\triangle t^{-\alpha}$***:*** In online social network, for the information has the time-sensitive features, the probability of node i to forward a message is affected by how much time has passed when node i received this message.

7.3.3 The overview of the blocking method

In this chapter, we try to limit the connection relationships between different users to block the privacy diffusion. To obtain these important connections from the complex network, we first invite a newly proposed complex network analysis tools, the network sparse representation to decompose the complex into several atoms and obtain these connections from these atoms. Then we block the privacy diffusion by limiting these important connections.

- Obtain the edges in atoms: By applying the sparse representation into this field, We decompose the network into several. we use these atoms from network sparse representation to discover the important edges in the topological structure of network. We divide the edges in the network into three categories according to the usage times when composed the original network. The more usage times, the more important these edges are in the complex network. We will analyse why the usage times are different in composing the network and why the edges with high usage times are important.

- Identifying important edge: By describing the detail of our improved sparse representation, we give how to find the high usage times edges in complex network. The usage times of different edges are determined by the number of splits for an element in the network matrix.

- Privacy blocking: we then block the information diffusion by limit these high usage times edges. We compared the result of these edges with two other method: blocking the random edges and the blocking the important nodes in the network. The result show that our method achieves a better result than the other two methods.

7.4 Sparse representation to obtain edges from atoms

7.4.1 The atoms in the network

The atoms are subgraphs, which represent networks by using their own special properties through self replication and sparse representation. In essence, they have the following characteristics: 1) they can generate complex networks by replicating themselves and linearly superposing them with other atoms, and this operation method remains in the sparse coding matrix; 2) Due to the optimization process in the network decomposition theory, the atomic set itself is a compressed representation of the network after the redundancy is removed; 3) Due to the strictness and consistency of the decomposition algorithm, the atomic and sparse codes of networks with similar structures are also similar, and have similar statistical characteristics. 4) Most of the complete connections of the network can be reconstructed by the inverse of the decomposition algorithm.

The atoms are the compressed representations of a network. These decomposed atoms could also be used to reconstruct the original complex networks. When the network sparse representation mechanism decompose the complex network, they will generate the atoms by using each nodes in the network graph. The size of each atoms n is determined by the average degree of the network. Normally, it will be slightly larger than the average degree to ensure that the generating atoms could cover the whole network. In the K-SVD mechanism, the algorithm will generate an atom for each ego node in the network. For an ego node i in the network, the atoms include its one-hop or two-hop neigborhood nodes with the number of n-1.

7.4.2 Finding the Redundant edges

When reconstructing the original network by the atoms, many edges on atoms has been used for multiple times during the process. This section will describe why this phenomenon happened.

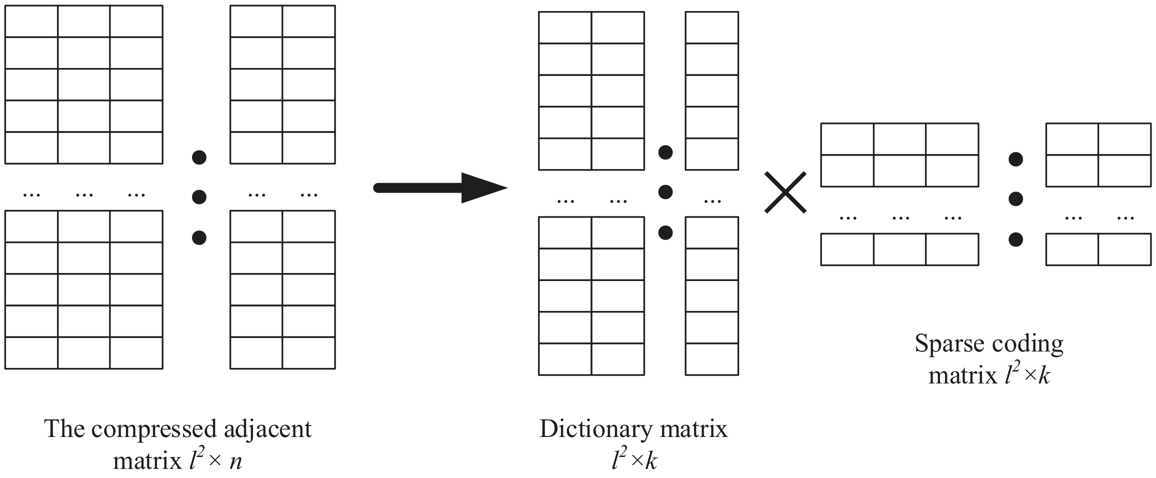

The compressed adjacent matrix $l^2 \times n$
Dictionary matrix $l^2 \times k$
Sparse coding matrix $l^2 \times k$

Figure 7.2: Decomposing process of the ego network matrix

The K-vsd mechanism, as a matrix decomposing algorithm, transforms the matrix into a linear combination of a set of base vector. figure 7.2 is an example, the K-SVD algorithm has decomposed the ego network matrix (with size of $l^2 \times n$) into two matrix. One is the dictionary matrix (With the size of $l^2 \times k$) which represent the different atoms we split from the original complex network. In such matrix, the column vector represent the atoms we split from the original network and the number k show how many categories of atoms construct this complex network. As the ego network matrix represent the topological matrix of the complex network, it is a binary matrix which has only one or zero element in the matrix. Therefore, we set the element of the dictionary

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 3 & 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 & 5 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

(A) (B)

Figure 7.3: The difference between the original ego network matrix and the reconstructing multiplication matrix.

matrix as the binary matrix during the iteration process. The other one is the sparse coding matrix (with the size of $k \times n$). In this matrix, we can get the position of each atoms by applying it with the dictionary matrix. This situation also happens in sparse coding matrix. It is a non-binary matrix and maps the atoms of each ego node in the network to their positions and the element should be set as the non-negative number.

As the non-zero element in the edges represent the connections between different nodes in the network. The formation process of the 1 element during the matrix multiplication of dictionary matrix and sparse coding matrix could be used to find the redundant edges. If an edges have been used only once in the sparse representation. This means that it present only by one single atoms of the ego node in the complex network. Therefore, the forming process could only be "1×1". The edges with multiple used times, on the contrary, have been used by several atoms of the ego node. For the element in the matrix is non-negative, the forming process should be "$1 \times a_1 + 1 \times a_2 + ...1 \times a_n$". In these process $a_1 + a_2 + ... + a_n = 1$. For the elements in the original network have only '0' or '1' element, these edges have been used several times and the number of the forming n are the usage times of them in the network for different ego nodes use these edges to generate the atoms. In the network sparse representation, the iteration process is to minimum the loss of the topological structures in the original network. Therefore, the different elements between the multiplication matrix and the original matrix have no influence on the topological structures of the network. We can also find the multiple usage times edges by using the elements from the multiplication matrix. Figure 5.2 is an example, these two matrix have the same topological structure when return to the original networks. Meanwhile, we can also get the usage times of different edges through the reconstructing multiplication matrix.

From the above discussion, we have give the explanation about how the different usage times of edges happened in the complex network during the sparse representation process. To find these edges, we can directly find the non-zero and none-one element during the process of forming an edges in sparse representations.

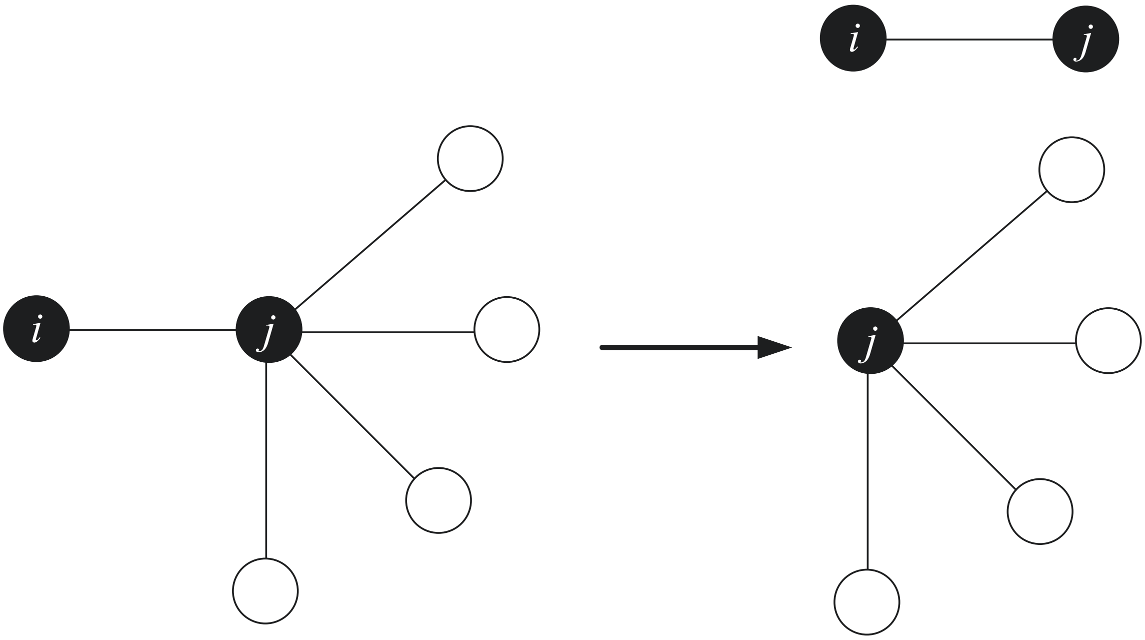


Figure 7.4: The situation that edges covered with only one time

7.5 The Redundant edges in atoms

The atoms we get from network sparse representation can help us to reconstruct the previous network. Due to the size of atoms is the same as the ego network which we sampled from the original network, it is larger than the average degree of the network. Therefore the network which reconstructed by the atoms has many edges which are repetitiously covered by several atoms in the graphs. Meanwhile, according the graph theory we discussed in Section 2, the product of number of nodes and average degree is 2 times of the edges in the graph. This means that all the edges should be present twice in the reconstruction process. However, some of the edges only present ones and some of them present more than twice during the process. As the complexity of this problem, in the following part of this section, we will divide the edges in the network into three categories. We will describe why they generate through the sparse representation and how they present in the complex network.

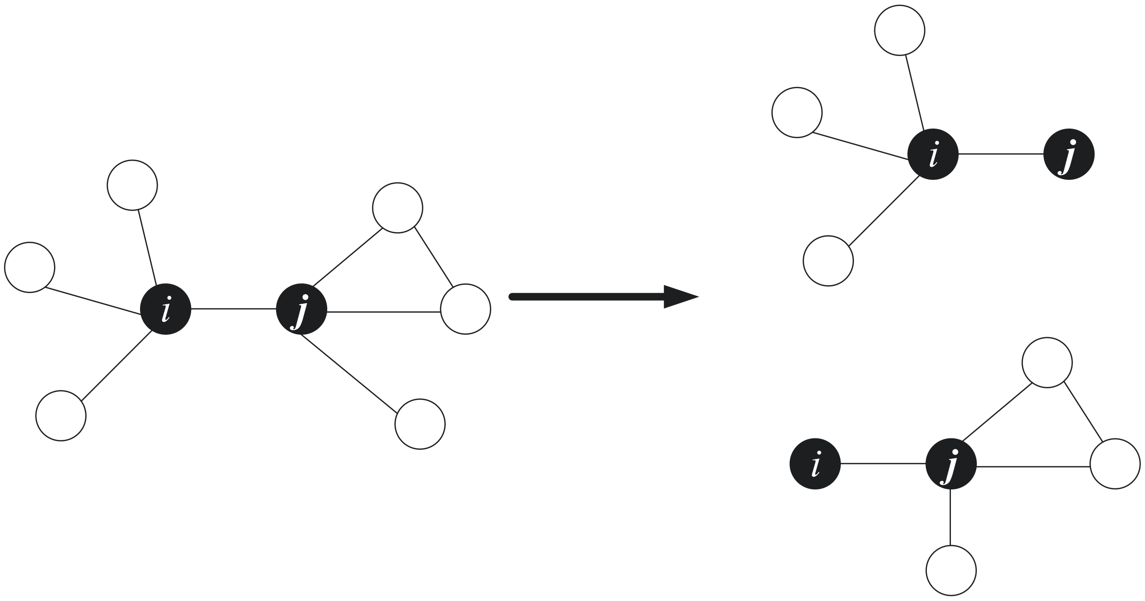


Figure 7.5: The situation that edges covered with two times

7.5.1 Redundant edges covered with only one time

This edges are often the connections between the nodes with high degree and nodes with low degree. According to the sparse representation, the size of an atom is a certain constant. This means that some atoms could present the whole structures of the ego node's neigborhood information. Some of them do not present the total information of the ego node's neigborhood structures. Therefore, some of the connections between the nodes with high degree and low degree will be presented by the low degree ego nodes but not presented by the high degree nodes in the atoms.

figure 7.4 is an example of this edges categories. the average degree of the original network is 1.66. We set the size of the atom at 3. As the result, the edges e_{ij} will be presented by the ego node i but not be presented by the ego node j in the atoms during the sparse representation process. This phenomenon leads to the situation that the edges only present ones in the process.

7.5.2 Redundant edges covered with two times

These edges are the most common edges in the network. The nodes connected by these edges have similar degree and they all present these edges in the generated atoms.

figure 7.5 is an example of this edges categories. the average degree of the original network is 2. We set the size of the atoms at 4. The edge between ego nodes i and j will be presented by both of the two nodes in the atoms in this situation.

7.5.3 Redundant edges covered with three or more times

These edges are the connections between the important nodes in the graph. The important nodes mean that these nodes have high degree and present a central roles in the complex network. Therefore, when the surrounding nodes of these important nodes generate atoms during the sparse representation process, these edges will be presented multiple times in different atoms.

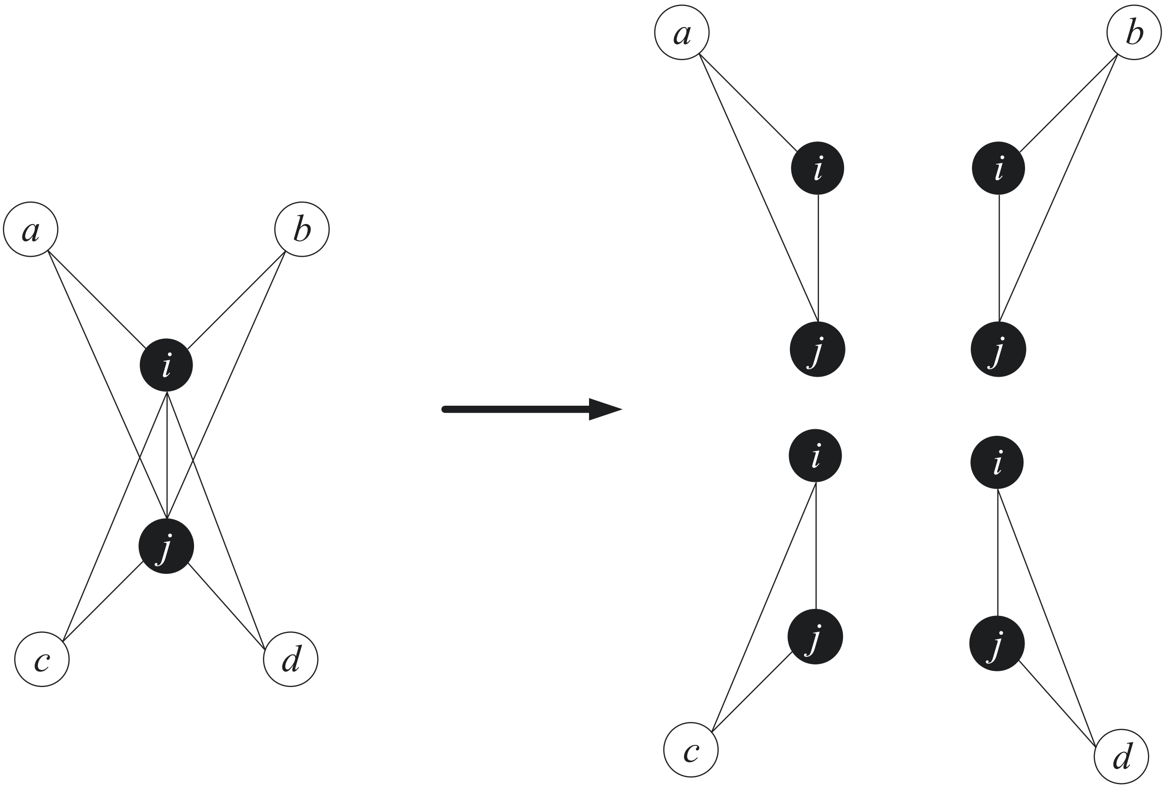


Figure 7.6: The situation that edges covered with three or more times

Figure 7.6 is an example of this edges categories. The average degree of the original network is 3, We set the size of the atoms at 4. We can see that when we generate the atoms of ego nodes a, b, c and d, all of them present the edge e_{ij} in the atoms for nodes i and j are the important nodes in the complex network. Edge e_{ij} is a significant edge in the graph, it is not only the connection between nodes i and j but also the directions from nodes a and b to nodes c and d.

```
Algorithm 3 Privacy Diffusion blocking mechanism
Input: Initial active nodes set $A^0$, follower-followee relationship network $G(V,E)$
Initialization
The inactive nodes set $N^t$ = V - $A^0$,
The set of nodes $V$ and set of edge $E$,
Process
Doing sparse representation for follower-followee relationship network $G(V,E)$ and obtain the atoms for each node in node set $V$
Reconstruct the network $G(V,E)$ and obtain the redundant edges set $E_d$ by the K-svd mechanism.
for edges in edge set $E$:
  If the edge is in the redundant edges set $E_d$
  delete this edge from edge set $E$
end for
Obtain the edge set $E_q$ without the redundant edges.
Obtain the novel follower-followee relationship network $G(V,E_d)$ by nodes set $V$ and edges set $E_q$
Simulate the diffusion process on the follower-followee relationship network $G(V,E_d)$
```

7.5.4 Discussion

In this section, we have analysis the edges in complex network and divided all the edges into three categories. In these three categories of edges, the edges covered only ones are the fringe edges. These edges often connected two nodes in which one is the node with a very low degree and another one with a high degree. In most cases, the low degree nodes have only one connections with other nodes in the complex network. For these edges often connected the common users and high influenced users in social media, they are very common and occupy a lot in the complex. The second category is the edges connected two nodes which are all with a average degree in the network. In other words, these two users are all the common users. The second category, on the contrary, connects two nodes with high degree which means that the users are all high influenced users. They are the important edges in the network and play an important role in information diffusion.

7.6 Information diffusion blocking

Following the information diffusion models we introduced in Section 2.2, we propose a mechanism to block the diffusion of privacy information in the model. Like we have discussed in the introduction part, we try to block the diffusion of privacy information by using the methods of drop the relationships between users. Obviously, the third category of edges we discussed in Section 4 are the main part of edges we limited in our blocking mechanism. We will first introduce how to find these edges in complex network, then we describe the privacy information blocking mechanism in the thesis.

With the important edges we get from the network sparse representations, we then aim to block the privacy information diffusion. For the difficulty of blocking the real information diffusion, in this thesis, we choose a privacy information diffusion model in the previous work [98] and simulate experiment on it. In this model, information diffuses through the online social network along the follower-followee relationships between different users. Therefore, our experiments will set on the follower-followee relationships networks and form the diffusion network. The probability of each users in the network to diffuse one information is based on four parameters which we have discussed in Section 2.4. Our experiment is based on the modeling of privacy information diffusion. The blocking mechanism is to drop these important edges from the original network and cut off the diffusion chain in social media. We have shown the pseudo code of this algorithm in ***Algorithm 1***.

Table 7.1: Network details

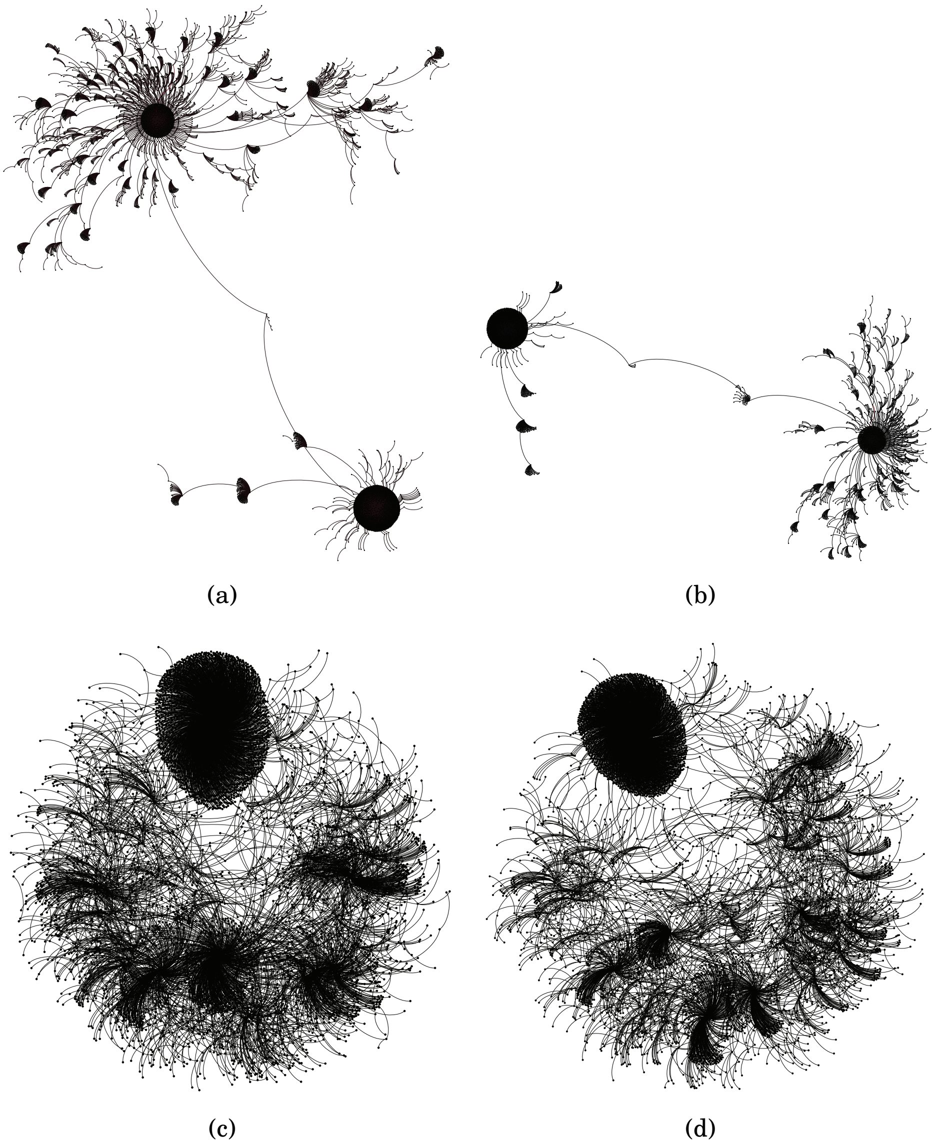
Network	**Number of nodes**	**Number of edges**
Twitter 1	6332	125114
Facebook 2	9345	198712
Random 3	6332	126475

7.7 Experiment

In this section, we describe our experiments and evaluate the performance of our privacy blocking mechanism. We first introduce the experimental settings, and then we carry out a set of experiments to show the effectiveness of the blocking method by comparing the results of the original diffusion model and the blocking diffusion model. We also compare our result with other two methods as well in the experiment.

7.7.1 Dataset description

We experimented on three sets of social media data in the sample area. Among these three datasets, two of them are the follower-followee relationships graph from the social media. One is the dataset we collected from Twitter and another one is the dataset from an open complex network data resource in SNAP. This datset are collected from the Facebook. The last one are the simulated random network which has the similar parameter with the above two network. We show the basic parameter of these three network are shown in the Table 6.1 as the following.



(a)

(b)

(c)

(d)

Figure 7.7: The simulated diffusion paths examples of Twitter data.

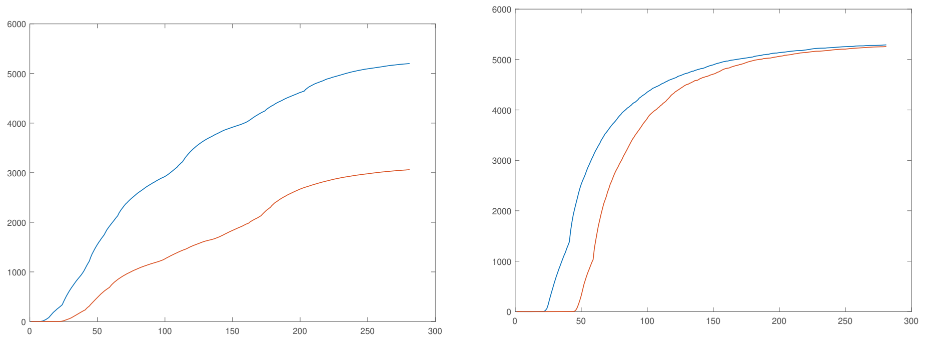
7.7.2 Experiment process

In the experiment, we first try to find the important edges in our three follower-followee relationships networks and drop them from the complex network. Then we simulate the diffusion model on both the original network without dropping the edges and the network which dropped the edges. Then we compare the results of these two simulated models. We then compare our blocking mechanism with other two methods which are randomly blocking edges or blocking nodes in follower-followee relationships networks to show the effectiveness of our methods.

7.7.3 Comparison of normal news and privacy information

We first compare the diffusion model on the privacy information and the normal news information. According to the previous study, the normal news are published by the users with a large number of followers while the privacy information often diffused from the nodes with a users who have some followers but do not hold a high influences [67]. Therefore, in this thesis, we choose the degree centrality as the different publisher on online social networks.

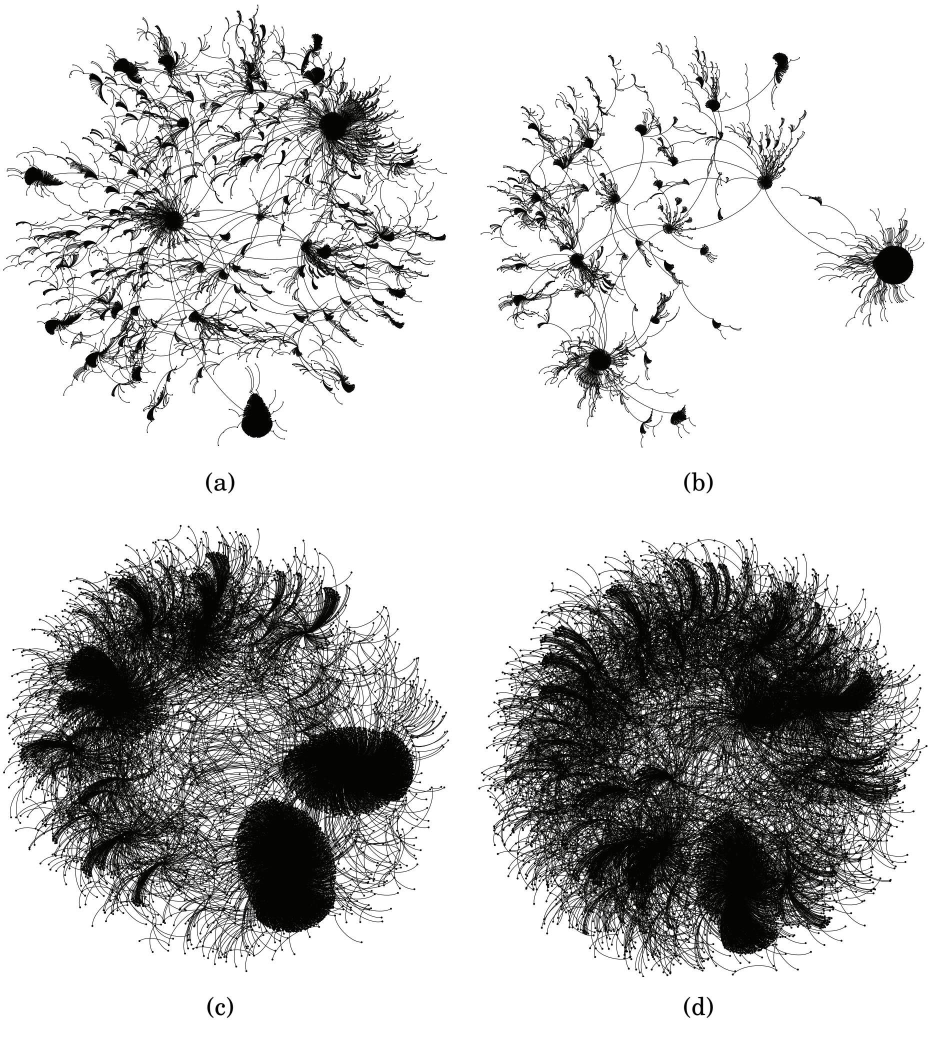
Definition (Degree centrality). The degree centrality is defined as the number of links that incident upon a node, *i.e* the paths of any pair of nodes passed one node in online . It indicate the important of one node in diffusion area. For the follower-followee relationships we used in this thesis, we choose the in-degree centrality to measure the nodes.



(a) (b)

Figure 7.8: The number of infected nodes in follower-followee relationship network during the simulated process.

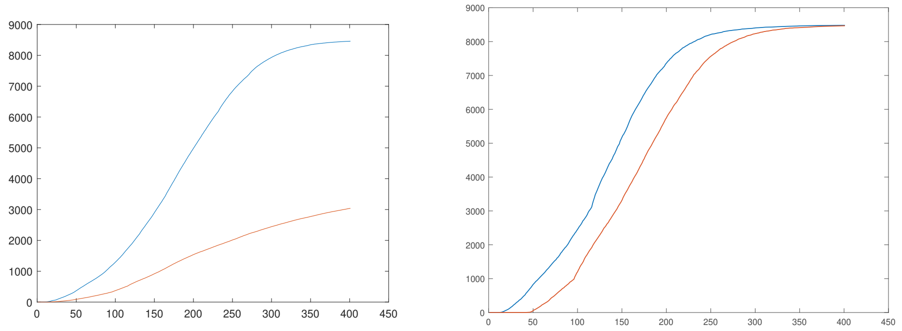
Based on the above definition, we simulate the diffusion process of both the privacy information and the normal news on the Twitter data and the Facebook data. Based on the previous study, we choose the high in-degree centrality nodes as the users who publish the normal news while the median in-degree centrality nodes as the privacy information publishers. In this section, we will simulate four categories of experiments: the privacy information diffusion on original network, the privacy information diffusion on the network without the redundant edges, the normal news on original network and the normal news diffusion on the network without the redundant edges. For one diffusion process experiment could not reflect the phenomenon, we simulated each category of experiment for 100 times and average them to get the result.



(a) (b) (c) (d)

Figure 7.9: The diffusion examples of the Facebook dataset.

Figure 7.7 shows the examples of our simulated privacy information and normal news diffusion process on the Twitter data. In figure 7.7, sub-figure 7.7(a) is the results of privacy information diffusion with the complete follower-followee relationships network and sub-figure 7.7(b) is the result with the network which block the third category of redundant edges. Sub-figure 7.7(c) and sub-figure 7.7(d) are the results of simulating normal news diffusion on the network with and without the redundant edges. Figure 7.8 shows the number of infected nodes in follower-followee relationship network during the simulated process. From these two figures, we can observe that the privacy can spread to over 5000 nodes in the network when the follower-followee relationship network is complete while it only spread to 3000 users when we block the third category of redundant edges in the follower-followee relationship network. It shows that deleting these important edges which connected important users can effectively blocking the privacy information diffusion while the normal news are less affected by this action.



(a)

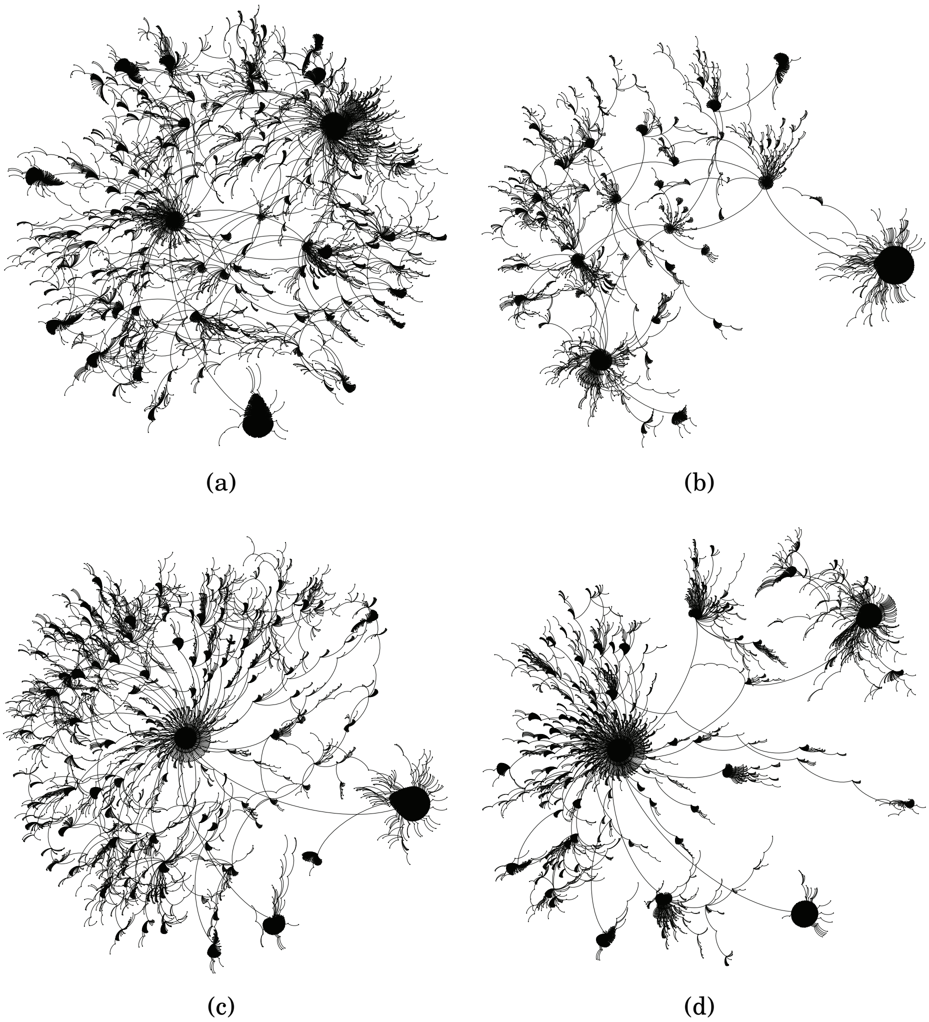
(b)

Figure 7.10: Comparison of privacy information and normal news diffusion process in Facebook dataset .

The similar situation happens on the Facebook we used in this thesis, Figure 7.9 shows the examples of the simulated privacy information and normal news diffusion process. Like in the previous examples. Sub-figure 7.9(a) and sub-figure 7.9(b) are the results of simulating normal news diffusion on the network with and without the redundant edges. Sub-figure 7.9(c) and sub-figure 7.9(d) show the simulating results of normal news diffusion. Figure 7.10 shows the the number of infected nodes in follower-followee relationship network during the simulated process. Similarly with Twitter dataset, the privacy diffusion has spread to 8000 users in the original graph while for the network without redundant edges it only diffuses to 5000. On the contrary, the diffusion of normal news are less affected by this action.

7.7.4 Comparison with other methods

In the last section, we compare the result of privacy information diffusion on the original follower-followee network and the network without the redundant edges. To better present the effectiveness of our mechanism, this section compares our method with other two methods which block the privacy diffusion: 1) randomly blocking the edges in the follower-followee relationship network and 2) blocking the nodes in the network. In the following part of this section, we will first introduce the above two methods and the metric to evaluate the efficiency of blocking mechanism, then we carry out with the simulated results. Like in the last section, we simulate each one method for 100 times on one graph to get the average result.



(a)

(b)

(c)

(d)

Figure 7.11: The comparison of diffusion examples by using the three categories of privacy information blocking mechanism.

Randomly blocking edges: The first kind of contrast method is to randomly block the edges from the follower-followee relationship network. In the previous experiment, we have blocked 8139 edges in Dataset A and 16807 edges in Dataset. Therefore, to control variables, we try to block the same number of edges but randomly chosen from the follower-followee relationships network then simulated experiment on the network.

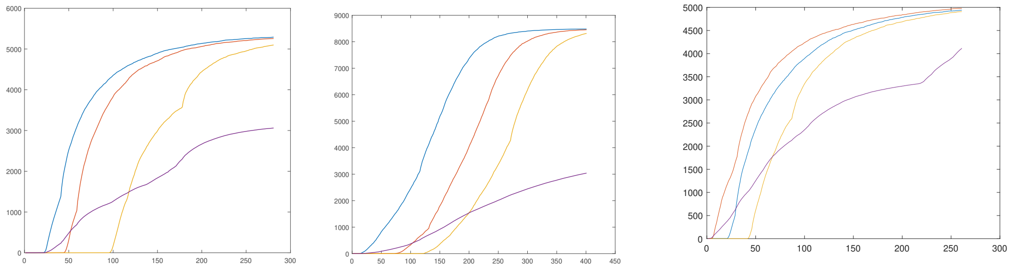
Table 7.2: Active rate of three methods

Dataset	Our method	Random	Nodes
Twitter dataset	61.45%	81.76%	80.32%
Facebook dataset	29.37%	86.71%	85.37%
Simulated dataset	30.42%	90.35%	88.63%

Randomly blocking nodes: The second kind of contrast method is to block the nodes in the original network. However, how to measure the number of blocking nodes which is equal to the blocking edges in the follower-followee relationship network is a problem. In this thesis, we try to block the nodes with high degree in the network. In this process we strike out of the nodes with high degree and block them from the biggest number to bottom. We cut off these nodes until the lost edges from the cutting nodes are similar to the blocking edges in our methods. Then we try to simulate the diffusion experiment on it.

Active rate: This metric is used to motify the efficiency of the above three method. It shows the how many percentage of nodes are affected in the simulating process. Obviously, the lower of this metric, the better blocking efficiency we get.

Figure 7.11 is an example result of the simulated diffusion result on the Twitter data (Referred as Dataset A) of the three method. Sub-figure 7.11(a) is the simulated result of the original diffusion results and Sub-figure 7.11(b) our edges blocking mechanism. Sub-figure 7.11(c) and Sub-figure 7.11(d) are the results of the Contrast methods. We can obviously observed from these sub-figures that the diffusion size of our blocking the redundant edges from follower-followee relationship network are smaller than the other two contrast methods. The average result of these three methods are shown in Figure 7.12 and the active rate of the three methods are shown in Table 6.2. We can conclude from the above resultd that the contrast methods which randomly blocking edges or limiting the nodes can only delay the diffusion of privacy but can only reduce a small size of the users who are affected by the privacy information. Meanwhile, our method could effectively constrain the affected users in the follower-followee relationship network.



(a) (b) (c)

Figure 7.12: Comparison of privacy information and normal news diffusion process in three datasets.

7.8 Summary

The above experiments have shown that the third category of redundant edges in the network have great influence on privacy information diffusion while the normal news did not affect by this factor. The reason for this phenomenon is that the normal news are often published by the users who have high influences while the users who release the privacy information are not very influential. Thus for the normal news the information spread very fast at the beginning for the high influenced users will quickly influence their followers in online social network and blocking the important edges could not stop this process. On the other hand, the diffusion of privacy information are based on users' interests. According to the previous study by [103], online social media users are more likely to diffuse a message which contains more interest than the normal news. Therefore, this kind of message will diffuse deeper and wider than the normal news and form more star structures in the network. In such situation, many users receive the message through the important connections between the influenced users in the follower-followee relationship network. Therefore, blocking the important can effectively cut off the diffusion paths.

For the blocking mechanism, blocking edges randomly from the follower-followee relationship network has the most worst effect on blocking the privacy information diffusion for it does not find the key edges in diffusion process. Limiting the nodes with high degree in the network has a better effect for these nodes play important role in the information diffusion process. However, in such situation, the privacy information still spread to the whole network and this method only reduces a little size of the diffusion scale. This means that without the high degree nodes in the follower-followee relationship network, the privacy information can still spread to the whole network. On the contrary, our method which block the important edges in the network can effectively cut off the diffusion paths. Compared with other two methods, it can not limit the privacy information diffusion at the beginning but can limit diffusion scale in the whole process.

;’/

CHAPTER 8

CONCLUSION

Privacy preservation has been an important issue in social media and cyber security. However, the propagation process of privacy information in the online social network still shows gap due to the lack of understanding of the propagation features of privacy information in social media. In this thesis, based on the case study of real diffusion paths in social media, we have deeply analysed how the information diffuse through online social network. our main contribution could be summarised as the following:

- we first collect the propagation data from Twitter and reconstruct the propagation process of both the normal news and privacy information into graph data. We then analyse the basic structure of these two categories of networks and discover that they are similar in topological features but different in propagation features. The result shows that the normal news often propagated in the first and second hops of the interpersonal network while the privacy information will spread deeper. We further classify these two categories of networks by using the Graphsage mechanism and try to find a method to stop the propagation process of privacy information. We preserve privacy information propagation in online social media by removing the high influence (CI) nodes from the social media. Extensive experiments show the efficiency of the Graphsage mechanisms and the CI nodes in networks.

- We compared the real diffusion paths of both the normal news and privacy information in the basic parameters and some high-order structures, i.e., the star structures in diffusion paths. We then build a new information diffusion model to simulate the diffusion process in social media. By comparing it with the traditional SI model, our model shows a closer topological structure with the real diffusion paths. Finally, we propose a mechanism to block the privacy diffusion in social media. The result shows that our mechanism has better effectiveness in congesting privacy diffusion and users' experience.

- we propose a novel method to block the privacy information in online social network. Unlike blocking users, our blocking mechanism blocking the connections between different users in the follower-followee relationship which avoids to violate the spirit of freedom of speech. In such method, we use the novel complex network decomposition, the sparse representation method to discover the important connections between different users in the complex network. We divide the edges in the network into three categories and proof that the edges which connect the nodes who have multiple surrounding nodes are the key connection in complex network. We then present the information diffusion model on the follower-followee relationship network with and without these edges in the network. The results show that our method could effectively reduce the diffusion of privacy information while the normal news are almost unaffected. We have also compared our methods with other two contrast information diffusion blocking mechanism. The results show that our method has the best blocking effectiveness.

Moreover, the analysis of real propagation structures has brought new insights in research area. Firstly, as we discussed before, the current studies of modelling information propagation are based on researchers' common sense and did not consider the real situations in online social networks. Therefore, we can build a new model which are closer to the actual situation to simulate the information propagation in social media. Secondly, for the reason that this thesis use the real data from social media, we can not analyse the privacy preserving method deeply. Therefore, we still needs research on how to propose a mechanism to preserve privacy leakage in social media. We believe that, based on the model of information propagation we mentioned before, we could give more details on the privacy preserving mechanism in a mathematic way. Thirdly, it is worth noting that the method of removing the CI nodes from the social media may not only stop the propagation of privacy information, but may also limit the propagation of normal news in online social network. How to preserve privacy information propagation while not limiting the propagation of normal news in social network will present opportunities for further research in this field. Meanwhile, our blocking method of information has given a new insight into this area. Firstly, as we discussed before, researcher try to block the rumors by spreading the truth information. How to spread these truth to against the rumors is the key elements. We can use these important connection between different users to strengthen this process. Secondly, we have proofed that these edges plays an important roles in information diffusion process. What the roles these edges are in the complex network? How they contribute to form the topological structures still need discovery. Thirdly, like in the rumor blocking, what kind of information is suitable for against the privacy information diffusion can be the future direction of this area.

APPENDIX A

APPENDIX

BIBLIOGRAPHY

[1] J. H. ABAWAJY, M. I. H. NINGGAL, AND T. HERAWAN, *Privacy preserving social network data publication*, IEEE communications surveys & tutorials, 18 (2016), pp. 1974–1997.

[2] M. AHARON, M. ELAD, AND A. BRUCKSTEIN, *K-svd: An algorithm for designing overcomplete dictionaries for sparse representation*, IEEE Transactions on signal processing, 54 (2006), pp. 4311–4322.

[3] P. BAJARI AND A. HORTAÇSU, *The winner's curse, reserve prices, and endogenous entry: Empirical insights from ebay auctions*, RAND Journal of Economics, (2003), pp. 329–355.

[4] A.-L. BARABÁSI AND R. ALBERT, *Emergence of scaling in random networks*, science, 286 (1999), pp. 509–512.

[5] A.-L. BARABÁSI AND E. BONABEAU, *Scale-free networks*, Scientific american, 288 (2003), pp. 60–69.

[6] A.-L. BARABÁSI, N. GULBAHCE, AND J. LOSCALZO, *Network medicine: a network-based approach to human disease*, Nature reviews genetics, 12 (2011), pp. 56–68.

[7] A.-L. BARABASI AND Z. N. OLTVAI, *Network biology: understanding the cell's functional organization*, Nature reviews genetics, 5 (2004), pp. 101–113.

[8] T. BERNERS-LEE, W. HALL, J. HENDLER, N. SHADBOLT, AND D. J. WEITZNER, *Creating a science of the web*, Science, 313 (2006), pp. 769–771.

[9] L. BIN, H. DONG, Z. DENGJI, AND Z. TAO, *Customer sharing in economic networks with costs*, arXiv preprint arXiv:1807.06822, (2018).

[10] J. Blocki, A. Blum, A. Datta, and O. Sheffet, *Differentially private data analysis of social networks via restricted sensitivity*, in Proceedings of the 4th conference on Innovations in Theoretical Computer Science, ACM, 2013, pp. 87–96.

[11] M. Boguá, R. Pastor-Satorras, and A. Vespignani, *Epidemic spreading in complex networks with degree correlations*, in Statistical mechanics of complex networks, Springer, 2003, pp. 127–147.

[12] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, *Network analysis in the social sciences*, science, 323 (2009), pp. 892–895.

[13] D. M. Boyd and N. B. Ellison, *Social network sites: Definition, history, and scholarship*, Journal of computer-mediated Communication, 13 (2007), pp. 210–230.

[14] C. Budak, D. Agrawal, and A. E. Abbadi, *Limiting the spread of misinformation in social networks*, in Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011, 2011.

[15] H. Chen, B. Perozzi, Y. Hu, and S. Skiena, *Harp: Hierarchical representation learning for networks*, in Proceedings of the AAAI conference on artificial intelligence, vol. 32, 2018.

[16] T. Chen, W. Liu, Q. Fang, J. Guo, and D.-Z. Du, *Minimizing misinformation profit in social networks*, IEEE Transactions on Computational Social Systems, 6 (2019), pp. 1206–1218.

[17] W. Chen, *An issue in the martingale analysis of the influence maximization algorithm imm*, in International Conference on Computational Social Networks, Springer, 2018, pp. 286–297.

[18] Z. Chen, T. Ni, H. Zhong, S. Zhang, and J. Cui, *Differentially Private Double Spectrum Auction with Approximate Social Welfare Maximization*, IEEE Transactions on Information Forensics and Security, 14 (2019), pp. 2805–2818.

[19] E. H. Clarke, *Multipart pricing of public goods*, Public choice, 11 (1971), pp. 17–33.

[20] A. K. DEY, Y. R. GEL, AND H. V. POOR, *What network motifs tell us about resilience and reliability of complex networks*, Proceedings of the National Academy of Sciences, 116 (2019), pp. 19368–19373.

[21] DINGDA, YANG, XIANGWEN, LIAO, HUAWEI, SHEN, XUEQI, CHENG, GUOLONG, AND CHEN, *Modeling the reemergence of information diffusion in social network*, Physica A Statistical Mechanics Its Applications, (2018).

[22] D.WORK AND A. ROTH, *The algorithmic foundations of differential privacy*, Foundations and Trends in Theoretical Computer Science, 9 (2014), pp. 1–277.

[23] C. D.WORK, *A firm foundation for private data analysis*, Commun. ACM, 54 (2011), p. 86,Äì95.

[24] C. DWORK, F. MCSHERRY, K. NISSIM, AND A. SMITH, *Calibrating noise to sensitivity in private data analysis*, in Theory of Cryptography, S. Halevi and T. Rabin, eds., Berlin, Heidelberg, 2006, Springer Berlin Heidelberg, pp. 265–284.

[25] B. EDELMAN, M. OSTROVSKY, AND M. SCHWARZ, *Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords*, American economic review, 97 (2007), pp. 242–259.

[26] P. ERDOS, A. RÉNYI, AND P. SZÜSZ, *On engel's and sylvester's series*, Ann. Univ. L. Eötvös (Sect. Math.), 1 (1958), pp. 7–32.

[27] N. FARAJIAN AND K. ZAMANIFAR, *Market-Driven Continuous Double Auction Method for Service Allocation in Cloud Computing*, in Proc. of ICAC3, 2013, pp. 14–24.

[28] M. FARAJTABAR, Y. WANG, M. GOMEZ RODRIGUEZ, S. LI, H. ZHA, AND L. SONG, *Coevolve: A joint point process model for information diffusion and network co-evolution*, Advances in Neural Information Processing Systems, 28 (2015).

[29] FIRDANIZA, B. N. RUCHJANA, AND D. CHAERANI, *Information diffusion model using continuous time markov chain on social media*, 1722 (2021), p. 012091.

[30] G. FU, F. CHEN, J. LIU, AND J. HAN, *Analysis of competitive information diffusion in a group-based population over social networks*, Physica A: Statistical Mechanics and its Applications, 525 (2019), pp. 409–419.

[31] A. GHOSH AND A. ROTH, *Selling privacy at auction*, Games and Economic Behavior, 91 (2015), pp. 334–346.

[32] S. GOEL, D. J. WATTS, AND D. G. GOLDSTEIN, *The structure of online diffusion networks*, in Proceedings of the 13th ACM Conference on Electronic Commerce, EC '12, New York, NY, USA, 2012, Association for Computing Machinery, p. 623‚Äì638.

[33] S. GOEL, D. J. WATTS, AND D. G. GOLDSTEIN, *The structure of online diffusion networks*, in Proceedings of the 13th ACM conference on electronic commerce, 2012, pp. 623–638.

[34] A. V. GOLDBERG AND J. D. HARTLINE, *Competitive auctions for multiple digital goods*, in European Symposium on Algorithms, Springer, 2000, pp. 416–427.

[35] A. V. GOLDBERG, J. D. HARTLINE, AND A. WRIGHT, *Competitive auctions and digital goods*, in Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, 2001, pp. 735–744.

[36] GRANOVETTER AND MARK, *Threshold models of collective behavior*, American Journal of Sociology, 83 (1978), pp. 1420–1443.

[37] T. GROVES ET AL., *Incentives in teams*, Econometrica, 41 (1973), pp. 617–631.

[38] J. GUO, T. CHEN, AND W. WU, *A multi-feature diffusion model: Rumor blocking in social networks*, IEEE/ACM Transactions on Networking, 29 (2021), pp. 386–397.

[39] X. HE, G. SONG, W. CHEN, AND Q. JIANG, *Influence blocking maximization in social networks under the competitive linear threshold model*, in Proceedings of the 2012 siam international conference on data mining, SIAM, 2012, pp. 463–474.

[40] X. HU, T. ZHU, X. ZHAI, H. WANG, W. ZHOU, AND W. ZHAO, *Privacy data diffusion modeling and preserving in online social network*, IEEE Transactions on Knowledge and Data Engineering, (2022), pp. 1–1.

[41] X. HU, T. ZHU, X. ZHAI, W. ZHOU, AND W. ZHAO, *Privacy data propagation and preservation in social media: a real-world case study*, IEEE Transactions on Knowledge and Data Engineering, (2021), pp. 1–1.

[42] Y. Hu, R. J. Song, and M. Chen, *Modeling for information diffusion in online social networks via hydrodynamics*, IEEE Access, PP (2017), pp. 1–1.

[43] M. Jackson, *Social and Economic Networks*, Princeton University Press, 2008.

[44] M. O. Jackson, *Social and economic networks*, Princeton university press, 2010.

[45] V. Karwa, S. Raskhodnikova, A. Smith, and G. Yaroslavtsev, *Private analysis of graph structure*, Proceedings of the VLDB Endowment, 4 (2011), pp. 1146–1157.

[46] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, *Topological generalizations of network motifs*, Phys Rev E Stat Nonlin Soft Matter Phys, 70 (2003), p. 031909.

[47] S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, and A. Smith, *Analyzing graphs with node differential privacy*, in Theory of Cryptography Conference, Springer, 2013, pp. 457–476.

[48] Kempe., *Maximizing the spread of influence through a social network*, Proc.of Acm Sigkdd Intl Conf.on Knowledge Discovery Data Mining, (2003).

[49] D. Kempe, J. Kleinberg, and É. Tardos, *Influential nodes in a diffusion model for social networks*, in International Colloquium on Automata, Languages, and Programming, Springer, 2005, pp. 1127–1138.

[50] M. Kimura, K. Saito, and R. Nakano, *Extracting influential nodes for information diffusion on a social network*, in Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2, AAAI'07, AAAI Press, 2007, p. 1371,Äì1376.

[51] V. Krishna, *Auction theory*, Academic press, 2009.

[52] J. Leskovec, J. Kleinberg, and C. Faloutsos, *Graph evolution: Densification and shrinking diameters*, ACM transactions on Knowledge Discovery from Data (TKDD), 1 (2007), pp. 2–es.

[53] B. Li, D. Hao, D. Zhao, and T. Zhou, *Mechanism design in social networks*, in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 586–592.

[54] D. LI AND J. LIU, *Modeling influence diffusion over signed social networks*, IEEE Transactions on Knowledge and Data Engineering, 33 (2019), pp. 613–625.

[55] D. LI, S. ZHANG, X. SUN, H. ZHOU, S. LI, AND X. LI, *Modeling information diffusion over social networks for temporal dynamic prediction*, IEEE Transactions on Knowledge and Data Engineering, 29 (2017), pp. 1985–1997.

[56] L. LI, D. WU, J. WU, H. LI, W. LIN, AND A. C. KOT, *Image sharpness assessment by sparse representation*, IEEE Transactions on Multimedia, 18 (2016), pp. 1085–1097.

[57] Y. LI, J. FAN, Y. WANG, AND K. L. TAN, *Influence maximization on social graphs: A survey*, IEEE Transactions on Knowledge & Data Engineering, (2018), pp. 1–1.

[58] Y. LI, J. FAN, Y. WANG, AND K.-L. TAN, *Influence maximization on social graphs: A survey*, IEEE Transactions on Knowledge and Data Engineering, 30 (2018), pp. 1852–1872.

[59] Z. LIANG AND W. YOUGUO, *Rumor diffusion model with spatio-temporal diffusion and uncertainty of behavior decision in complex social networks*, Physica A: Statistical Mechanics and its Applications, 502 (2018), pp. 29–39.

[60] Y. LIN, W. CHEN, AND J. C. LUI, *Boosting information spread: An algorithmic approach*, in 2017 IEEE 33rd International Conference on Data Engineering (ICDE), IEEE, 2017, pp. 883–894.

[61] L. LIU, B. QU, B. CHEN, A. HANJALIC, AND H. WANG, *Modelling of information diffusion on social networks with applications to wechat*, Physica A: Statistical Mechanics and its Applications, 496 (2018), pp. 318–329.

[62] C. MA ET AL., *Dynamical analysis of rumor spreading model with impulse vaccination and time delay*, Physica A: Statistical Mechanics and its Applications, 471 (2017), pp. 653–665.

[63] Y. MATSUBARA, Y. SAKURAI, B. A. PRAKASH, L. LI, AND C. FALOUTSOS, *Rise and fall patterns of information diffusion: model and implications*, in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 6–14.

[64] F. McSherry and K. Talwar, *Mechanism design via differential privacy.*, in FOCS, vol. 7, 2007, pp. 94–103.

[65] R. Milo, S. Shen-Orr, S. Ltzkovitz, N. Kashtan, and U. Alan, *Network motifs: Simple building blocks of complex networks*, (2011).

[66] Y. Moreno, M. Nekovee, and A. F. Pacheco, *Dynamics of rumor spreading in complex networks*, Phys Rev E Stat Nonlin Soft Matter Phys, 69 (2004), p. 066130.

[67] F. Morone and H. A. Makse, *Influence maximization in complex networks through optimal percolation*, Nature, 524 (2015).

[68] R. B. Myerson, *Optimal auction design*, Mathematics of operations research, 6 (1981), pp. 58–73.

[69] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili, *Theory of rumour spreading in complex social networks*, Physica A: Statistical Mechanics and its Applications, 374 (2007), pp. 457–470.

[70] M. Newman, *Newman mej.. the structure and function of complex networks. siam rev 45: 167-256*, SIAM Review, 45 (2003).

[71] M. E. Newman, *Modularity and community structure in networks*, Proceedings of the national academy of sciences, 103 (2006), pp. 8577–8582.

[72] H. T. Nguyen, T. P. Nguyen, T. N. Vu, and T. N. Dinh, *Outward influence and cascade size estimation in billion-scale networks*, Proceedings of the ACM on Measurement and Analysis of Computing Systems, 1 (2017), pp. 1–30.

[73] H. T. Nguyen, T. P. Nguyen, T. N. Vu, and T. N. Dinh, *Outward influence and cascade size estimation in billion-scale networks*, ACM SIGMETRICS Performance Evaluation Review, 44 (2017), pp. 63–63.

[74] K. Q. Nguyen and J. Traoré, *An online public auction protocol protecting bidder privacy*, in Australasian Conference on Information Security and Privacy, Springer, 2000, pp. 427–442.

[75] R. Pastor-Satorras and A. Vespignani, *Epidemic spreading in scale-free networks*, Physical review letters, 86 (2001), p. 3200.

[76] L. Pei, X. Kai, D. Wang, Z. Xin, and W. Hui, *Information diffusion in facebook-like social networks under information overload*, International Journal of Modern Physics C, 24 (2013), pp. 1350047–.

[77] H. Peng, J. Li, Q. Gong, Y. Ning, and L. He, *Motif-matching based subgraph-level attentional convolutional network for graph classification*, Proceedings of the AAAI Conference on Artificial Intelligence, 34 (2020), pp. 5387–5394.

[78] Z. Qin, J. Fan, Y. Liu, Y. Gao, and G. Y. Li, *Sparse representation for wireless communications: A compressive sensing approach*, IEEE Signal Processing Magazine, 35 (2018), pp. 40–58.

[79] S. Raskhodnikova and A. Smith, *Efficient lipschitz extensions for high-dimensional graph statistics and node private degree distributions*, arXiv preprint arXiv:1504.07912, (2015).

[80] F. D. Sahneh, C. Scoglio, and P. Van Mieghem, *Generalized epidemic mean-field model for spreading processes over multilayer complex networks*, IEEE/ACM Transactions on Networking, 21 (2013), pp. 1609–1620.

[81] A. Sankar, J. Wang, A. Krishnan, and H. Sundaram, *Beyond localized graph neural networks: An attributed motif regularization framework*, in 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 2020, pp. 472–481.

[82] S. Sathe, *Rumor spreading in livejournal*, Scientific production and competences, (2008).

[83] F. Schweitzer, G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani, and D. R. White, *Economic networks: The new challenges*, science, 325 (2009), pp. 422–425.

[84] E. Serrano, C. Á. Iglesias, and M. Garijo, *A novel agent-based rumor spreading model in twitter*, in Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 811–814.

[85] J. Shin, L. Jian, K. Driscoll, and F. Bar, *The diffusion of misinformation on social media: Temporal pattern, message, and source*, Computers in Human Behavior, 83 (2018), pp. 278–287.

[86] J. Shin and K. Thorson, *Partisan selective sharing: The biased diffusion of fact-checking messages on social media*, Journal of Communication, 67 (2017), pp. 233–255.

[87] E. Stai, E. Milaiou, V. Karyotis, and S. Papavassiliou, *Temporal dynamics of information diffusion in twitter: Modeling and experimentation*, IEEE Transactions on Computational Social Systems, 5 (2018), pp. 256–264.

[88] Q. Su, J. Huang, and X. Zhao, *An information propagation model considering incomplete reading behavior in microblog*, Physica A Statistical Mechanics Its Applications, 419 (2015), pp. 55–63.

[89] Y. Su, X. Zhang, S. Wang, B. Fang, T. Zhang, and P. S. Yu, *Understanding information diffusion via heterogeneous information network embeddings*, in International Conference on Database Systems for Advanced Applications, Springer, 2019, pp. 501–516.

[90] A. Subramonian, *Motif-driven contrastive learning of graph representations*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 15980–15981.

[91] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, *Want to be retweeted? large scale analytics on factors impacting retweet in twitter network*, in 2010 IEEE second international conference on social computing, IEEE, 2010, pp. 177–184.

[92] L. Sun, W. Huang, P. S. Yu, and W. Chen, *Multi-round influence maximization*, in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2249–2258.

[93] L. Sun, W. Huang, P. S. Yu, and W. Chen, *Multi-round influence maximization*, KDD ’18, New York, NY, USA, 2018, Association for Computing Machinery, p. 2249,Äì2258.

[94] A. Susarla, J.-H. Oh, and Y. Tan, *Social networks and the diffusion of user-generated content: Evidence from youtube*, Information systems research, 23 (2012), pp. 23–41.

[95] A. Tafti, R. Zotti, and W. Jank, *Real-time diffusion of information on twitter and the financial markets*, Plos One, 11 (2016), p. e0159226.

[96] C. TASK AND C. CLIFTON, *A guide to differential privacy theory in social network analysis*, in 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE, 2012, pp. 411–417.

[97] G. TONG, W. WU, L. GUO, D. LI, C. LIU, B. LIU, AND D.-Z. DU, *An efficient randomized algorithm for rumor blocking in online social networks*, IEEE Transactions on Network Science and Engineering, 7 (2017), pp. 845–854.

[98] G. A. TONG, W. WU, G. LING, D. LI, AND D. Z. DU, *An efficient randomized algorithm for rumor blocking in online social networks*, IEEE, (2017).

[99] O. TSUR AND A. RAPPOPORT, *What's in a hashtag? content based prediction of the spread of ideas in microblogging communities*, in Proceedings of the fifth ACM international conference on Web search and data mining, 2012, pp. 643–652.

[100] W. VICKREY, *Counterspeculation, auctions, and competitive sealed tenders*, The Journal of finance, 16 (1961), pp. 8–37.

[101] M. VIDAL, M. CUSICK, AND A. BARABASI, *Interactome networks and human disease.*, Cell, 144 (2011), pp. 986–998.

[102] M. VIDAL, M. E. CUSICK, AND A.-L. BARABASI, *Interactome networks and human disease*, Cell, 144 (2011), pp. 986–998.

[103] S. VOSOUGHI, D. ROY, AND S. ARAL, *The spread of true and false news online*, Science, 359 (2018), pp. 1146–1151.

[104] J.-W. WANG AND L.-L. RONG, *Cascade-based attack vulnerability on the us power grid*, Safety science, 47 (2009), pp. 1332–1336.

[105] J. W. WANG AND L. L. RONG, *Cascade-based attack vulnerability on the us power grid*, Safety Science, 47 (2009), pp. 1332–1336.

[106] Y. WANG, D. CHAKRABARTI, C. WANG, AND C. FALOUTSOS, *Epidemic spreading in real networks: An eigenvalue viewpoint*, in 22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings., IEEE, 2003, pp. 25–34.

[107] D. J. WATTS AND S. H. STROGATZ, *Collective dynamics of ‚Äòsmall-world'networks*, nature, 393 (1998), p. 440.

[108] Y. H. Wen, L. Gao, H. Fu, F. L. Zhang, and S. Xia, *Graph cnns with motif and variable temporal block for skeleton-based action recognition*, Proceedings of the AAAI Conference on Artificial Intelligence, 33 (2019), pp. 8989–8996.

[109] K. Words, M. Y. Li, and J. S. Muldowney, *Global stability for the seir model in epidemiology*, Mathematical Biosciences, 125 (1995), pp. 155–164.

[110] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, *Robust face recognition via sparse representation*, IEEE transactions on pattern analysis and machine intelligence, 31 (2008), pp. 210–227.

[111] M. Wright and M. P. Wellman, *Evaluating the Stability of Non-Adaptive Trading in Continuous Double Auctions*, in Proc. of AAMAS, 2019, pp. 614–622.

[112] X. Wu, L. Fu, H. Long, D. Yang, Y. Lu, X. Wang, and G. Chen, *Adaptive diffusion of sensitive information in online social networks*, IEEE Transactions on Knowledge and Data Engineering, (2020), pp. 1–1.

[113] P. Xiong, L. Zhang, T. Zhu, G. Li, and W. Zhou, *Private collaborative filtering under untrusted recommender server*, Future generation computer systems, (2018), pp. 1–10.

[114] Q. Xu, Z. Su, K. Zhang, P. Ren, and X. S. Shen, *Epidemic information dissemination in mobile social networks with opportunistic links*, IEEE Transactions on Emerging Topics in Computing, 3 (2015), pp. 399–409.

[115] A. Xz, B. Wz, A. Gf, L. A. Cai, and C. Gha, *Network sparse representation: Decomposition, dimensionality-reduction and reconstruction*, Information Sciences, 521 (2020), pp. 307–325.

[116] C. Yang, M. Sun, H. Liu, S. Han, Z. Liu, and H. Luan, *Neural diffusion model for microscopic cascade study*, IEEE Transactions on Knowledge and Data Engineering, 33 (2019), pp. 1128–1139.

[117] J. Yang, J. Wright, T. S. Huang, and Y. Ma, *Image super-resolution via sparse representation*, IEEE transactions on image processing, 19 (2010), pp. 2861–2873.

[118] L.-X. Yang, X. Yang, and Y. Y. Tang, *A bi-virus competing spreading model with generic infection rates*, IEEE Transactions on Network Science and Engineering, 5 (2017), pp. 2–13.

[119] R. YANG, B. H. WANG, J. REN, W. J. BAI, Z. W. SHI, W. X. WANG, AND Z. TAO, *Epidemic spreading on heterogeneous networks with identical infectivity*, Physics Letters A, 364 (2007), pp. 189–193.

[120] Z. YANG, J. GUO, K. CAI, J. TANG, J. LI, L. ZHANG, AND Z. SU, *Understanding retweeting behaviors in social networks*, in Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 1633–1636.

[121] D. YE, T. ZHU, W. ZHOU, AND P. S. YU, *Differentially Private Malicious Agent Avoidance in Multiagent Advising Learning*, IEEE Transactions on Cybernetics, (2019), pp. 1–14.

[122] L. YU, P. CUI, F. WANG, C. SONG, AND S. YANG, *From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics*, in 2015 IEEE international conference on data mining, IEEE, 2015, pp. 559–568.

[123] N. J. YUAN, Y. ZHONG, F. ZHANG, X. XIE, C.-Y. LIN, AND Y. RUI, *Who will reply to/retweet this tweet? the dynamics of intimacy from online social interactions*, in Proceedings of the ninth ACM international conference on web search and data mining, 2016, pp. 3–12.

[124] L. A. ZAGER AND G. C. VERGHESE, *Graph similarity scoring and matching*, Applied Mathematics Letters, 21 (2008), pp. 86–94.

[125] A. ZAREZADE, A. KHODADADI, M. FARAJTABAR, H. R. RABIEE, AND H. ZHA, *Correlated cascades: Compete or cooperate*, in Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[126] J. ZHANG, J. TANG, J. LI, Y. LIU, AND C. XING, *Who influenced you? predicting retweet via social influence locality*, ACM Transactions on Knowledge Discovery from Data (TKDD), 9 (2015), pp. 1–26.

[127] J. ZHANG, P. S. YU, Y. LV, AND Q. ZHAN, *Information diffusion at workplace*, in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, 2016, pp. 1673–1682.

[128] T. ZHANG, T. ZHU, P. XIONG, H. HUO, Z. TARI, AND W. ZHOU, *Correlated Differential Privacy: Feature Selection in Machine Learning*, IEEE Transactions on Indistrial Informatics, 16 (2019), pp. 2115–2124.

[129] D. ZHAO, B. LI, J. XU, D. HAO, AND N. R. JENNINGS, *Selling multiple items via social networks*, in Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 68–76.

[130] L. ZHAO, H. CUI, X. QIU, X. WANG, AND J. WANG, *Sir rumor spreading model in the new media age*, Physica A: Statistical Mechanics and its Applications, 392 (2013).

[131] L. ZHAO, J. WANG, Y. CHEN, Q. WANG, J. CHENG, AND H. CUI, *Sihr rumor spreading model in social networks*, Physica A: Statistical Mechanics and its Applications, 391 (2012), pp. 2444–2453.

[132] L. ZHAO, Q. WANG, J. CHENG, Y. CHEN, J. WANG, AND W. HUANG, *Rumor spreading model with consideration of forgetting mechanism: A case of online blogging livejournal*, Physica A: Statistical Mechanics and its Applications, (2011).

[133] L. ZHAO, Q. WANG, J. CHENG, Y. CHEN, J. WANG, AND W. HUANG, *Rumor spreading model with consideration of forgetting mechanism: A case of online blogging livejournal*, PHYSICA A, 390 (2011), pp. 2619–2625.

[134] C. ZHENG, C. XIA, Q. GUO, AND M. DEHMER, *Interplay between sir-based disease spreading and awareness diffusion on multiplex networks*, Journal of Parallel and Distributed Computing, 115 (2018), pp. 20–28.

[135] J. ZHOU, Z. LIU, AND B. LI, *Influence of network structure on rumor propagation*, Physics Letters A, 368 (2007), pp. 458–463.

[136] J. ZHOU, Z. LIU, AND B. LI, *Influence of network structure on rumor propagation*, Physics Letters A, 368 (2007), pp. 458–463.

[137] J. ZHOU, J. SHEN, AND Q. XUAN, *Data augmentation for graph classification*, CIKM '20, New York, NY, USA, 2020, Association for Computing Machinery, p. 2341,Äì2344.

[138] X. ZHOU, M. ZHU, S. LEONARDOS, AND K. DANIILIDIS, *Sparse representation for 3d shape estimation: A convex relaxation approach*, IEEE transactions on pattern analysis and machine intelligence, 39 (2016), pp. 1648–1661.

[139] H. Zhu, C. Huang, and H. Li, *Information diffusion model based on privacy setting in online social networking services*, The Computer Journal, 58 (2015), pp. 536–548.

[140] L. Zhu and Y. Wang, *Rumor spreading model with noise interference in complex social networks*, Physica A: Statistical Mechanics and its Applications, 469 (2017), pp. 750–760.

[141] L. Zhu, H. Zhao, and H. Wang, *Complex dynamic behavior of a rumor propagation model with spatial-temporal diffusion terms*, Information Sciences, 349 (2016), pp. 119–136.

[142] T. Zhu, G. Li, W. Zhou, and S. Y. Philip, *Differentially private data publishing and analysis: A survey*, IEEE Transactions on Knowledge and Data Engineering, 29 (2017), pp. 1619–1638.

[143] T. Zhu, J. Li, X. Hu, P. Xiong, and W. Zhou, *The dynamic privacy-preserving mechanisms for online dynamic social networks*, IEEE Transactions on Knowledge and Data Engineering, PP (2020), pp. 1–1.

[144] T. Zhu and S. Y. Philip, *Applying differential privacy mechanism in artificial intelligence*, in 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), IEEE, 2019, pp. 1601–1609.

[145] T. Zhu, M. Yang, P. Xiong, Y. Xiang, and W. Zhou, *An iteration-based differentially private social network data release*, Computer systems science and engineering, 33 (2018), pp. 61–69.