國 立 交 通 大 學

電控工程研究所

碩士論文

社群網站資料流動分析模型之研究:以Facebook為例

Information Diffusion Model of Online Social Networks: A Case Study of Facebook

研 究 生: 陳玟媗

Student: CHEN, WEN-HSUAN

指導教授: 黃育綸 博士

Advisor: Dr. Yu-Lun Huang

中華民國 101 年 6 月

社群網站資料流動分析模型之研究:以Facebook為例

Information Diffusion Model of Online Social Networks: A Case Study of Facebook

研 究 生：陳玟媗　　Student: CHEN, WEN-HSUAN

指導教授：黃育綸　博士　　Advisor: Dr. Yu-Lun Huang

國 立 交 通 大 學

電控工程研究所

碩士論文

A Thesis

Submitted to Institute of Electrical Control Engineering

College of Electrical Engineering

National Chiao Tung University

in partial Fulfill of the Requirements

for the Degree of

Master

in

Institute of Electrical Control Engineering

June, 2012

Hsinchu, Taiwan, Republic of China

中華民國 101 年 6 月

社群網站資料流動分析模型之研究:以Facebook為例

學生：陳玟媗　　指導教授：黃育綸　博士

國 立 交 通 大 學 電控工程研究所（研究所）班

摘　　要

社群網站(例如 Facebook、Twitter、及Google+等)提供使用者一個分享及取得資訊的新平台，使用者之間的互連性因其互動行為變動而展現不同的強度關係。現有研究主要探討如何透過使用者在虛擬世界的互動關係，及個人資訊，逆推得到真實環境之人際關係。此外，也有其他研究嘗試探討社群網路在資訊擴散中所扮演的角色。然而，這些研究並未討論虛擬人際關係對資訊流動的影響，例如當某個使用者被植入殭屍病毒，則該使用者有可能會變成攻擊者，散播惡意連結給他的朋友，並進而造成大範圍的網路感染。 在此論文當中，我們針對資料流動提出一個分析方法。我們的方法包含兩個階段：第一階段我們先建立分析模型，利用使用者間的互動，計算出兩兩使用者的關係強度與回應比例。透過第一階段所得到的關係強度與回應比例，我們在第二階段進一步地建立單點及多點資訊擴散模型，並預測資料的流動路徑、評估該使用者對資訊擴散的影響。以Facebook為研究案例，我們利用Facebook上實際的數據，以驗證我們所提方法的正確性。實驗結果顯示，我們的方法可以預測每個受測者所張貼的資訊的可能擴散範圍。 相較於現有的研究，我們的方法可以找出影響資料擴散的弱連結，並篩選出對資訊擴散影響力最大的使用者群。

Information Diffusion Model of Online Social Networks: A Case Study of Facebook

Student: CHEN, WEN-HSUAN Advisor: Dr. Yu-Lun Huang

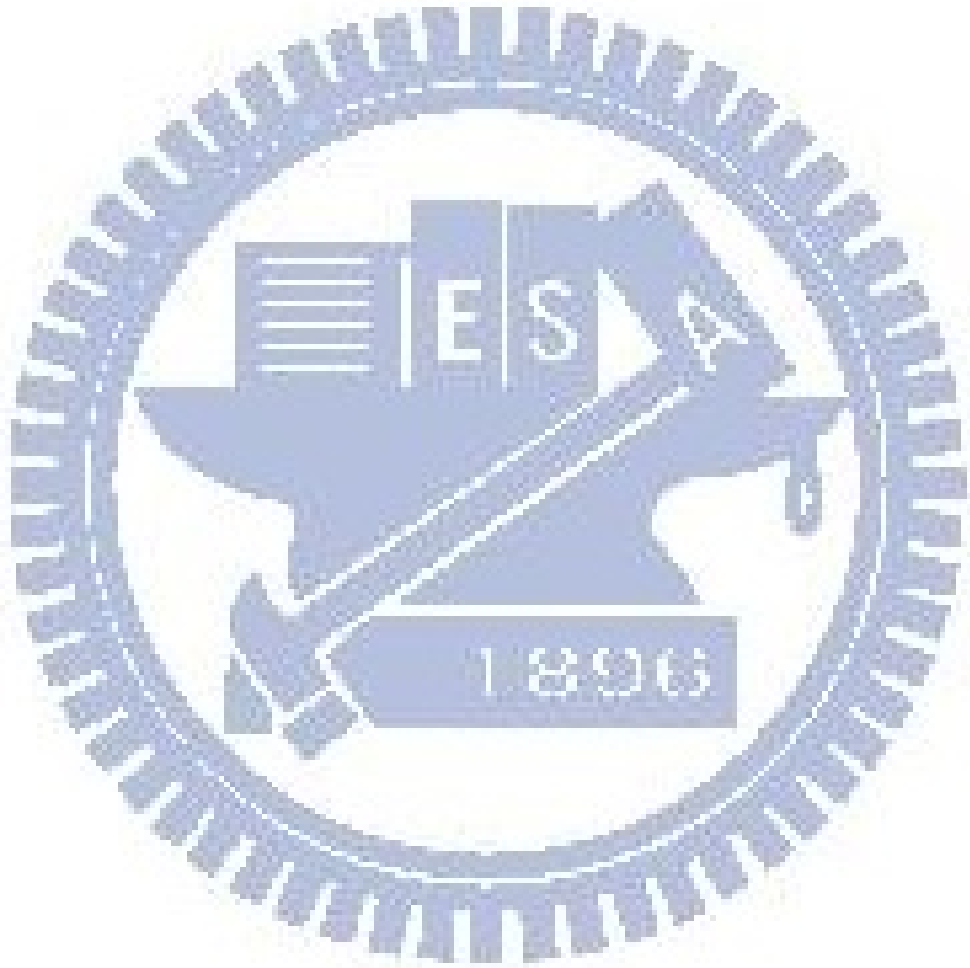
Institute of Electrical Control Engineering

National Chiao Tung University

Abstract

Online social networks (OSNs), like Facebook, Twitter and Google+, have created a novel way for people to connect with each other by sharing and obtaining information. As user behaviours vary, their inter-connectivities also reveal different levels of tie strengths within the OSNs accordingly. By researching the similarities of user profiles and interactions, some existing works have presented the formal analysis to explain dyad relationships; some others have explained OSNs roles in information diffusion. However, none of them addresses the security impact resulted from a user who tries to spread his information. If such a user is attacked with his post injected by any malicious link, the user then becomes a bot master to disperse the information to his friends and potentially taint the network at large scale. In this thesis, we propose a method to predict the information diffusion of a post within an OSN. In addition to tie strengths of dyads, we claim that responding patterns and rates, along with the involvements of friend's friends, should also be taken in account when predicting the possible delivering paths of information. In our method, we first estimate the one-hop information diffusion, then we can derive multi-hop information diffusion accordingly. We conduct several experiments to verify the validity of our predictions with the real data obtained from the Facebook website. The experiment results show that our method predicts the diffusion coverage of a piece of information from a

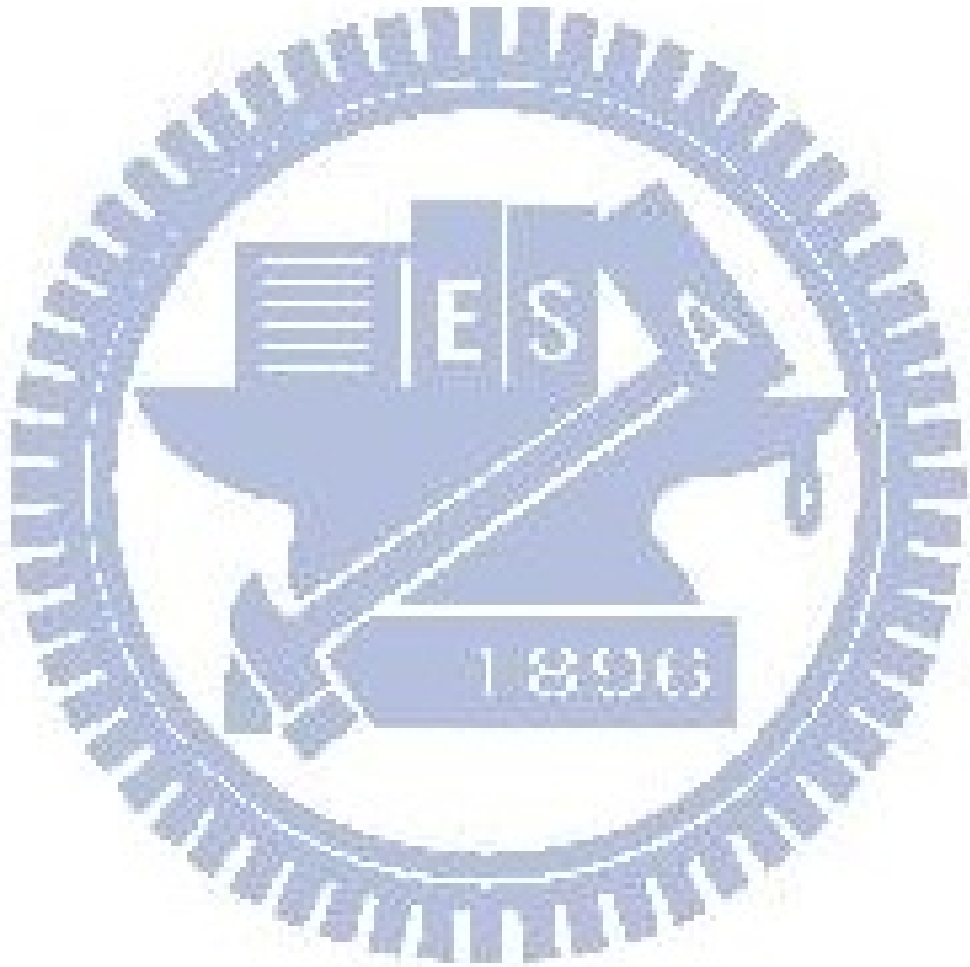
specified individual.



Contents

List of Figures

List of Tables



Chapter 1

Introduction

Online social networks (so called OSNs), such as Facebook, LinkedIn, Google+, Tiwtter and Plurk, have created a great way for people to connect with each other by sharing and obtaining information. People can connect to hundreds, even more than thousands of people even more in online social networks. As user behaviours vary, their inter-connectivities also reveal different levels of tie strengths within the OSN accordingly.

1.1 Online Social Network

The Internet consists of different information sharing systems, including the Web, e-mail, etc. Unlike traditional Web sites providing contents only, online social networks further comprise user interactions. An online social network, one of web-based services, is defined by Danah M. Boyd and Nicole B. Ellison [1]. An online social network lets users to hinge a list of other users with whom they share a connection within this bounded system. Users view and bind their connections and the lists of connections made by others within this system. Also, users of an online social network are able to post the information with multimedia contents about themselves, their interests or concerns to their friends [1].

Internet users can not only communicate with their friends, but also search for new friendships in the online social network. Online social networks are dramatically drawing people into the online world for two most significant reasons. One is the ability to form different networks with people having different backgrounds. The other is that users can freely create and share

the information with low cost [2]. The popularity of online social networks attracts the attention of researches such that many studies intend to understand its impact on human life.

In recent years, online social networks have become part of people's lives because of its convenience. More and more people use OSNs to keep in touch with each other. 57% of people talk more online than in real life [3]. Several stunting statistics remind us to pay attention to the cause of online social networks. In [4], Mislove *et al.* presented the large-scale measurements of several online social networks and thus analyzed their structures. The structure of online social networks has following characteristics: *power-law*, *scale-free* and *small-world*. The online social network can be represented by graph. The power-law property means that, the network has a large number of highly connected clusters consist of relatively low-degree nodes, while only 10% of the nodes have high degree. The scale-free property indicates that, nodes with high in-degree tend to have high out-degree. The small-world property shows a high degree of reciprocity. That is, the network exists a tight core consisting of high-degree nodes and a strong positive correlation in the link degrees for connected users. The study of Mislove *et al.* differs social networks from general networks by Newman and Park [5] as well. Besides, social networking sites are being used for advertisement and e-commerce.

Granovetter [6] indicated conceivably that who know few about each other forming a weak tie plays a important role in information adoption because of the specialization in a variety way of the users' relationship. Also, Manuel E. Sosa [7] claimed that sporadic and distant dyadic relationships foster individual creativity. In this thesis, we focus on Facebook, which is one of the most famous online social network sites in the world.

1.2 Facebook

From the statistics announced in OnlineSchool site, the increment of Facebook users exceeds 200 million in less than one year in 2011 [3]. In March 2012, Facebook has 901 million users and 125 billion friendships. Mashable, the largest independent online news site, released a piece of news in April 2012 that Facebook had 526 million daily active users [8].

Every day, more than one billion pieces of information have been shared on Facebook. The average post of information by OSNs users is 90 piece per month. On average, they spend 23 minutes on each visit to Facebook. 70% of local businesses use Facebook for marketing and promotion. Rather than traditional news broadcast, over 50% of people have learned about breaking news via social media, while 59.5% of Facebook users do as well. The breaking news that have been heard by 49.1% of people via social media turned out to be false.

These shocking statistics bring about many discussions in diverse fields, such as marketing, sociology and network security. CNN, a news media with high credibility, reported that "*With a user base of some 800 million users, Facebook is fertile hunting ground for scammers and hackers. Often, users who click bad links will be infected with malware that causes them to, in turn, share the bad link with their friends.*" cited from *Facebook: Are you sure you want to click on that?* [9]. Facebook even provides a massive defence system of social networks, called *Facebook immune system* [10] to keep the users safe from spam and cyber attacks. These phenomena make people aware ofthe network security issue. Hence, several models, *independent cascade model* and *linear threshold model*, for information diffusion have been proposed as the effect of "word of mouth" [11]. As a consequence of viral marketing strategies [12] and problem of cascading fail in power systems [13], many researches posed the fundamental algorithmic problems of information diffusion in online social network. Also, these researches built *descriptive* models from mathematics sociology [14] to expicitly represent the dynamics

of information adoption step by step.

Moreover, privacy and safety issues are rising on OSNs. Many young parents flood photo-sharing and social network sites with the photos of their children. This behavior gives the chance for criminals to fake profile using childrens' photos. In the New York Times report, a stranger created a fake profile using children's photo. Therefore, the private information are exposed to strangers indirectly. "*The real danger is that a photo is appropriated and mis-treated.*"cited from New York Times: *Is It Safe to post Children's Images on Online Photo Sites* [15]. Therefore, much of the existing works have focused on analyzing the relational patterns of social networks. By using the social media data sufficiently, the existing works have found the connections between the practical experiment and sociological theory. Some research have extracted the features influencing the tie strengths [16] [17]. Some others have shed light on how the tie strength affecting the whole networks [18].

1.3 Tie Strength

In Granvotter's study [6], "*tie strength*" is a linear combination of *the amount of time*, *the emotional intensity*, *the intimacy* and *the reciprocal services*. The *amount of time* means the time two individuals spending with each other. The longer time the dyad know each other, the stronger tie strength they have. The *emotional intensity* indicate the frequency that two individuals interact with each other. The higher frequency the dyad interact with each other, the stronger sentiments of friendship for the dyad have. The *intimacy* symbolizes the mutual friends two individuals have or the intimacy words they use. The more intimacy the dyad show, the higher mutual confiding they have. The *reciprocal services* intend the informational, social or economic goods exchanged between two individuals. The more reciprocal services the dyad exchange, the more trustness they have.

In Granovetter's study, a 'tie' is defined as positive and symmetric relationship between two individuals. Tie strength is roughly divided into three levels: strong, weak and absent. Besides strong ties described above, a weak tie represents the dyad with distant relationship. An "absent" tie means "nodding" relationship. Two people "knowing" each other by name do not need to move their friendship out of this level if their interactions are negligible. Granovetter's study also indicates the strength of weak ties relateing to varied macro phenomena as diffusion, social mobility, political organization and social cohesion in general. In this thesis, we redefine tie strength in a different way for information diffusion. The scenario that two individuals are not friends on Facebook is not discussed. If two friends have no interactions on Facebook, an absent tie is assigned. If two friends have interactions on Facebook, either a strong tie or a weak tie is used to represent the strength (relationship) of the two individuals. Except absent ties, we discuss the impacts on information diffusion caused by strong and weak ties in this thesis.

1.4 Motivation

Unlike other Web sites, social network sites not only allowed users to obtain information but also to share the information with friends. Social network plays a dominate role in the spread of various information, such as innovations, hot topics, personal profiles and malicious links. Online social networks hence become a mean to disseminate information. However, people are used to receive information from online social networks, only parts of them would propagate the information to others.

Consequently, in this thesis, we blend tie strength into a diffusion model in order to establish an assessment tool for social network vulnerability. When an idea comes out in an online social network, the idea may either die out soon or be spread out quickly. Similar diffusion can be applied to malicious attacks. Under the premise that network services are available, we study

how information dissemination revel within an OSN: the extent of which users are likely to be influenced by their friends, or the extent to which ”word-of-mouth” effects will occur. Although the basic diffusion models mentioned above are widely-used, the user behaviours on online social networks change by the new functions provided by OSNs. In this thesis, we simplify the estimation of tie strengths according to our method. Then, we have adapted the diffusion model for network vulnerability by redefining the extent of a user affected by another user. Motivated by the role of social network in information diffusion [19], we intend to predict a user set that triggers a maximizing spread of influence through a social network.

1.5 Synopsis

The remaining sections are organized as follow: Chapter 2 introduces the existing works. In Chapter 3, we explain our method in detail. Chapter 4 shows the design of our experiment designs and verifications to prove the validity of our assumptions. We also demonstrate the results of information diffusion in the same chapter. Subsequently, in Chapter 5, we depict the proposed tie strength and responding rate for information diffusion in an OSN. We compare the proposed method to the existing works. Then, we discuss several phenomena found in the experiments. Finally, we conclude the thesis in Chapter 6.

Chapter 2

Existing Work

As the growth and popularity of OSNs such as Facebook, Google+, Tiwtter and LinkedIn in recent years, many research focused on extracting the features which influence the relationship between two individuals and building a model to describe the user's relationship by different inferences. Some works have discussed the network structure along with tie strengths. Taking Kivran-Swaine's study [18] as an example, they explore the network structure influencing the period that tie lasts on Twitter. Tie of two users breaks when these two individuals are not friends any more. They analyzed media data which varied with time by multilevel logistic regression. The study of Kivran-Swaine was showed that tie breaks depend negatively on reciprocity, the seed's follow-back rate, the follower's follower-to-followee ratio, the seed's network density and the dyad's number of common neighbors. Although it works well on Twitter, it may have different influential factors on different social networks, such as Facebook and Google+.

Some works have analyzed the growth patterns of social networks. Some other works have focused on the social network structure with auxiliary survey. Some works have figured out that the the time of articles shared distributes from five days to nine days. And the amount of information changing repeats per week by statistics. Furthermore, Facebook, Inc. also explored the role of social network in information diffusion by analyzing tie strengths and information exposure time. In this chapter, we introduce the Xiang's and Gilbert's methods and the role of social network in detail.

2.1 Xiang's Method

Xiang [20] estimated the tie strength strength between two users on Facebook and LinkedIn based on *homophily* [21] from sociology. The homophily principle indicates that people who have much more similar background tend to form ties with each other. According to Granovetter's definition [6], the stronger the ties are (e.g., close friends), the greater similarity the dyad has. Otherwise, those with weak ties (e.g., acquaintances) are inclined to interact rarely. Xiang built a link-based latent variable and unsupervised model to estimate link weights result from user interactions and similarities. The tie strength was considered as a *hidden effect* of user profile similarities and as the *hidden cause* of interactions as shown in Figure 2.1. In this Figure,

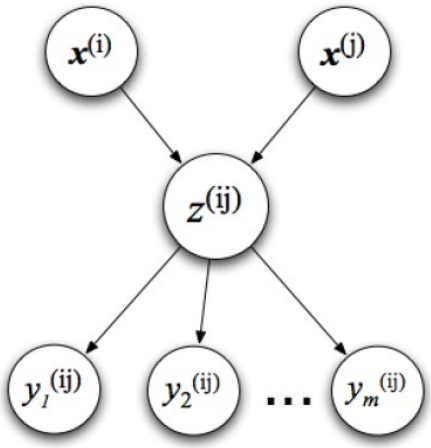


Figure 2.1: Graphical model representation of the general relationship strength model. Adopted from [20].

$x^{(i)}$ and $x^{(j)}$ represent the profile vectors of two individuals i and j. $z^{(ij)}$ is defined as a latent relationship. The model implied the influence of $x^{(i)}$ and $x^{(j)}$ on $z^{(ij)}$, as well as the influence of $z^{(ij)}$ on $y_t^{(ij)}$, $t = 1, 2, ...m$. The latent variable was estimated by Gaussian distribution given profile similarities and by joint probability as maximum likelihood given interactions, where

$$P(z^{(ij)}, \mathbf{y}^{(ij)}|x^{(i)}, x^{(j)}) = P(z^{(ij)}|x^{(i)}, x^{(j)}) \prod_{t=1}^{m} P(y_t^{(ij)})|z^{(ij)})$$

As a result, the estimation of tie strength was in higher autocorrelation of user profiles and can be used to improve the classification accuracy of user profiles. It was found that the autocorrelation varies with friendship density. However, the study of Xiang wasn't considered the dependencies between adjacent edges and the relationship evolving with time.

2.2 Gilbert's Method

According to Granovetter's definition [6], tie strength can basically be categorized into four dimensions: Intensity, Intimacy, Duration and Reciprocal Services. For instance, Gilbert [17] mapped media data on Facebook, which was divided into seven dimensions defined by Marsden [22], to tie strength and proposed a predictive model to distinguish *strong* and *weak* ties. Besides the Granovetter's definition, the additional dimensions of tie strength defined by Marsden are Social Distance, Emotional Support and Structural. Gilbert explored their model using original least square regression and compared the significant variables affecting the tie strength with auxiliary survey, which asked participants to rate their friendships as shown in Figure 2.2. Expanding the dimensions of tie strength [6], this work guided the feature selection. Gilbert's method found that more variables does not always represent exactly greater impact. Listed 74 variables to build a predictive model at first, the result of Gilbert's method showed that only 15 variables are needed to distinguish strong and weak tie strengths. The mapping accuracy is higher than 85%. Gilbert conducted the follow-up interviews about the most difficulty predicting cases to understand the limitations. The error anlysis shows that friendships are actually asymmetric. Also, the education difference, which is in 'social distance' dimension, impacts tie strengths little than the assumptions. Nevertheless, the observations with each participant were dependent. This observation is a common obstacle for building a model with ego-centric survey.

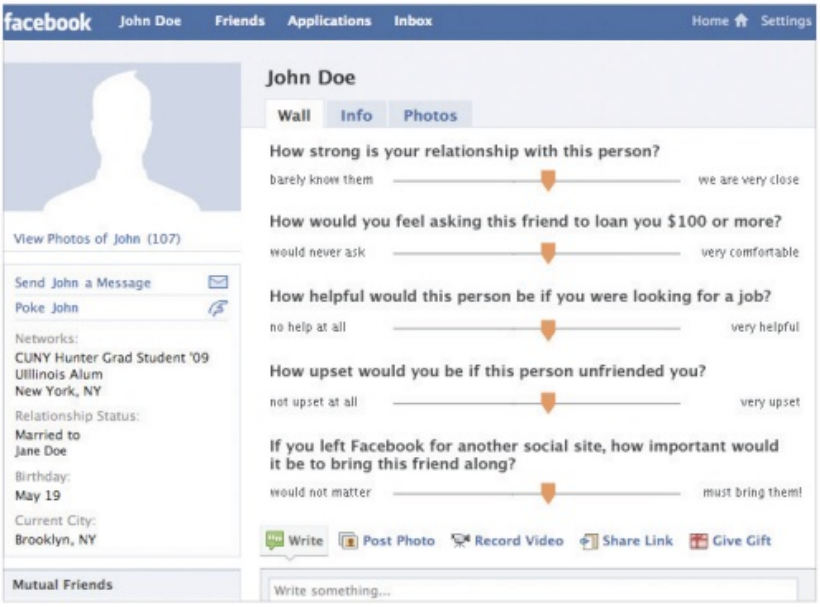


Figure 2.2: The auxiliary questions used to assess participants describing their friendships. Adopted from [17].

2.3 The Role of Social Network

Facebook released the latest noted news on **Facebook newsroom** [23] : *Rethinking information diversity in network.* The release pointed out that there is another important issue, which is information diffusion. Instead of spotlighting the tie strength modeling, Bakshy *et al.* [19] conducted a study on Facebook to examine the nature of information dissemination in social network. Modeling the effects on spread of information in a social network requires not only identifying the influences, but also whether an individual would still disseminate information in the absence of online social network. Bakshy *et al.* examine the role of online social network in information diffusion with large-scale field experiments. Those who are exposed to the information are dramatically more likely to propagate information, and also more quickly than those who are not exposed as shown in Figure 2.3. It was found that even though stronger ties are more influential, weak ties are still responsible for the dissemination of novel information as shown in Figure 2.4 . This work distinguished that weak tie plays a leading role in information diffusion online. Bakshy *et al.* shed light on trend over a large population and this study

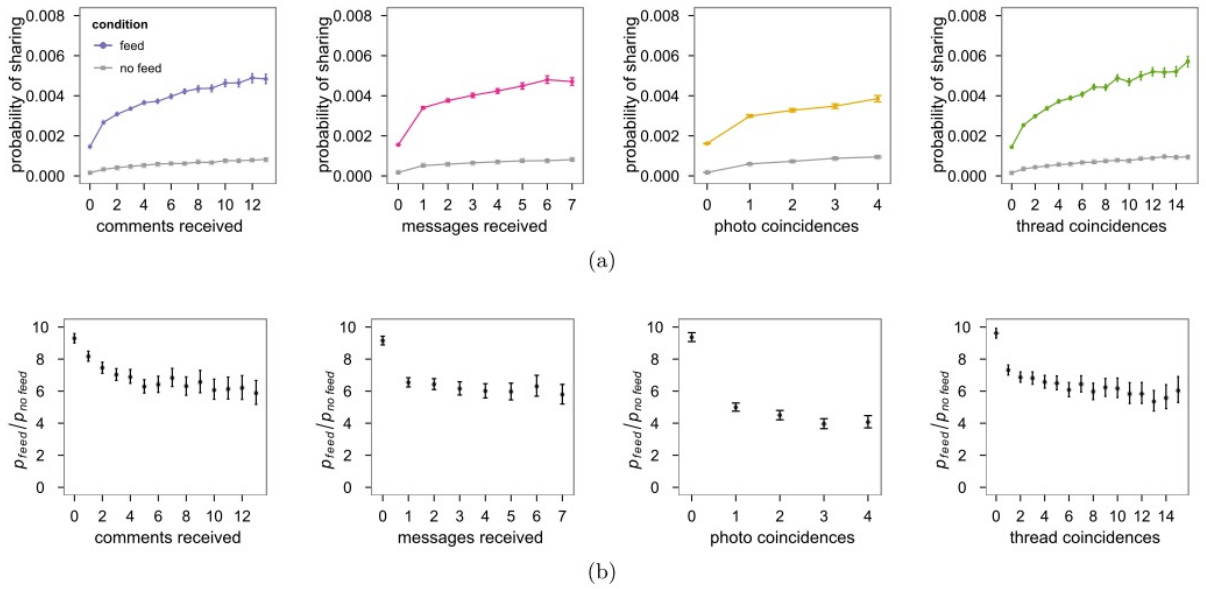


Figure 2.3: (a) The probability of sharing a link that a friend shared in the *feed* (exposed) and *no feed* (not exposed) conditions. (b) The multiplicative effect of feed decreases with tie strength. Adopted from [19].

painted a different picture of the world. Nevertheless, habitual patterns of individuals were not considered as the cause of information diffusion. The classification of strong and weak ties was also imprudent in this work.

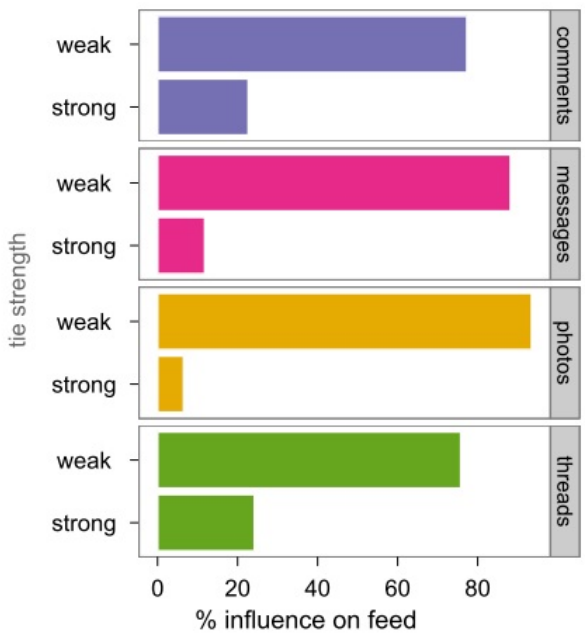


Figure 2.4: Weak ties play a more dominate role in information diffusion than strong ties do. Adopted from [19].

Chapter 3

Proposed Method

For evaluating the information diffusion, we proposed a method consisting of two phases: modeling phase and evaluation phase. In this Chapter, we first introduce the definitions and notations used in our method. Then, we portray our design concept in detail. The proposed method can realize the role of tie strength in the information diffusion.

3.1 Definitions & Notations

This thesis adopts the definition of 'stream' and 'ticker' given by Facebook Inc. [24] and a common sociology noun, 'dyad', to explain the sociology interactions of Facebook users.

- ***Stream***: so called ***news feed***, which represents the way to keep up with people in our lives. 'Stream' exhibits status updates, links, photos and app activity from our friends and group.
- ***Ticker***: a faster vision of news feed such a live stream of activity appears in real-time.
- ***Dyad***: a pair of individuals which is the smallest social group connects with each other via relationship . 'Dyad' is a common noun used to describe the type of interaction in sociology.

For each story on stream, users can have four actions: like, comment, click and share.

- ***Like***: a way to give positive feedback and connect with the stories users care about. Users click the "Like" button at the bottom of the content. This makes the content appear in their friends' News feeds.

- ***Comment***: stream allows users to post messages for their friends to read. In turn, friends can respond with their own comments.
- ***Share***: a simple way to share the information with users' friends on Facebook. Users click the "Share" button at the bottom of the content. This makes the content appear in users' own wall.
- ***Click***: a particular action for cyber links. Users open the cyber links in a new tab to view the content of links by clicking.

According to different interactions, we further classify online users into three types, including contagious users, inactive users and active users.

- ***Contagious users***: users who post information to their own walls on Facebook.
- ***Inactive users***: users who either are exposed to the information but ignore the information, or are not exposed to the information posted by a contagious user.
- ***Active users***: users who are exposed to the information and get the information posted by a contagious user. An active user may not become a contagious user. Once a user becomes active, the user will never be inactive for a specific post.

Except contagious users in a set of users, a user is either an inactive user or an active user for a specific wall post. To clearly explain the proposed method for evaluating the information diffusion, we define the following notations to represent the user sets of different types.

- U: a set of users.
- F_c: a set of user c's friends.
- $\mathcal{A}$: a set of active users, where $\mathcal{A} \subset U$.
- $\mathcal{I}$: a set of inactive users, where $\mathcal{I} \subseteq U$.
- $\mathcal{C}$: a set of contagious users, where $\mathcal{C} \subset U$, $\mathcal{A} \cap \mathcal{I} = \phi$, $\mathcal{A} \cap \mathcal{C} = \phi$, $\mathcal{C} \cap \mathcal{I} = \phi$ and $\mathcal{A} \cup \mathcal{I} \cup \mathcal{C} = U$.

Taking Figure 3.1 as an example, the whole block represents the whole user set U, the white part represents an inactive user set $\mathcal{I}$, the gray part represents an active user set $\mathcal{A}$ and the black part represents a contagious user set $\mathcal{C}$. For a contagious user c, its friend set F_c is depicted. As the result of influence cause by user c, F_c can be divided into different subsets, $\mathcal{A}, \mathcal{I}$ and $\mathcal{C}$, such that $\mathcal{A} \cup F_c$, $\mathcal{I} \cup F_c$ and $\mathcal{C} \cup F_c$.

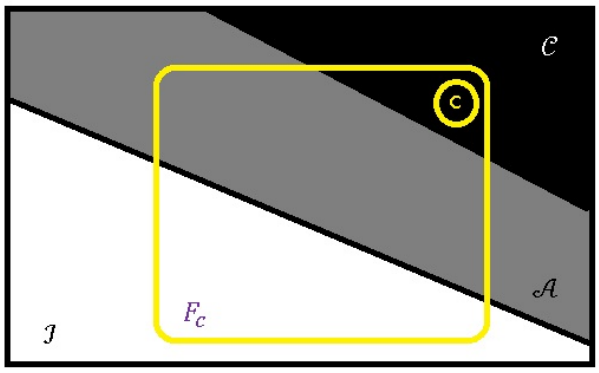


Figure 3.1: The graph representation of user set.

3.2 Design Concept

By observation, we figure out that users have their own behavioural posting pattern and responding pattern. Most people have daily routine and routine cycles by week. Then, users get online in a regular way. The time at which a post is created by a user varies little. Refresh frequency of News Feed on the user's wall alters little as well. Take a dyad, user i and user j, as an example, if user j's wall refreshes frequently, the factor of the time a post created by user i influences user j's responding pattern more. If user j's wall refreshes scarcely, the factor of the time posted by user i influences user j's responding pattern rarely.

Users have their own interests and share the similar information. McPherson, M. *et al.* indicated [21] "homophily" in a social network. A users' posts appeal to friends of the user by

topics. For examples, a user who shares the travel information appeals to the user who likes to travel. The user who is interested in finance watches out for the user who shares the business essays. And they may have more interactions by either information exchanges or unidirectional responses.

Additionally, we consider that while the Internet grows rapidly and widely in recent years, users aware of online security issues gradually. Users decide to whether click a cyperlink or not more carefully than before. Although a user get interested in the title of a cyperlink, the user may think twice before clicking this information right away.

Therefore, the impact factors of habitual pattern contain the posting timing, the refresh frequency of News Feed, the attractiveness of the topics and the security awareness of cyperlinks, as illustrated in Figure 3.2.

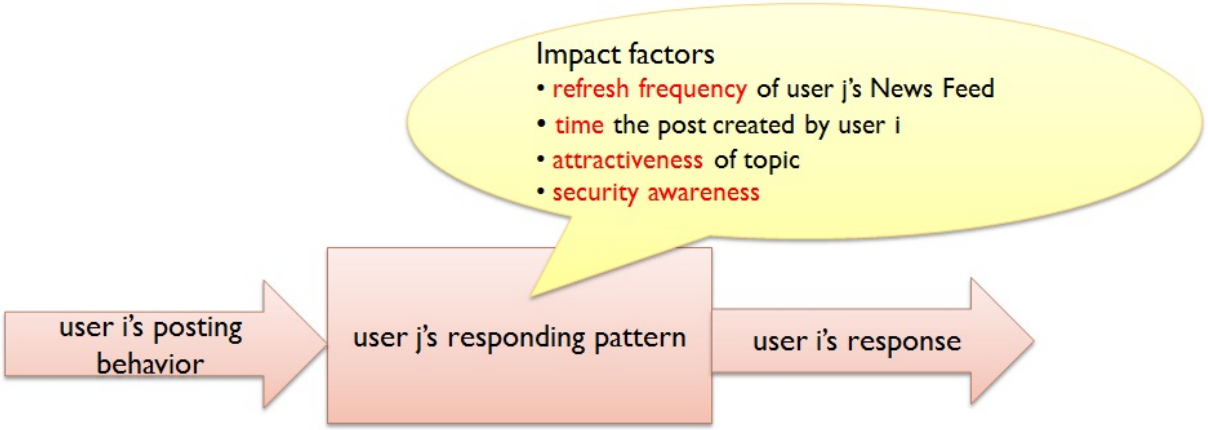


Figure 3.2: The concept of our design.

3.3 Phases

In the proposed method, we design two phases, modeling and evaluation, to model the tie strength and estimate the influence of information. In the first phase, we first model the tie strength to indicate the potential influence of dyad. Then, we model the individual responding

rate for the opportunity of a user triggered by another user. Secondly, we use the dyad's responding rate estimated in first phase to predict the information delivery path, along wtih which the information on different individual diffuses in the social network. In the second phase, we evaluate the information diffusion in a social network with the tie strength and the responding rate.

3.3.1 Modeling

We assume that the nature and frequency of online interactions between two users directly affect their tie strength. The individual responding rate implies the opportunity that one user gets the information posted by another user. The larger the individual responding rate is, the higher likelihood interaction will occur. The responding rate implies that a specific piece of information (e.g., either malicious links or rumours) has been delivered successfully between a pair of users. The stronger the tie strength is, the stronger virtual relationship between the dyad exists. We illustrate the individual responding rate and the tie strength using directed graphica representation as illustrated in Figure 3.4. For example, user g, j and k are friends of user i, i.e., $g, i, j, k \in U$ and $g, j, k \in F_i$. User i influences user j to the extent $s^{(ij)}$ as node i points toward node j. User i has an opportunity $r^{(ij)}$ to affect user j as node i points toward node j. In the same way, we describe the opposite situation with user j influencing user i to the extent $s^{(ji)}$ as node j points toward node i. And user j has an opportunity $r^{(ji)}$ to affect user i as node j pointing toward node i.

We propose a model to estimate the tie strength as a linear combination of stream updates in Eq. (3.1).

$$s^{(ij)} = \alpha_0^{(ij)} + \alpha_1^{(ij)} \mathrm{x}^{(i)} + \alpha_2^{(ij)} \mathrm{y}^{(i)} + \epsilon^{(ij)}, \tag{3.1}$$

where $\alpha_1^{(ij)}$ is the weight of $\mathrm{x}^{(i)}$, $\alpha_2^{(ij)}$ is also the weight of $\mathrm{y}^{(i)}$, and $\alpha_0^{(ij)}$ is the estimated constant

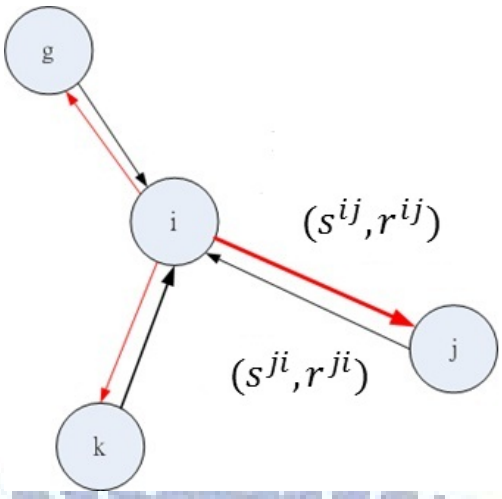


Figure 3.3: Graphical individual responding rate and tie strength representation.

and error term $\epsilon^{(ij)}$ is i.i.d. (independent and identically distributed). The individual responding rate can thus be modeled as a linear combination in Eq. (3.2).

$$r^{(ij)} = \beta_0^{(ij)} + \beta_1^{(ij)} \mathrm{x}^{(i)} + \beta_2^{(ij)} \mathrm{y}^{(i)} + \epsilon^{(ij)}, \tag{3.2}$$

where $\beta_1^{(ij)}$ is the weight for $\mathrm{x}^{(i)}$ to be estimated, $\beta_2^{(ij)}$ is also the weight for $\mathrm{y}^{(i)}$ to be estimated, and $\beta_0^{(ij)}$ is estimated constant and error term $\epsilon^{(ij)}$ is i.i.d.

Then, we define $s^{(ij)}$ as the tie strength that user i toward user j and $r^{(ij)}$ as the responding rate that user i toward user j. Let $\mathrm{y}^{(i)}$ denote a non-link stream variable by user i and $\mathrm{x}^{(i)}$ represent a link stream variable by user i.

Distinguishing from other research, we separate the individuals' wall data into two categories: link and "non-link". A non-link updated stream consists of either texts or photos. Even though all the stream data represent the importance of interaction behaviours, only the links posted in the stream have potential aggression upon other users. We consider that the interactions are directional, since two individuals do not have the same affectivity on each other (e.g., user i commented on user j's stream, but user j might not comment on user i's stream). Due to the sparsity if the data, we only consider the information posted on a wall by its owner, instead of friends of the owner.

We consider that each user has his or her own behavioural pattern in a social network. Most ousers have daily routine and the routine repeats by week as mentioned in section 3.2. We consider the timing user presence online every week is similar. Hence, we divide the real data by week, depicted in Table 5.1 and Table 5.2. $s^{(ij)}$ is the real number of user i's posts with comments received from user j. We use Least-Squares Regression to estimate the expected tie strength $\mathbf{s}^{(ij)}$ for every week.

With such a design, even though the same variables are used to calculate the responding rate and tie strength, we might still obtain different results, For example, user g has updated 10 pieces of information, and user j has commented on 5 of them. User i has updated 20 pieces of information, and again user j has commented on 9 of them. Although user j has commented more on user i's posts, we might get dissimilar responding rates. According to Eq. (3.2), we obtain $r^{(gj)} > r^{(ij)}$ instead. Hence, we conclude that user j has more opportunity to be influenced by user g. We also consider the frequency of the user j's news feed in the same way. Because if user i updated stream frequently, user j may not have a chance to see the feed and get a small $r^{(ij)}$. We calculate the numerator of $r^{(ij)}$, which is number of stream updated if user j has commented on user i's each stream updated on user i's own wall per week, and the denominator of $r^{(ij)}$, which is number of total stream updated on user i's own wall by user i per week, of the fraction as the responding rate. Also, we use Least-Squares Regression to estimate the expected responding rate $\mathbf{r}^{(ij)}$ for user i from user j. The responding rate would be in the interval [0,1].

3.3.2 Evaluation

In the previous section, we have tie strength and responding rate estimated between individuals. The estimation forms a one-hop network, as illustrated in Figure 3.4, where user *i* is

Table 3.1: Non-link stream shared for user i during week w,$w = \{1, 2, ..., m-1, m\}$.

$x_1^{(i)}$	number of i's non-link stream updated on i's own wall during week 1
$x_2^{(i)}$	number of i's non-link stream updated on i's own wall during week 2
.....	
$x_{m-1}^{(i)}$	number of i's non-link stream updated on i's own wall during week $m-1$
$x_m^{(i)}$	number of i's non-link stream updated on i's own wall during week m

Table 3.2: Link stream for user i during week w,$w = \{1, 2, ..., m-1, m\}$.

$y_1^{(i)}$	number of i's link stream shared on i's own wall during week 1
$y_2^{(i)}$	number of i's link stream shared on i's own wall during week 2
.....	
$y_{m-1}^{(i)}$	number of i's link stream shared on i's own wall during week $m-1$
$y_m^{(i)}$	number of i's link stream shared on i's own wall during week m

the central node and has latent relationships with his or her friends. Even though all users are in user c's friend list, some of them might not have connections (absent ties). The links between each dyad can be also viewed as zero if no friend responds to any stream posted by user c. Taking Figure 3.4 as an example, one node represents one user. User $g, h, k, j, p, q, r, l, m, n$ are user c's friends. User g, h, k, j, p, q, r have interactions based with user c as node c points toward them. User l, m, n have no interaction with user c. In this example, user c is a contagious user, user g, h, k, j, p, q, r are active and user l, m, n are inactive. The graph shows $g, h, k, j, r, p, q \in \{A \cap F_c\}$ and $l, m, n \in \{I \cap F_c\}$. By overlapping the one-hop networks, we can construct a connected component to represent the whole network as shown in Figure 3.5

By adapting the *Independent cascade model*, we design a model to estimate the diffusion of information. In the simplest independent cascade model [25], when node v becomes active, user v attempts to activate user v's neighbor w, succeeding with probability $p_{v,w}$. If a malicious link or a rumour lets an affected user publish this information automatically, the active user

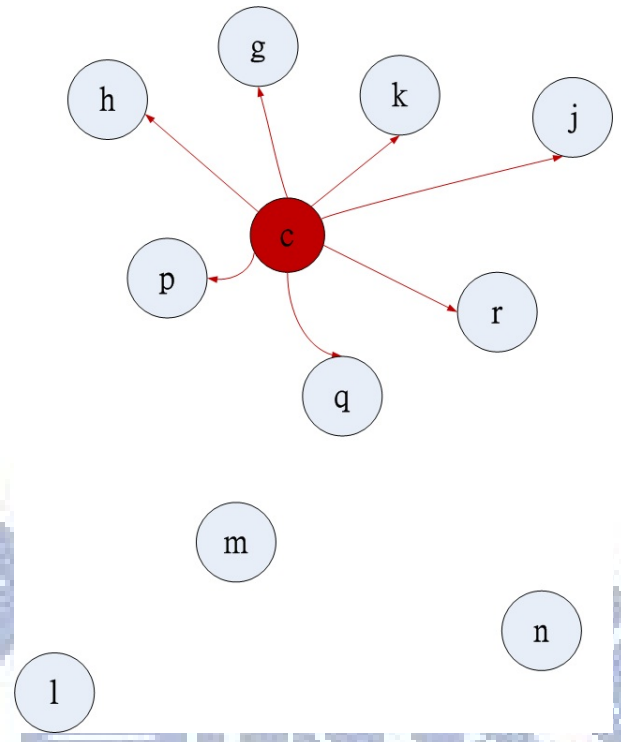


Figure 3.4: Tie strength forms one-hop network individually.

becomes a contagious user as well. The 1-hop information delivering model can then be derived as Eq. 3.3

$$\sigma^{(1)}(c, th) = \sum_{v \in F_c} s^{(cv)} \quad \text{if} \quad r^{(cv)} \geq th, \tag{3.3}$$

where F_c is a set of user c's friend v, user c is a contagious user, user v is inactive user and th is responding threshold.

The responding threshold is a threshold that activates a user $(v \in F_c)$ to become active. The activation succeeds when the responding rate $r^{(cv)}$ is higher than responding threshold th. The responding threshold is affected by the attractiveness of a post topics and the timing the post created. $\sigma^{(1)}(c, th)$ predicts the possible delivering path of 1-hop and the potentially affected users, when the contagious user c posts information with responding threshold th. When the responding threshold is equal to or larger than the responding rate $r^{(cv)}$, the infectious information would activate the user v to become an active user. This means that the opportunity of satisfying the responding threshold rises as user has larger responding rate.

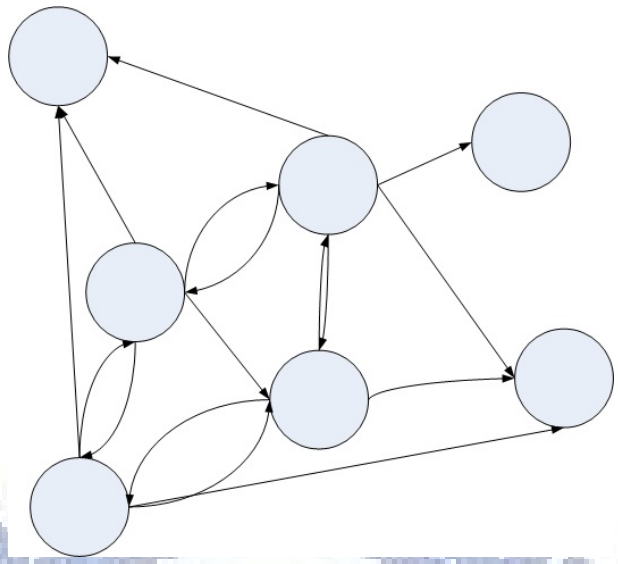


Figure 3.5: A connected component in a graph is a set of nodes where each node has path using both forward and reverse links to every other node in the set.

Consequently, if the affected user automatically publishes some information, the user would immediately become a contagious user. In such a scenario, we define the m-hop $(m > 1)$ information delivering model in Eq. 3.4.

$$\sigma^{(m)}(c, th) = \sum_{p \in F_c} s^{(cp)} + \sum_{p \in F_c} \sum_{v \in F_p} s^{(pv)} \tag{3.4}$$

$$\text{if } r^{(cp)} \geq th \ \ \& \ \ r^{(pv)} \geq th \ \ \& \ \ (r^{(pv)} > r^{(qv)} \| (r^{(pv)} = r^{(qv)} \& \& r^{(cp)} > r^{(cq)})),$$

where $p, q \in F_c$, $v \in F_p$, user c is a contagious user, user p,q are affected by user c and user v is the potentially affected user affected by user p. $\sigma^{(m)}(c, th)$ represents the m-hop influence of a piece of information shared by user c, where the responding threshold is th. In this scenario, a user has a chance to be affected by more than one users. Besides achieving the responding threshold $r^{(cp)} \geq th$ and $r^{(pv)} \geq th$, we consider two situations: (1) the user p with larger responding rate $r^{(pv)}$ would deliver the information to user v; (2) user p and q have the same responding rate with user v, $r^{(pv)} = r^{(qv)}$ where user p with larger responding rate $r^{(cp)}$ would deliver the information to user v. And we illustrate two situations in Figure 3.6, where for the affected users p and q belong to different friend sets (e.i., $q \in F_p$ and $q \notin F_p$). In brief, $\sigma^{(1)}$

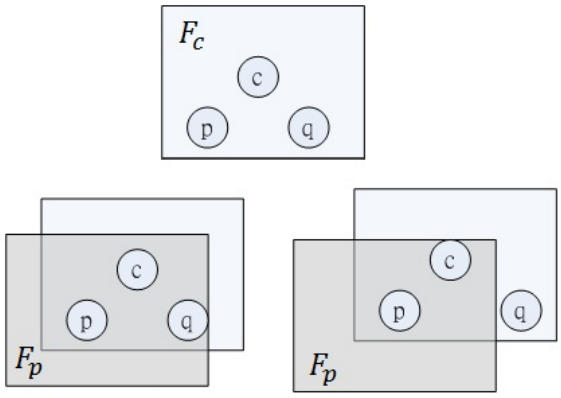


Figure 3.6: Two situations for the friend set of user p

portrays 1-hop information delivery and $\sigma^{(m)}$ pictures m-hop information delivery.

Moreover, we illustrate two scenarios in Figure 3.7 and Figure 3.8 for m-hop information delivering path prediction. Each node represents one user, while the edge represents the directional response form one user to another. The number labelled beside the arrows are the estimation of the responding rate of two users.

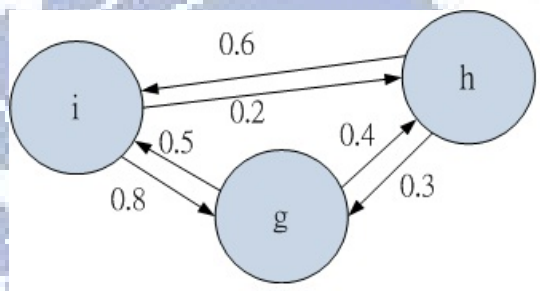


Figure 3.7: Information Diffusion Representation for the first case

- Scenario I:

  In the first scenario, user g is a contagious user carrying a piece of information with responding threshold 0.5 to influence friends of user g. The information would be propagated along the directed paths toward users who are triggered with $r^{(ij)} > 0.5$ such as user i but not user h. Due to $r^{(hi)} \geq 0.5$, $r^{(ig)} \geq 0.5$ and $r^{(hg)} < 0.5$, when user h is a

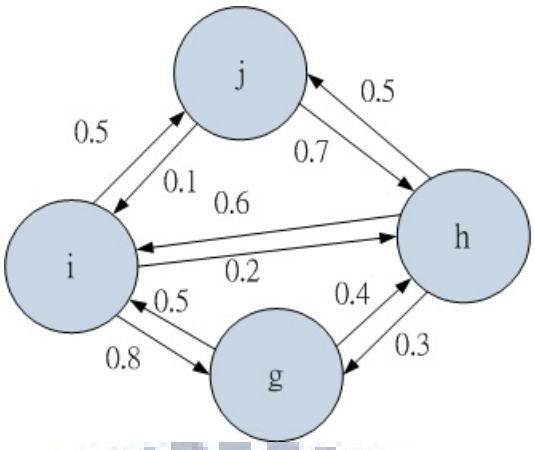


Figure 3.8: Information Diffusion Representation for the second case

contagious user, the information flows through user i toward user g, instead of directly influence user g. As a result, we illustrate information delivering path for the first scenario in Figure 3.9.

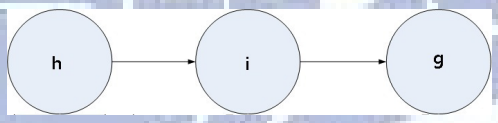


Figure 3.9: The predicted m-hop information delivering path for the first scenario.

- Scenario II:

  For a smaller responding threshold (0.4), user h and user i are influenced in the first hop. In the second hop, for both user h's and user i's turns to disseminate information, user j, a mutual friend of user h and user i, approaches both the responding threshold from both user h and user i. The credit for who to trigger user j depends on who has larger $r^{(ij)}$ from the previous hop. Therefore, the information would be delivered along the path form user i to user j, instead of along the path form user h to user j. Also, we illustrate information delivering path for the second scenario in Figure 3.10. It would be the same way to predict the information delivering path with different individual as the first information sharing user. The different responding threshold would change the information delivery range.

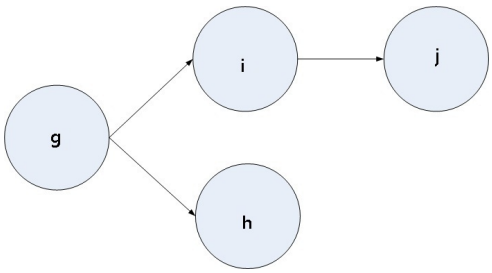


Figure 3.10: The predicted m-hop information delivering path for the second scenario.

Furthermore, we evaluate each user's influence of diffusion by both tie strength and responding rate. Information diffusion can be classified into two types: 1-hop diffusion and m-hop diffusion. Normally, we select some users as a target set. Then, we adjust the size of the target set and observe the impacts of the users in target set on the whole network . For 1-hop diffusion, we define that the extent of user v's friends tricked by user c's post. Also, user c is initially targeted by our attack. We propose a model to value 1-hop diffusion for any responding threshold th, $th \in [0, 1]$, in Eq. (3.5)

$$\sigma^{(1)}(\mathcal{C}) = \sum_{c \in \mathcal{C}} \sigma^{(1)}(c) \tag{3.5}$$

For m-hop diffusion, we define that extent of users tricked by user c's post. Also, user c's post spreads with m-hop diffusion when user c is initially targeted in our target set. We propose a model to evaluate m-hop diffusion for any responding threshold th, $th \in [0, 1]$, in Eq. (3.6).

$$\sigma^{(m)}(\mathcal{C}) = \sigma^{(1)}(\mathcal{C}) + \sum_{v \in F_p, p \in F_C} (r^{(cp)} \times s^{(pv)}) \tag{3.6}$$

Since the responding rates may affect the exposing opportunity of a particular post, in addition to tie strengths, we also consider the responding rates when estimating the diffusion of the post. Therefore, in the above equation, $r^{(cp)}$ is defined to represent the opportunity that user p (influenced by user c) makes a impact on user v. Different from Eq. 3.4, no particular responding threshold th is specified, but all possible delivering paths for a particular piece of information are considered in Eq. 3.6. In short, let user c be the initial attacking target and $\mathcal{C}$ be the initial

contagious set. By Eq. 3.6, we can obtain the impact brought by $\mathcal{C}$. Therefore, we consider that the higher $\sigma(C)$ a user has, the higher information diffusion range through this connected component network users C have potential to reach.

Concisely, with the proposed method, if we get the access tokens for one of the dyad, then we can estimate the each dyad's responding rate and tie strength in the first step. The responding rate helps us to understand the individual's influential generation with his or her friends by habitual sharings. Also, strength of a tie aims us to comprehend the individual's influential intensity with his or her friends. In second step, we predict the information delivering path, according to the responding rate obtained in the first step. In final step, we evaluate the diffusion not only by tie strength but also by the responding rate. Once we have permissions of some users to fetch the essential data, these users form a connected component. For a connected component, the above steps can be used to assess network information diffusion, such as malicious attacks and rumours.

3.4 Examination

In general diffusion model, the signal decays while it disseminates due to the power consumption, the environment disturbance, etc. However, we evaluate the multi-layer effects on a piece of information diffusion by summing up all the responding rates multiplying tie strengths among the dyads, instead of by series multiplying the responding rate and the tie strengths. The way of information spreading is distinct from general signal transmitting. For the power consumption, there is different concept of infromation spreading on Facebook. For example, the particular information posted by user c is shared by user v. To the friends of user v, the particular information is first time shared by user v. User v's friends consider that user v is the original poster of the particular information. Therefore, we claim that there is no signal

decaying when the particular information diffuses.

Since definition and selection of variables are considerable in designing a tie strength model, we examine the proposed method in terms of different considerations, including homophily, the user groups and the inequality of information exposure.

"Homophily" implies that users with high similarities gather together. Taking students studying at the same school as an example, students may know each other due to the same 'education' in their user profiles. Hence, we say these students are homophilous. In the existing work [17] [20], 'user profile' or 'background' is adopted as an important variable on modeling tie strengths among users. However, students studying in the same school, in fact, only know a small portion of students in the school. And, users who are interested in the same topics may naturally talk and interact with each other. So, we consider the 'user profile' or 'background' a redundant variable. In the thesis, we measure tie strength between two users with the dyadic interactions, which somehow implies their similarities.

Different users create friend lists in a diverse way. The average Facebook use is connected to approximately 80 commumity pages, groups and events. Considering users exposed to different custom social group, we record whether the interaction exists between the dyad. The record helps to understand the extent that groups users belong to overlapping with one another.

Moreover, to avoid that a user is flooded with massive activities on News Feeds, Facebook is using the **EdgeRank** algorithm that determines what stories populate users' News Feeds. The EdgeRank value is assigned to every story. This value is based on *affinity*, *weight* and *time*. Affinity is the score between viewing user and edge creator. Weight is affected by the type of a story such as post, comment, like, tag, etc. These actions have different weight: **Share** > **Comments**> **Like**> **Click**. However, it's not accessible to retrieve the real data of the sharers of each links via FQL now. In the future, we planned to consider the sharer to complete our

method since we believe that there is a correlation between dyad's interactions and the spread of information via social network. Thus, we adopt `comments` for the proposed method. Time is a freshness factor. The more recently users post, the higher EdgeRank scores. EdgeRank is the reason that users don't see every post from their news feed. However, the more users interact with each other, the greater the affinity score becomes. Thy are more likely to see users' posts in the future. Hence, we not only count up the amount but also compute the proportion of comments received.

Chapter 4

Experiments

Facebook is a popular online social network site with 845 million monthly active users at the end of December 2011 [23]. Users can not only create and update their personal profiles but also upload their photos and tag others in pictures. Actions a user makes can show either on their own friends' wall or on public, which means it can be seen and shared by strangers. The users' friendships are undirected because a user only invites other users whom they want to make friend with and get invitee's confirmation. Our experiments evaluate the proposed method on Facebook data (www.facebook.com).

4.1 Preliminary

To get the real data from Facebook, we register a Facebook Developer account and create a facebook application, called "Hare" (deep-robot-3593.herokuapp.com), on Heroku. Heroku is a cloud application platform and cooperates with Facebook platform. We install Herokutoolbelt, which contains a CLI tool. On the Heruku platform, we can access PHP SDK and Javascript SDK directly. We deploy our Hare using the Git (version 1.7.6) revision control system. Our Hare establishes in October 2011 and uses OAuth 2.0 protocol, which is announced in May 2011. For querying facebook database, we use Facebook Query Language (FQL) to get the real data for modeling tie strengths and responding rates. We analyze the data queried from facebook database with Matlab R2010a. The proposed method is simulated by Matlab R2010a as well.

The Figure 4.1 shows how Facebook applications interact with Facebook Graph API. We request Hare on Heroku. Heroku would send FQL as a API call to Facebook. Then, Facebook responds in JSON format. Hare builds an HTML response to show the desired data.

Figure 4.1: How Facebook applications interact with Graph API.

4.2 Dataset

Because the permissions of some media data have privacy issues, we have little chance to conduct large-scale experiments. In the beginning, we invite 36 users to participant our experiments. But only 24 Facebook users allow our Hare to access their data on Facebook. Most of these participants are students from National Chiao Tung University. 5 out of 24 users are non-students. These 24 participants with their friends form 9550 friendships.

We collect the media data from September 5th, 2011 to February 19th, 2012. And we separate the data into two periods: the first duration T_1 started from September 5th to November 27th in 2011 and the second duration T_2 started from November 28th in 2011 to February 19th in 2012. Data in first time set is used to estimate the proposed method. Data in second time set is used for verifying the correctness of our method.

For the proposed method, we need "read_stream" permissions to access table stream for all posts in the users' **News Feed**. Also, we need "read_friendlists" permissions to read table friend for any friend lists the user created [24]. Even though the Facebook developer site shows that an extended permission is no need for table friend, Facebook only provides to retrieve the friends of current session user by default.

In the following paragraphs, we list the entries of table stream, comment and friend defined by Facebook, Inc. [24]. In the table stream, five major columns are defined to maintain the wall post of a user.

```
/* Table stream */
    post_id       // the ID of the post.

    source_id     // the ID of user, page, group,
                  // or event whose wall the post is on.

    filter_key    // the filter key to fetch data with.

    type          // the type of the story,
                  // such as 80 for link and 46 for status update.

    created_time  // the time the post was published.
```

In the table comment, two major columns are defined to maintain the comments received in a post.

```
/* Table comment */
    post_id  // the ID of the post.
    fromid   // the ID of user who submits a comment.
```

In the table friend, two major columns are defined to maintain the list of a user's friend.

```
/* Table friend */
    uid1  // the user ID of the first user in a particular friendship link.
    uid2  // the user ID of the second user in a particular friendship link.
```

To avoid confusion and respect to Facebook copyright, we do not reword the definitions of each entries. We only list parts of the entries which are used in the proposed method.

4.3 Validation

Friendship varies with time in real world, not to mention in the virtual social network. By our assumptions, users have daily routine and also have influence toward their friends. We verify this idea with the real data we collected to make our method more persuasive.

4.3.1 Tie Strength

Using the first time set, the proposed method evaluates the participants with tie strengths. For the purpose of the validation, we consider the average of the every estimated tie strength $s^{(ij)}$ per week as the tie strength for every i toward j in a particular duration. We list the order of users have most potential influence intensity toward their friends as shown in Table 4.1. Then, we verify our target set with the second time set to check whether these users have more interactions with their friends. Concerning the user privacy, we replace their facebook identities with their nicknames.

$$fluctuation\ ratio = \frac{time\ set\ T_2 - time\ set\ T_1}{time\ set\ T_1} \tag{4.1}$$

Comparing with two time sets, there are 3 of 5 matched within two target sets. We find out that even though average fluctuation ratio of individual user is 0.21, the fluctuation ratio of whole target set is 0.0102 only. In our dataset, notwithstanding number of friends, number of friends

Table 4.1: Order of users have the most potential influence toward their friends by tie strength when tie strength fluctuates with time varying. (T_1:[Sept.-Nov.] T_2:[Dec.-Feb.])

user	**integrated tie strength**		**fluctuation ratio** (4.1)
	$\mathbf{T_1}$ (order)	$\mathbf{T_2}$ (order)	
alphar	*29.424 (1)*	*28.196 (1)*	*↓0.0417*
benben	*23.051 (2)*	*14.941 (5)*	*↓0.3518*
doo	*17.768 (3)*	*21.538 (2)*	*↑0.2122*
mei	*17.107 (4)*	*12.774 (9)*	*↓0.2533*
ann	*16.194 (5)*	*14.449 (7)*	*↓0.1077*
viola	*14.476 (6)*	*20.403 (3)*	*↑0.4094*
mao	*14.015 (7)*	*11.827 (13)*	*↓0.1561*
stone	*13.039 (8)*	*14.651 (6)*	*↑0.1236*
claire	*11.713 (9)*	*15.057 (4)*	*↑0.2855*
channing	*10.981 (10)*	*8.977 (17)*	*↓0.1824*
average	*11.122*	*11.236*	*↑0.0102*

who interact with our participants is around 100 users. According to the estimation results, 0.21 average fluctuation ratio of individual user means about one user variation of the potentially affected users. The variation is in a sensible and prospected range. We figure out that if we get the information about participants' career and re-list the target set as two groups. One is students and the other is non-students, showed in Table 4.2 and in Table 4.3.

In Table 4.2 and Table 4.3, we notice that match percentage increases both for students and non-students. And the non-students having daily routine more approximates to our assumption than students. Therefore, we can evaluate the target sets totally matched between two periods. These phenomenon implies the career have slight impact on our assumptions.

4.3.2 Responding Rate

Subsequently, using the first time set, the proposed method evaluates the participants with responding rates. For the purpose of the validation, we consider the average of the every esti-

Table 4.2: Order of students have most potential influence toward their friends by tie strength with time varying.

order	**user**	**integrated tie strength** T_1	**user**	**integrated tie strength** T_2
1	*benben*	*23.0501*	*doo*	*21.5382*
2	*doo*	*17.7676*	*viola*	*20.4031*
3	*viola*	*14.4763*	*claire*	*15.0568*
4	*mao*	*14.0148*	*benben*	*14.9411*
5	*stone*	*13.0389*	*stone*	*14.6504*
6	*claire*	*11.7128*	*minhsi*	*13.4869*
7	*channing*	*10.9798*	*flower*	*12.3364*
8	*ryan*	*10.8396*	*terry*	*12.0245*
9	*minhsi*	*10.3403*	*mao*	*11.8268*
10	*jason*	*9.1983*	*jason*	*10.2462*

Table 4.3: Order of non-students have most potential influence toward their friends by tie strength with time varying.

order	**user**	**integrated tie strength** T_1	**user**	**integrated tie strength** T_2
1	*alphar*	*29.424*	*alphar*	*28.196*
2	*mei*	*17.107*	*ann*	*14.449*
3	*ann*	*16.194*	*mei*	*12.774*
4	*lun*	*6.954*	*lun*	*12.208*
5	*gk*	*6.455*	*gk*	*5.411*

mated responding rate $\mathbf{r}^{(ij)}$ per week as the responding rate for every i toward j in a particular duration. We list the order of users have most potential influence toward their friends as shown in Table 4.4. Then, we verify our target set with the second time set to check whether these users have higher responding rate with their friends. Because of the privacy issue, we replace their Facebook identities with their nicknames.

Comparing with two time sets, there are 4 of 5 matched within two target sets. We find out that even though average fluctuation ratio of individual user is 0.29, the fluctuation ratio of whole target set is 0.0084 only. Except $minhsi$ who has the most fluctuation ratio of responding rate, the average fluctuation ratio of individual user is 0.09. As the estimation results, 0.09

Table 4.4: Order of users have the most potential influence toward their friends by responding rate when responding rate fluctuates with time varying. (T_1:[Sept.-Nov.] T_2:[Dec.-Feb.])

user	**integrated responding rate**		**fluctuation ratio**(4.1)
	$\mathbf{T_1}$ (order)	$\mathbf{T_2}$ (order)	
alphar	*7.798 (1)*	*6.682 (4)*	*↓0.1431*
max	*6.357 (2)*	*7.133 (3)*	*↑0.1221*
flower	*5.977 (3)*	*5.89 (5)*	*↓0.0146*
doo	*5.667 (4)*	*7.368 (2)*	*↑0.3001*
bzero	*5.6 (5)*	*2.0 (21)*	*↓0.6429*
benben	*4.736 (6)*	*3.114 (11)*	*↓0.3425*
mei	*4.694 (7)*	*3.642 (10)*	*↓0.2241*
stone	*4.183 (8)*	*4.709 (7)*	*↑0.1257*
minhsi	*4.067 (9)*	*8.024 (1)*	*↑0.9729*
claire	*3.765 (10)*	*3.836 (9)*	*↑0.0189*
average	*5.284*	*5.239*	*↓0.0084*

average fluctuation ratio of individual user means about one user variation of the potentially affected users. The variation is in a sensible and prospected range. According to our estimated responding rate for potential targets, we only need two variables about the users and we can list the target set for attacker with 80% matched while time varies. We figure out that if we get the information about participants' career and re-list the target set as two group. One is students and the other is non-students, showed in Table 4.5 and in Table 4.6. In Table 4.5 and Table 4.6, we notice that most students have an increasing fluctuation ratio since students are in their winter vacation during T_2. And the non-students having daily routine more approximates to our assumption than students. Therefore, we can evaluate the target set totally matched between two periods. More tie strength and responding rate estimations of participants are in appendix.

Table 4.5: Order of students have most potential influence toward their friends by responding rate with time varying.

order	user	integrated responding rate T_1	user	integrated responding rate T_2
1	*max*	*6.357*	*minhsi*	*8.024*
2	*flower*	*5.977*	*doo*	*7.368*
3	*doo*	*5.667*	*max*	*7.133*
4	*bzero*	*5.6*	*flower*	*5.89*
5	*benben*	*4.736*	*terry*	*5.794*
6	*stone*	*4.183*	*stone*	*4.709*
7	*minhsi*	*4.067*	*hsuan*	*4.517*
8	*claire*	*3.765*	*claire*	*3.836*
9	*min*	*3.333*	*benben*	*3.114*
10	*ryan*	*3.249*	*mao*	*2.96*

Table 4.6: Order of non-students have most potential influence toward their friends by responding rate with time varying.

order	user	integrated responding rate T_1	user	integrated responding rate T_2
1	*alphar*	*7.7979*	*alphar*	*6.682*
2	*mei*	*4.6937*	*mei*	*3.6417*
3	*ann*	*2.9623*	*ann*	*2.7641*
4	*lun*	*2.0861*	*lun*	*2.6083*
5	*gk*	*1.7578*	*gk*	*2.4735*

4.4 Analysis

Verifying the validation of our design concept, we confirm that users have particular and distinct influence toward their friends on Facebook in a short duration. Once our method permits valid, our experiments continue examining the phenomena which the tie strength and responding rate causing on Facebook. In this section, we employ the responding rate to infer the possible information delivering path on Facebook. According to the validation of proposed method and the information delivering path result, we evaluate the information diffusion through Face-

book by our m-hop diffusion model.

4.4.1 Information Delivering Path

We assume that attackers can assess the desired data, and the phishing link disguises interesting website successfully. Or, people feel free to assess the desired data, and spread the rumours intentionally. We trace the real data whether the information disseminates as we anticipate. While the information originates from different users, several phenomena deserve people's attention.

In this thesis, we demonstrate four cases of our method. Then, we utilize the real data to testify our method. Besides, we define that information delivering path as the average of path starting from contagious user to every active user. We record the information delivering paths to comprehend the effect of different users on information dissemination. For instance, we illustrate every participant as a node and the edge between two user as 'friends' in Figure 4.2. The edges between users represent the real connections on Facebook. Since the number of the participants and their friends are too many to display, we only illustrate the participants in this thesis. The properties can be observed in the same way for all the participants and their friends.

For case study, we define $\mathcal{D}$ as the average length of information delivering path (IDP).

$$\mathcal{D}(c) = \frac{\sum_{v \in affected\ users} IDP_{(c \to v)}}{number\ of\ affected\ users},$$

where user c is contagious, user v is an affected user and $c \to v$ represents the directed information delivering path from user c to user v.

In case 1, we presume that contagious user max carries a specific piece of information with responding threshold 0.1. When $r^{(ij)} \geq th$, we consider information would be delivered from user i toward user j. Marking the contagious user as black node and the affected user as gray node, we illustrate the result of case 1 in Figure 4.3. In this case, while

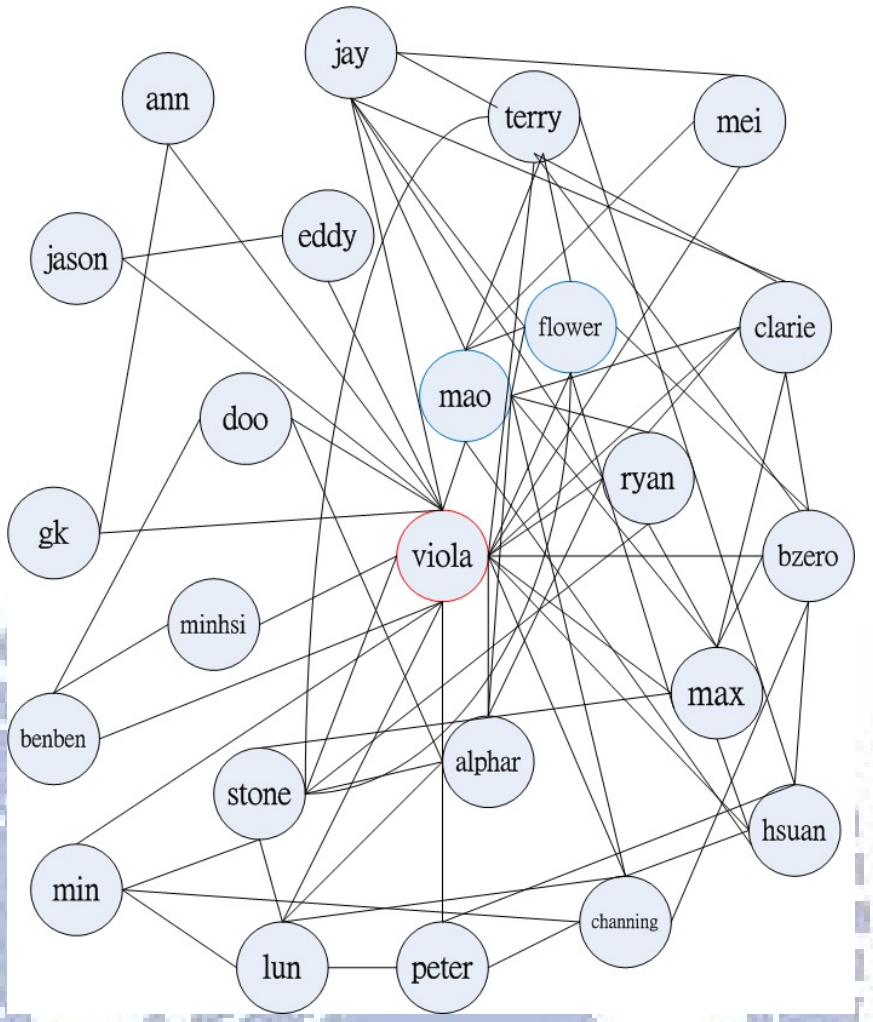


Figure 4.2: The real connections on Facebook.

$stone$, $viola$, $ryan$, mao, $bzero$, $flower$, $claire$, $terry$, mei, $jay \in F_{max}$, max spreads a specific piece of information through $viola$ to $stone$ indirectly. Real data reveal this predicted information delivering path existing, showed in Figure 4.4 and Figure 4.5. Therefore, although $stone$, $viola$, $alphar$, $peter$, $channing$, $hsuan$, $min \in F_{lun}$, lun, who is not a friend of max, has an opportunity to be affected by max via $stone$ and $viola$. Whereas not all actions are recorded in accessible database that some users get used to click the share button to share a piece of information and the others don't, we do our best to prove our predicted path set is efficient.

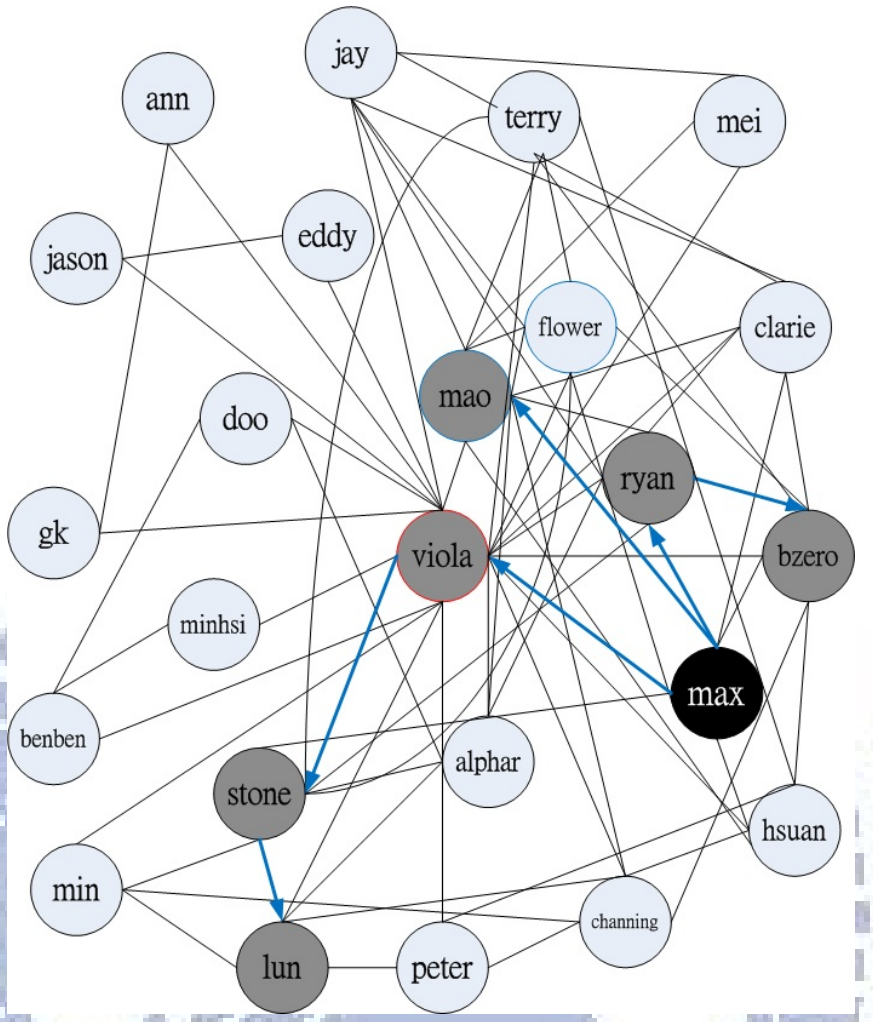


Figure 4.3: The expected information diffusion of case one.

As a result, there are 6 affected user and the information delivering path is $1\frac{2}{3}$.

$$\mathcal{D}(max) = \frac{1_{(max \to viola)} + 1_{(max \to ryan)} + 1_{(max \to mao)} + 2_{(max \to bzero)} + 2_{(max \to stone)} + 3_{(max \to lun)}}{6}$$
$$= 1\frac{2}{3}$$

In case 2, we presume that contagious user max carries a specific piece of information with responding threshold 0.05. We illustrate the result of case 2 in Figure 4.6. Besides the duplicate paths we have testified in case one, we show the rests in Figure 4.7, in Figure 4.8 and in Figure 4.9. According to the proposed method, $viola$ affect lun directly because of $r^{((viola)(lun))}$ reaching the responding threshold. Also, $bzero$ is affected by max, instead of indirectly by $ryan$. As a result, there are 14 affected user and the information delivering path is

Figure 4.4: Real data testify for the diffusion path existing from max to $ryan$.

Figure 4.5: Real data testify for the diffusion path existing from max to lun.

$1\frac{5}{12}$.

$$\mathcal{D}(max) = \frac{5 \times 1 + 8 \times 2 + 3 \times 1}{14} = 1\frac{5}{12}$$

max's influence has a chance to cause the information diffusion in breadth through our participant network.

According to case 1 and case 2, the number of affected users rises as the responding threshold descends. The number of users directed affected increases as well. By verifying with the real data, the proposed method discovers the possible delivering paths from 9550 relationships. We claim that the information certainly propagates in a selected set. On the other hand, we can derive an efficient target set for diffusion of a specific piece of information. we can derive the

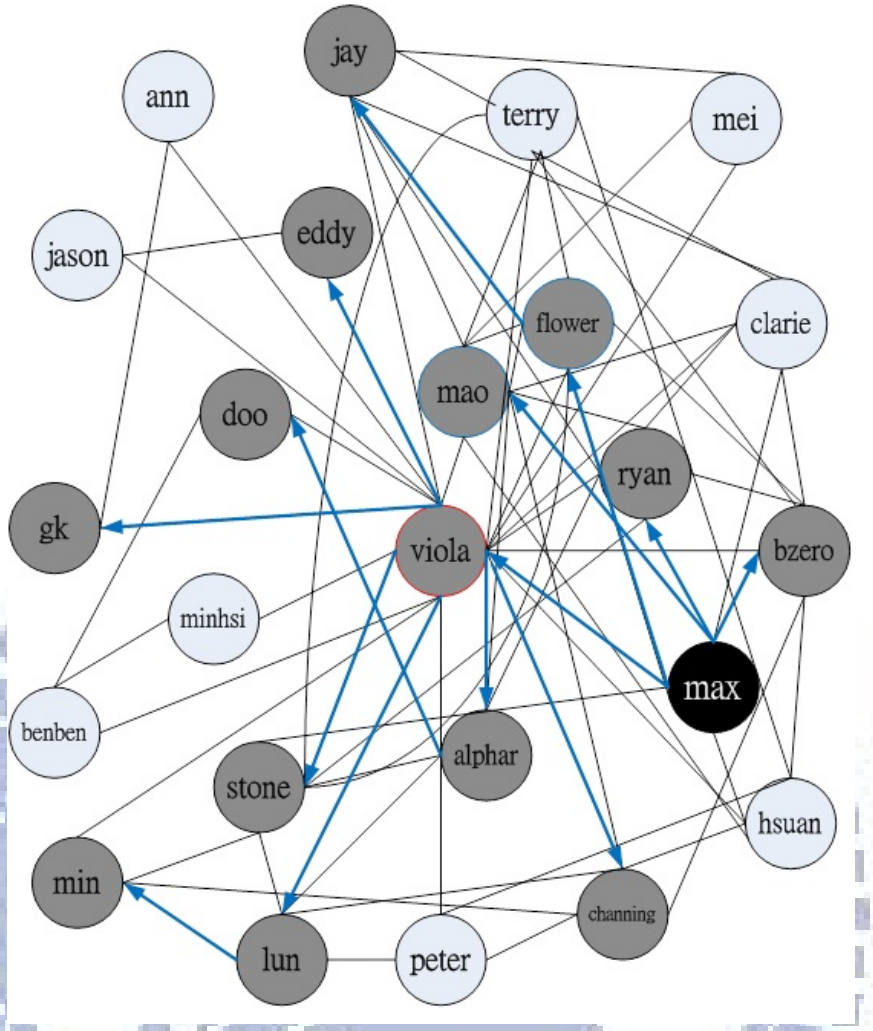


Figure 4.6: expected information diffusion of case two

diffusion coverage of a piece of information from a specified individual as well.

In case 3, we presume that the contagious user $alphar$ carries a specific piece of information with responding threshold 0.03. We illustrate the result of case 3 in Figure 4.10. Besides the duplicate paths we have testified in case 1, we show the rests in Figure 4.11. We examine a strong individual responding rate $r^{((alphar)(viola))}$, which represents the top 3 responding rate for $alphar$, to form the information delivering path steadily. Even though the content of topics are not similar, the information spreads along this path more times than others as shown in Figure 4.12. As a result, there are 6 affected users and the information delivering path is $1\frac{2}{3}$.

In the past few years, researchers have studied scale-free networks and have found the

Figure 4.7: Real data testify for the diffusion path existing from $viola$ to $alphar$.

vulnerabilities to attacks are rooted in the inhomogeneity of the connectivity distribution. In such networks, removing some highly connected nodes, which ensure the connectivity, may alter the network's topology and decrease the communication abilities of the remaining nodes dramatically [26]. Due to the testifications above, we figure out that attack survivability is not equivalent to the connectivity any more while human behaviours are included in scale-free network if the information users carry with contains malicious link. Take Figure 4.13 as an example, $viola$ is the most connected user to all other 23 participants while the information with $th = 0.1$ from $viola$ affects 2 users among the participants and the information delivering path is $1\frac{1}{2}$. Although $terry$ is not the most connected user, the information with $th = 0.1$ from $terry$ affects 2 users among the participants and the information delivering path is $3\frac{5}{9}$.

Analyzed by different angle, $stone$, who is the most connected user to his friends among the participants, has 960 friends. $alphar$ has 325 friends and $terry$ has 276 friends only. However, the information with $th = 0.1$ from $alphar$ affects 72 users among this connected component and the information with $th = 0.1$ from $terry$ affects 94 users among this connected component. But the information with $th = 0.1$ from $stone$ affects 27 users only, due to fewer interactions between $stone$ and his friends. More detail will be introduced in following subsection.

Figure 4.8: Real data testify for the diffusion path existing from $viola$ to gk.

4.4.2 Diffusion Evaluation

Due to the desired data permissions, our dataset contains 24 participants and their friends only. Regardless of the diffusion situation toward friends more than directed two hops from participants, we give a list of our participants' contagious potential through our dataset, depicted in Table 4.7. We not only evaluate the participants' influence toward this connected network but also predict who are the potentially affected users. On account of strong ties and weak ties, greater $\sigma^{(m)}(c)$ is not equal to more potentially affected users user c could affect. The greater $\sigma^{(m)}(c)$ is, the more potential influence user c has. If user c has larger $\frac{\sigma^{(m)}(c)}{number\ of\ potentially\ affected\ users}$, user c has a better opportunity to affect others successfully. However, user c with smaller $\frac{\sigma^{(m)}(c)}{number\ of\ potentially\ affected\ users}$ is not negligible. User c with smaller $\frac{\sigma^{(m)}(c)}{number\ of\ potentially\ affected\ users}$ still has an opportunity to affect others, because of two

Figure 4.9: Real data testify for the diffusion path existing from $viola$ to lun.

Table 4.7: Order by users have contagious potential through our dataset.

user	$\sigma^{(\mathbf{m})}(\mathbf{user}, 0.1)$	**maximum cumulative number of potentially affected users**	$\frac{\sigma^{(\mathbf{m})}}{\textbf{number of potential victims}}$
alphar	*26.1806*	*72*	*0.3636*
flower	*14.4183*	*88*	*0.1638*
max	*13.3293*	*64*	*0.2082*
ryan	*12.6004*	*64*	*0.1969*
hsuan	*11.2753*	*73*	*0.1545*
terry	*10.7332*	*94*	*0.1142*
bzero	*10.0351*	*64*	*0.1568*
mei	*9.6781*	*11*	*0.8799*
claire	*9.2474*	*39*	*0.2371*
benben	*8.5961*	*11*	*0.7815*
stone	*7.7616*	*27*	*0.2875*
mao	*6.7921*	*31*	*0.2191*
min	*6.0786*	*32*	*0.2101*

reasons. On one hand, according to one of Facebook functionality, once an information receives a like or a comment, a piece of information would display on the top of not only the **News Feed** pages but also **Ticker** bar. While interactions exist between user c and his or her friends factually, the information would get more exposure. On the other hand, the responding threshold depends on subjective elements of different users. In our method, responding threshold is inherited when an information is re-shared. But in the real world, the definition of responding threshold varies with different users.

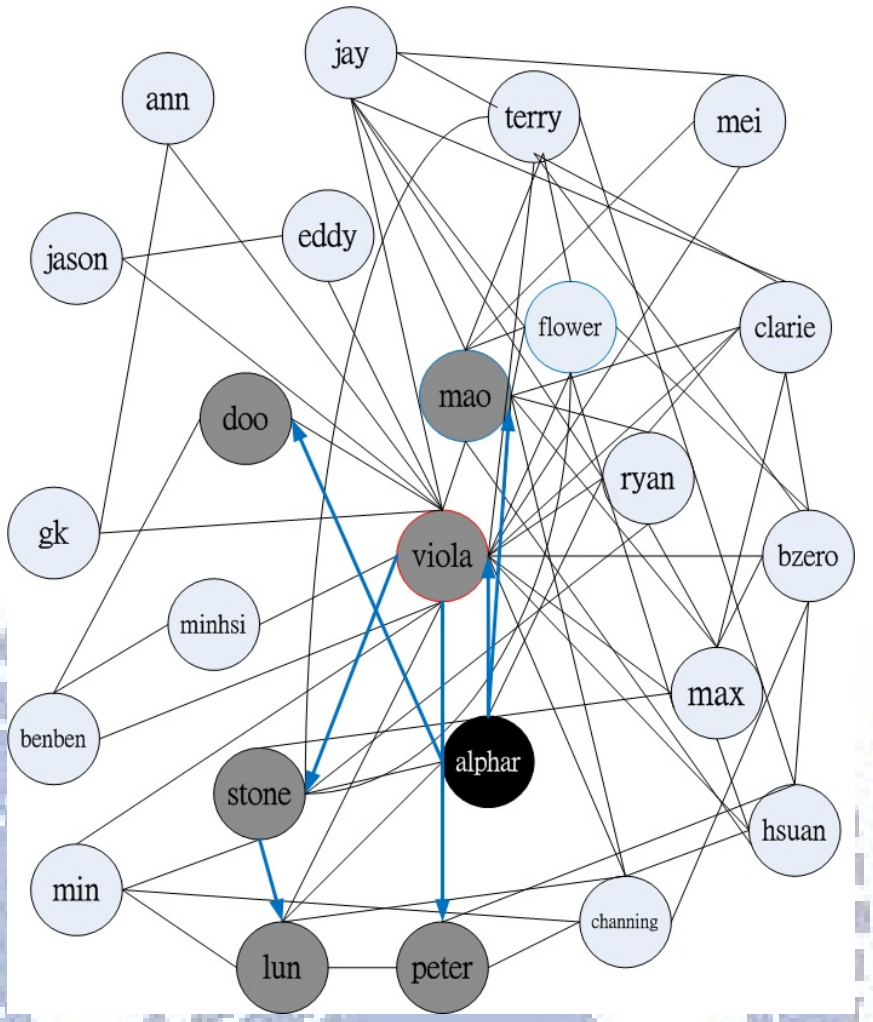


Figure 4.10: The expected information diffusion of case three.

4.5 Summary

According to our experiments, we validate our method by estimating tie strength and responding rate with real data in the first place. Secondly, we predict possible information delivering path with individual responding rate. Then, real data testify that the individual responding rate can really infer information delivering path. By the proposed method, we can select a possible information delivering path set out of all 9550 friendships. The executing reduction in selecting efficient target set is remarkable. In other words, we can predict the diffusion coverage of a piece of information from a specified individual. By giving several cases, since different users have different tie strengths with their friends and the characteristics of scale-free

Figure 4.11: Real data testify for the information delivering path existing from $alphar$ to $peter$.

network impact the information diffusion, information diffuses in a different way as contagious user alters. Thirdly, although the vulnerability of scale-free network is rooted in connectivity distribution, we observe a noticeable phenomena that the most connected users does not affect the whole network most with network structure and affinities of users. Finally, we evaluate all our participants' potential influence to indicate that who is the best choice in terms of disseminating information extensively on Facebook. Also, we provide several indexes of who is the most powerful user for information diffusion in social network: $\sigma^{(m)}(user, th)$ and maximum cumulative number of potentially affected users.

Figure 4.12: Different real data testify for the information delivering path existing steadily from *alphar* to *peter*.

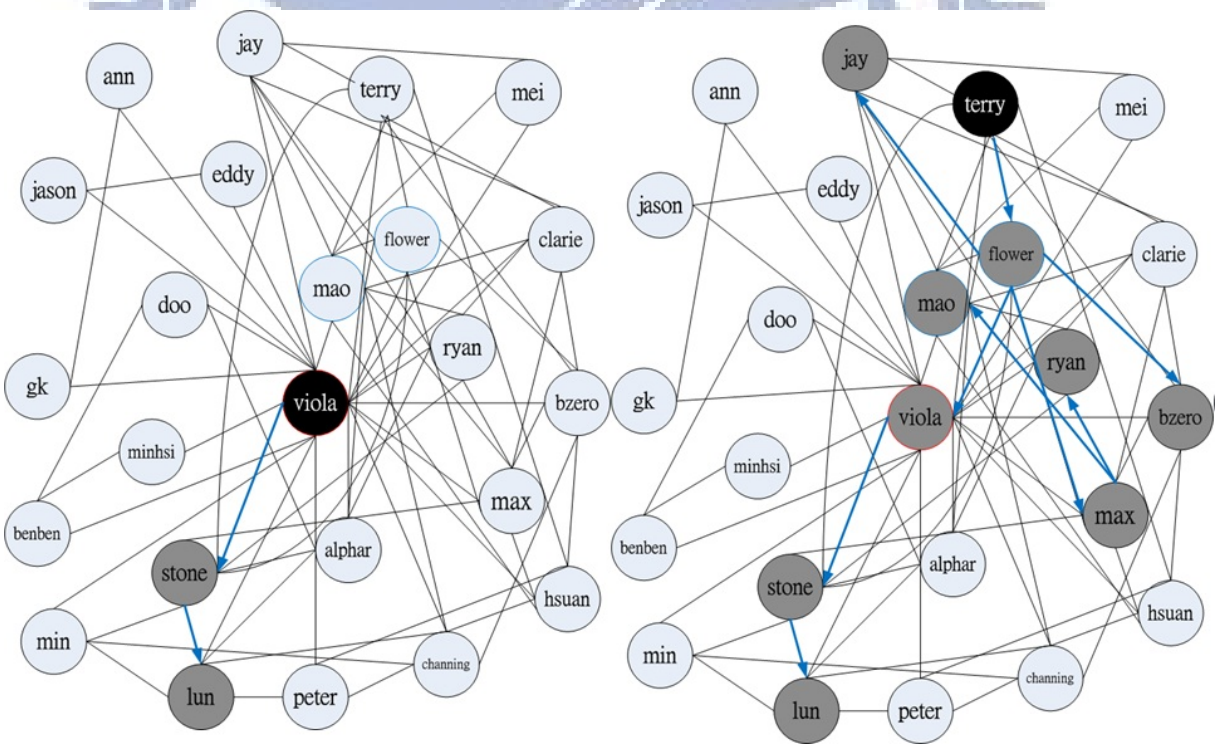


Figure 4.13: Attack Survivability $\neq$ Connectivity in Online Social Networks

Chapter 5

Discussion

In this chapter, we discuss different representations of tie strength from the existing works. Also, we compare the proposed method to existing works. Then, we examine several phenomena with the proposed method.

5.1 Estimation

In this section, we compare the proposed tie strength with the definition used in [19] in terms of similarities and differences. We observe the habitual behaviours of Facebook users to infer the cause of users re-sharing the information on Facebook. In this thesis, we learn three factors of information propagation from the experiments: (1) trust in information sharers, (2) novelty of information, (3) exposure to information.

The first factor, trust in information sharers, is the strength of ties between dyads. By the definitions of "*homophily*" and *tie strength*, a strong tie is individually influential unquestionably. Also, according to Granovetter [6], Manuel E. Sosa [7] and Bakshy *et al.* [19], strength of a weak tie is a critical bridging of diffusion in a general tie model. Because of the second factor, novelty of information, some weak ties are responsible for information dissemination. The users with weak ties have more diverse social networks that provide access to novel information. To increase the weighting of the weak ties in [19], we focus on the number of post with response from others rather than the number of response. Instead of concerning about how profoundly a user gets attracted in topics posted by other users, we intensify how diversely a

user gets interested in topics posted by other user. This represents 'intensity' and 'reciprocal services' of tie strength in another way. Then, the third factor, exposure to information, has already been proved with large-scale experiments by Facebook, Inc in Figure 5.1. The sharing

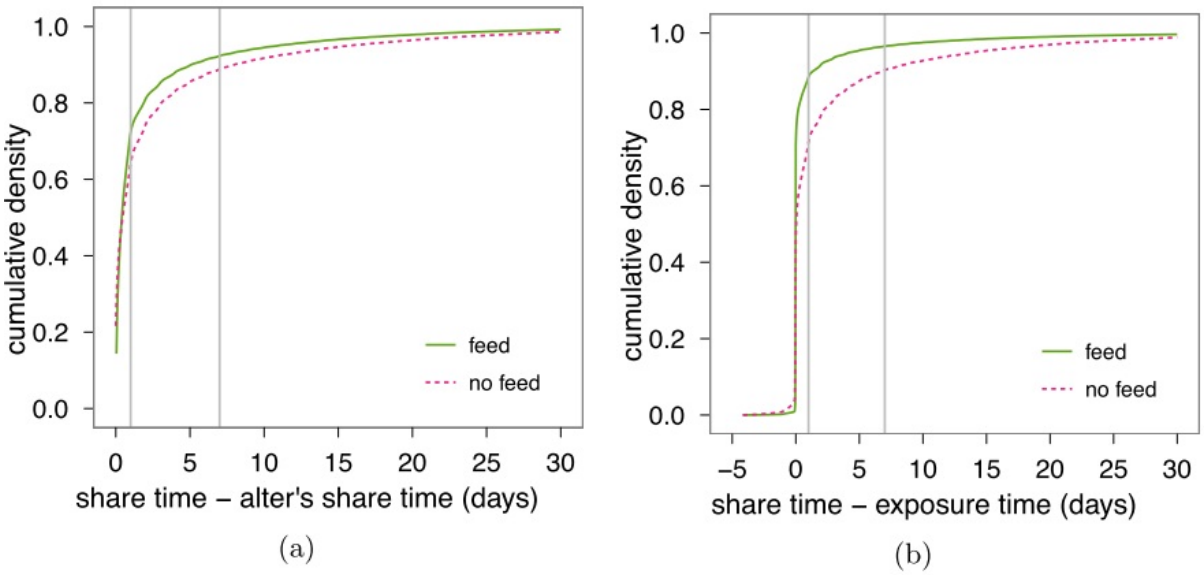


Figure 5.1: (a) The difference in sharing time between a user and their first sharing friend. (b) The difference between the time at which a user was first to exposed (or was to be exposed) to the link and the time at which they shared. Adopted from [19].

latency after a friend has already shared the information is conspicuous within one day, even within one week. After a week the information has been shared, the probability of sharing latency is small enough to endure the case in our method. Since no access to get the time a user exposed to the information, we consider these two scenarios as one scenario. Hence, we devise the *responding rate* to consider the refresh frequency and attraction of information. By the definitions of Granovetter, we divide the dataset by week to reveal the frequency and the 'time' of interactions between individuals. In the proposed method, we do not consider the 'intimacy' of tie strength. Because we only concern about whether information "delivering" or not, regardless positive or negative of the information.

We compare tie strength estimation by the proposed method to the one by [19], illustrated

in Figure 5.2. We show that tie strengths connect our participants $viola$ and $alphar$ with

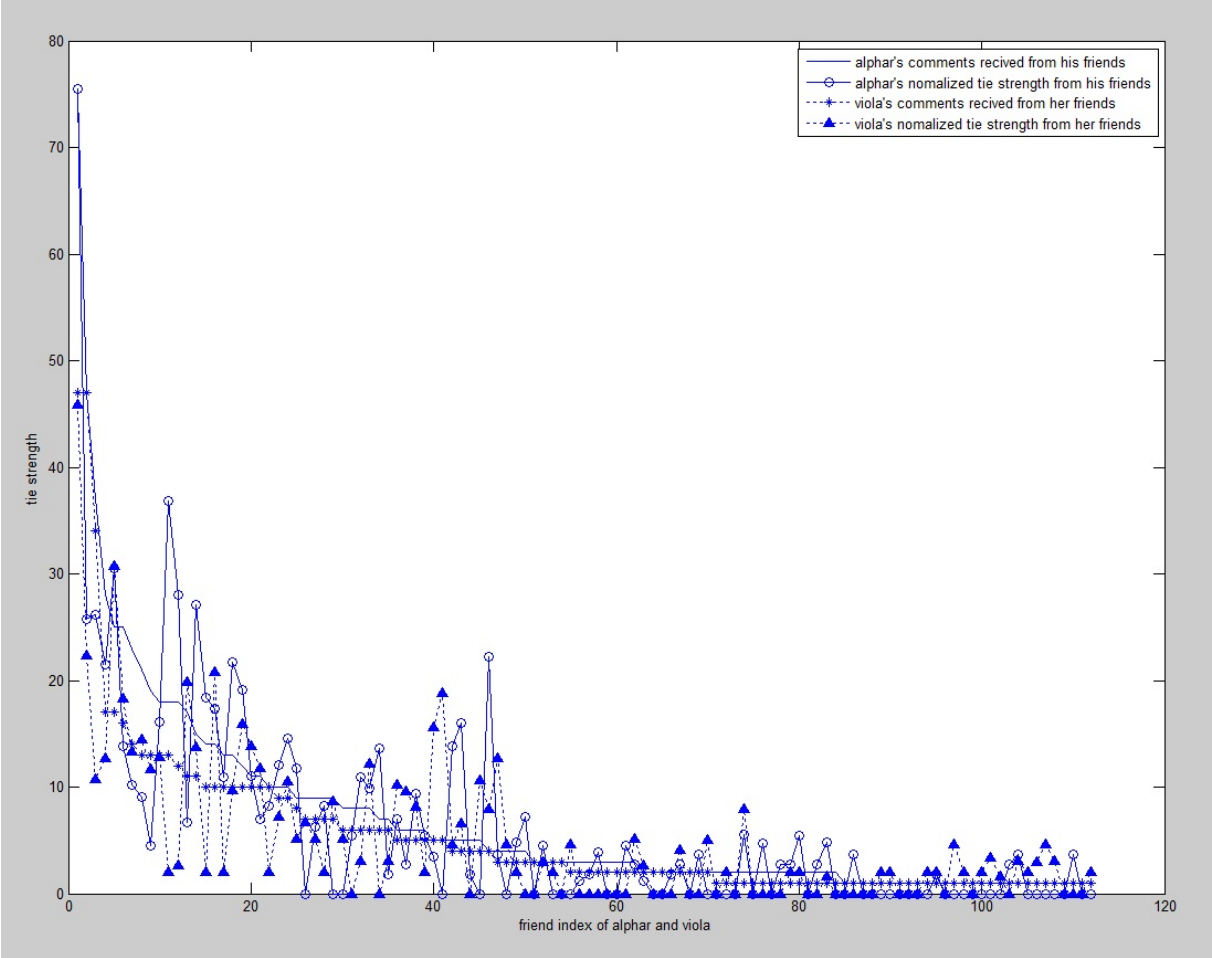


Figure 5.2: Tie strength estimation by the proposed method compares to the one by previous method.

their friends. The tie strength estimated by the proposed method is ordered by comments received. While, overall, the normalized tie strength decays similarly to comments received, the proposed method can identify the weak ties, defined by [19], which is critical for information dissemination. We adjust the corresponding weights for tie strengths. Also, we lighten the ties which belong to users having interested in some specific topics deeply but not in diverse topics. The users attracted by diverse topics are more likely to rise the propagation opportunity of information exposure within a broader context.

For a specific diffusion purpose, we do not consider "homophly" in our method. For example, $viola$ and gk are family. They have little interactions on Facebook although they are

familiar than others in real world. Besides, $viola$ and jay are in a relationship. They have either few interactions on Facebook even though they are closer than others in real world. However, $viola$ is one of lun's student. Though they have difference in 'social distance', the interactions between $viola$ and lun is apparent.

Bakshy *et al.* [19] defined 'weak' ties to friends with no interaction. However, in accordance with the definition by Granovetter, no interaction between individuals is called 'absent' tie, even though they "know" the name of each other. Since no interaction exists, the proposed method does not describe the influence of absent ties. In Figure 5.2, we do not list the friends having no interaction with our participant $viola$ during the two time sets.

5.2 Comparison

In this section, we compare the existing works to the proposed method for information propagation. In recent years, new research studies have put forward different issues of tie strength in a social network. Reviewing the previous works, we contrast the existing works mentioned in Chapter 2 with the proposed method in Table 5.1. We describe more as follows.

For offering a fine representation of relationship than binary friendship indicator, Xiang [20] proposed a homophily-based model using the similarities of user profiles and dyadic directional interactions. Xiang utilized Gaussian prior probability to explain users' relationships. The study presented the higher autocorrelation of profile attributes for relationship weights than for binary friendship.

For mapping social media data to real tie strength, Gilbert advocated Marsden's definition of tie strength using the similarities of user profiles and dyadic undirected interactions. Gilbert adopted OLS regression to approximate subjective tie strength of participants. The study ex-

Table 5.1: Existing works compare to the proposed approach.

	Xiang's method	**Gilbert's method**	**Bakshy's method**	**Proposed method**
OBJECTIVE	Finer granularity for relationship	Mapping social media data to real tie strength	Role of social network in information diffusion	Tie strength for information diffusion
PRINCIPLE	Homophily	Tie strength by Marsden	Homophily vs. Tie strength	Tie strength by Granovetter
MODEL	Gaussian prior	OLS regression	Raw data	OLS regression
FACTOR	Profile similarity, Directional interaction	Profile similarity, Undirected interaction	Information sharing event	Directional interaction
RESULT	Higher autocorrelation	Significant variables	strong tie $\rightarrow$ influential, weak tie $\rightarrow$ responsible	Delivering Path prediction, Diffusion evaluation
DIFFUSION ANALYSIS GRANULARITY	N/A	N/A	Coarse	Fine

tracted 15 significant factors out of 74 variables for tie strength.

For exploring the role of social network in information diffusion, Bakshy *et al.* designed several homophily-based experiments using information sharing events. Bakshy *et al.* employed raw data to inspect the relations between user tie strength and information propagation. The study demonstrated not only strong ties are influential but also weak ties are responsible for dissemination of novel information.

Moreover, based on research of Bakshy *et al.*, we draw attention on tie strength for information diffusion in social network. We adopt the Granovetter's definition of tie strength using dyadic directional interactions. Also, we estimate tie strength by OLS regression. Our study predict possible information delivering path and evaluate possible information diffusion.

However, that there have been few attempts to establish a direct relations between informa-

tion diffusion and tie strength. With Gilbert's method, we cannot analyze information delivering path since undirected interaction leading no directions. With Xiang's method, we can neither do information diffusion analysis owing to similarities of user profiles in lower positive correlation than the similarities that are currently believed. Although Bakshy *et al.* performed the large-scale experiments of diffusion analysis, this study regarded tie strength as a factor and gave a coarse view of diffusion in a social network. However, we provide a fine granularity, not only tie strength but also responding rate, of the dyads' relations for information diffusion. Also, we inspect a finer way of diffusion in social network such as information delivering paths.

The proposed method provides an assessment of general information diffusion. Also, we do not need too many permissions for the proposed method to disturb our participants. For the network security issue, the information may consist of malicious links. If the malicious phising link disguises successfully, the proposed method considers the potentially affected users as hidden victims. For the privacy and safety issues, the information may display personal messages. The proposed method regards the potentially affected users as likely viewers and probable spreaders.

5.3 Issues

We record the average responses of each post from participants. The participant who has the most responses are defined as the most influential user. We compare the target set with the real influential set. There are 3 out of 5 matched within two sets. And we define predicting error as follows.

$$predicting\ error = |average\ responses - predicting\ responses| \tag{5.1}$$

The predicting responses is equivalent to $\sigma^{(m)}(user, 0.1)$. The average predicting error is 3. The maximal predicting error is 17. The minimal predicting error is lower than 1. In this paper, we define 'distributional similarity' as the coverage ratio of active participant friends to active friends. For example, jay has the largest distributional similarity. He has 3 active participant friends among his 9 active friends. Distributional similarity of jay is 0.33. $viola$ has 17 active participant friends among her 74 active friends. Distributional similarity of $viola$ is 0.23.

If user has larger distributional similarity, the user gets smaller predicting error. In our dataset, the distributional similarity of user c is larger than 20%, and the predicting error of user c is lower than 2. Otherwise, the distributional similarity of user c is lower than 1%, and the predicting error of user c is between 3 and 17. Take max for instances, his distributional similarity is 20% and his predicting error is 1.15.

Going over our experiments, we examine several circumstances in this section. In comparison with the outcomes of evaluation in Table 4.7, the remaining results discriminate the property from every participants indistinctly. Taking results with $th = 0$ and $th = 0.7$ in Table 5.2 as examples, if th is too small, almost all participants pass the responding threshold to unfold the information among the whole connected network. Furthermore, if th is too large, almost all participants fail to extend the information through the connected network. We display other remaining results in Appendix A.

However, the diffusion evaluation results with $th = 0.1$ differentiate the property from every participants apparently. The evaluation results with $th = 0.1$ distribute more diversely than other results. Taking $max, ryan, bzero$ in Table 4.7 as examples, we distinguish the inconsistent $\sigma^{(m)}$ from these participants due to their tie strength with friends, although these participants have the same maximum cumulative number of potentially affected users. Also, taking $benben, mao$ as an example, even though both participants have similar $\sigma^{(m)}$, $benben$'s

Table 5.2: Diffusion evaluation with $th = \{0, 0.7\}$ ordered by $th = 0.1$. (In case $th = 0$, users get activated if $r^{(ij)} > 0$, but if $r^{(ij)} = 0$.)

user	$\sigma^{(\mathbf{m})}(\mathbf{user}, \mathbf{0})$	**maximum cumulative number of potentially affected users for** $th = 0$	$\sigma^{(\mathbf{m})}(\mathbf{user}, \mathbf{0.7})$	**maximum cumulative number of potentially affected users for** $th = 0.7$
alphar	*43.1866*	*1086*	*3.0188*	*1*
flower	*28.5267*	*1086*	*0.0*	*0*
max	*23.9585*	*1086*	*1.0567*	*1*
ryan	*24.0749*	*1086*	*0.0*	*0*
hsuan	*26.0559*	*1086*	*0.0*	*0*
terry	*25.2049*	*1086*	*0.0*	*0*
bzero	*20.5095*	*1086*	*0.0*	*0*
mei	*26.8897*	*1086*	*0.0*	*0*
claire	*29.1228*	*1086*	*0.0*	*0*
benben	*23.593*	*217*	*0.0*	*0*
stone	*27.4131*	*1086*	*0.0*	*0*
mao	*28.0695*	*1086*	*0.0*	*0*
min	*26.5757*	*1086*	*0.0*	*0*

tie strengths with friends is quite stronger than mao's, which leads to different maximum cumulative number of potentially affected users obviously. Hence, while we do our experiments for diffusion evaluations with responding threshold $th = \{0, 0.1, 0.2, ..., 0.7\}$, we only display the evaluation results with $th = 0.1$ in section 4.4.2.

Since the invitations of our experiments are sent to users not only who frequently use but also who infrequently use Facebook, the behaviour of inactive participants is hard to predict. In our dataset, it is common situation that participants who infrequently visit Facebook post only one piece of information on their wall every week or every few weeks. However, according to the proposed method, no interactions within few weeks is treated as sparsity of training data in our dataset. Thus, the participants who infrequently visit Facebook get the regression coefficient α, β with large variance due to insufficient data. The cases of participants infrequently using

Facebook guide various outcomes. Sparsity also causes the large predicting error, even though the user has larger distributional similarity.

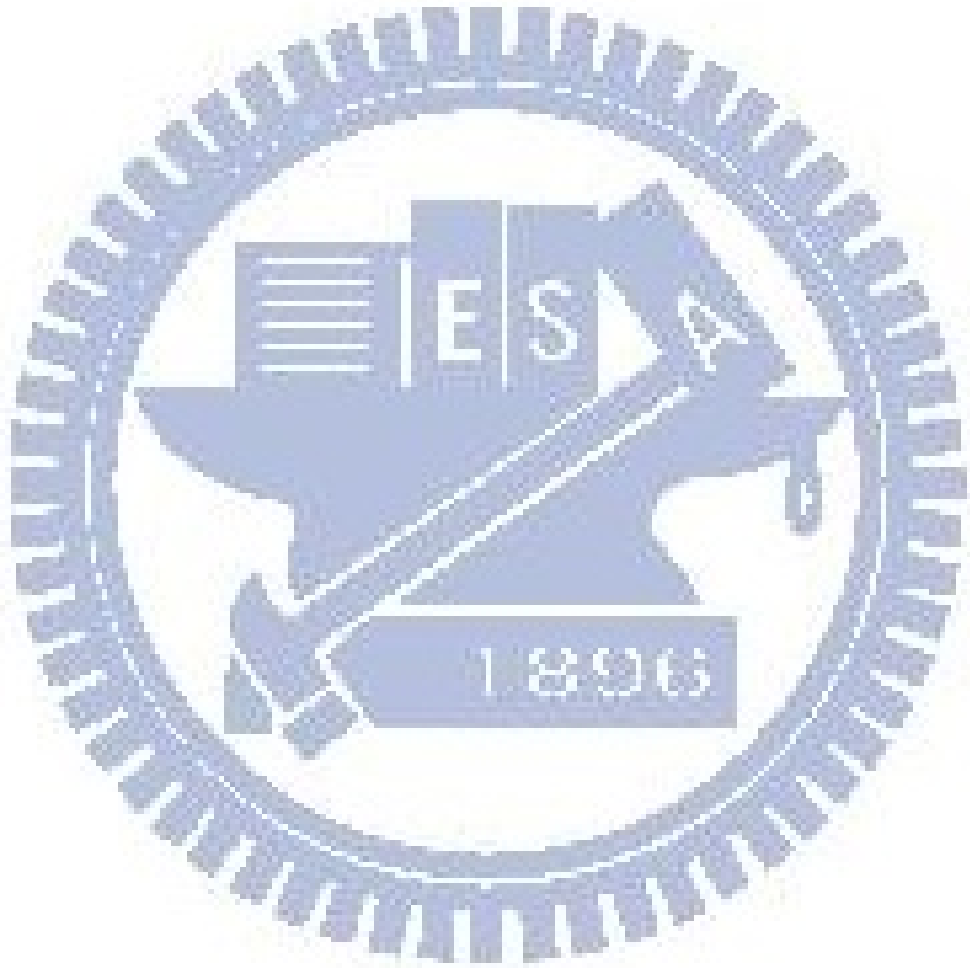
During the time sets in our experiments, Facebook originally published **Timeline** beta on Sept. 22th, 2011. Timeline is a new kind of user profile. The way of Timeline displaying the stories changes the way of user behaviours. It's much easier to get the information through Timeline than order user profile. Timeline updated on Dec. 6th, 2011 and roll it out in New Zealand. Some of our participants get Timeline for their profiles before the experiments or during the experiments, but the others don't. Hence, the inconsistency on user wall displaying may involve with the proposed method.

Also, **Ticker** is introduced in Sept. 2011. The users with English version get this product on their Facebook first. The introduce of Ticker to users in de-synchrony involves with the proposed method as well. The observations on new Facebook products deserve people's attentions.

Even though Facebook strengthens security with partnerships in attempt to protect its 900 million users from spam and malicious content, malicious Web links still pop up on the social network [27]. Besides, many issues rise due to the popularity of social network such as marketing, privacy and safety issues. The proposed method delineating the way of information spreading on Facebook is noteworthy although only less than 4 percent of content shared is spam now.

The proposed method provides a preliminary idea of information diffusion in social network for the condition that information is shared on user i's wall by user i. Future work will hopefully examine the condition that information is shared on user i's wall by user j. An additional interesting avenue of investigation might be to consider the condition that the information is shared as a comment. However, the EdgeRank that Facebook use to determine which story

comes out on the News Feed is not announced in detail. Also, no entry is for sharer of each link via FQL now. We only can access the `like` data as a list of the viewing user's friends who like the post. Future studies should be alerted to the desired data limitation.



Chapter 6

Conclusion

In this research, we propose a method to measure tie strength of the dyads for strength of interaction, instead of for relationship. We define a responding rate to represent the opportunity for information propagation. We further predict the information delivering path of a wall post. Also, we model information diffusion for 1-hop dissemination and m-hop spreading.

Sequentially, we conduct experiments to estimate the tie strength and responding rate for our participants by analyzing 6 months data. By verifying the estimation with the characteristics of user behaviours, we build the connections between the participants. With the proposed method, we can select an efficient target set from 9550 relationships. The executing reduction is noticeable. In other words, we can predict the diffusion coverage of a piece of information from a specified individual. After analyzing the tie strengths and responding rates in the real data, we verify the existence of the information delivering path predicted by our method. According to our prediction, we find out that attack survivability is not equivalent to the connectivity while human behaviours are included in a scale-free network, especially when the information contains malicious links. Furthermore, we provide a preliminary model to evaluate information diffusion within a broad context.

An area of future research that should be considered is the information topic attractiveness in a finer granularity. In this method, we use one out of the three elements for EdgeRank so far. Adding the other two elements into diffusion model is obviously required in the future work, but this is an exciting first step for analysis of information diffusion.

Appendix A

As mentioned in section 4.3.2, the remaining verification of tie strengths and responding rates are in Table A.1 and Table A.2, respectively. We figure out that the results with fewer interactions fluctuates larger. Larger fluctuation does not represent the lower accuracy of our method. The infrequent users use Facebook irregularly. Hence, they are much harder to predict their behaviours. We will consider another model to describe infrequent users in the future work.

Table A.1: Order of users have potential influence toward their friends by tie strength when tie strength fluctuates with time varying. (T_1:[Sept.-Nov.] T_2:[Dec.-Feb.])

user	**integrated tie strength**		**fluctuation ratio** (4.1)
	$\mathbf{T_1}$	$\mathbf{T_2}$	
ryan	*10.839*	*12.208*	*↓0.1262*
minhsi	*10.34*	*12.025*	*↑0.1629*
jason	*9.198*	*11.827*	*↓0.2858*
flower	*8.933*	*10.246*	*↑0.1471*
hsuan	*8.93*	*9.71*	*↓0.0872*
max	*8.844*	*9.054*	*↑0.0238*
terry	*7.695*	*8.977*	*↓0.1666*
lun	*6.954*	*7.259*	*↓0.0439*
eddy	*6.587*	*6.777*	*↓0.0288*
gk	*6.455*	*5.411*	*↑0.1618*
bzero	*5.6*	*4.613*	*↑0.1762*
peter	*3.454*	*2.0*	*↑0.4209*
min	*3.333*	*1.714*	*↑0.4857*
jay	*2.0*	*0.0*	*↑1.0*

As mentioned in section 5.3, the remaining diffusion evaluations are in Table A.3 and Table A.4. We figure out that the influence of participants with high $\sigma^{(1)}$ and low integrated responding rates is enclosed. For instances, mao and $stone$ are the top 10 participants for influencing their friends. Since their integrated responding rates are lower than those who are

Table A.2: Order of users have potential influence toward their friends by tie strength when responding rate fluctuates with time varying. (T_1:[Sept.-Nov.] T_2:[Dec.-Feb.])

user	**integrated responding rate**		**fluctuation ratio** (4.1)
	$\mathbf{T_1}$	$\mathbf{T_2}$	
min	*3.333*	*1.714*	*↑0.4857*
ryan	*3.249*	*2.474*	*↑0.2388*
ann	*2.962*	*3.114*	*↓0.0513*
hsuan	*2.917*	*7.133*	*↓1.4457*
terry	*2.85*	*2.283*	*↑0.199*
eddy	*2.785*	*2.221*	*↑0.2025*
viola	*2.25*	*4.709*	*↓0.8835*
jason	*2.409*	*2.96*	*↓0.2289*
channing	*2.313*	*5.89*	*↓1.5469*
mao	*2.244*	*2.764*	*↓0.2315*
lun	*2.086*	*2.453*	*↓0.1759*
jay	*2.0*	*0.0*	*↑1.1*
gk	*1.758*	*2.608*	*↓0.4839*
peter	*0.439*	*2.0*	*↓3.5485*

the top 10 for influencing their friends, their influences are smaller for multi-hop diffusion. However, the influence of participants with higher either integrated or individual responding rates are more likely to be disclosed. For examples, $flower$, max and min are not the top 10 for influencing their friends. Since their either integrated or individual responding rates are higher than those who are the top 10 for influencing their friends, their influences last longer through multi-hop diffusion.

Table A.3: Diffusion evaluation with $th = \{0.2, 0.3\}$ ordered by $th = 0.1$.

user	$\sigma^{(\mathbf{m})}(\mathbf{user}, \mathbf{0.2})$	**maximum cumulative number of potentially affected users for** $th = 0.2$	$\sigma^{(\mathbf{m})}(\mathbf{user}, \mathbf{0.3})$	**maximum cumulative number of potentially affected users for** $th = 0.3$
alphar	*14.4411*	*17*	*6.8323*	*4*
flower	*9.4475*	*43*	*5.1920*	*18*
max	*10.4899*	*33*	*5.2539*	*8*
ryan	*9.4759*	*33*	*1.5300*	*1*
hsuan	*2.7257*	*3*	*1.3090*	*1*
terry	*0.0*	*0*	*0.0*	*0*
bzero	*7.6855*	*33*	*4.1016*	*13*
mei	*8.1314*	*8*	*8.1314*	*8*
claire	*4.1938*	*10*	*0.9621*	*1*
benben	*2.2242*	*2*	*1.3383*	*1*
stone	*2.5193*	*5*	*0.7377*	*1*
mao	*4.2908*	*9*	*1.7632*	*1*
min	*4.8088*	*11*	*3.3333*	*5*

Table A.4: Diffusion evaluation with $th = \{0.4, 0.5, 0.6\}$ ordered by $th = 0.1$. (**MCNPAU** for maximum cumulative number of potentially affected users)

user	$\sigma^{(\mathbf{m})}(\mathbf{user}, \mathbf{0.4})$	**MCNPAU** for $th = 0.4$	$\sigma^{(\mathbf{m})}(\mathbf{user}, \mathbf{0.5})$	**MCNPAU** for $th = 0.5$	$\sigma^{(\mathbf{m})}(\mathbf{user}, \mathbf{0.6})$	**MCNPAU** for $th = 0.6$
alphar	*3.0188*	*1*	*3.0188*	*1*	*3.0188*	*1*
flower	*0.6313*	*1*	*0.0*	*0*	*0.0*	*0*
max	*3.3084*	*4*	*2.6599*	*3*	*1.9093*	*2*
ryan	*1.5300*	*1*	*0.0*	*0*	*0.0*	*0*
hsuan	*1.3090*	*1*	*0.0*	*0*	*0.0*	*0*
terry	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*
bzero	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*
mei	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*
claire	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*
benben	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*
stone	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*
mao	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*
min	*3.3333*	*5*	*3.3333*	*5*	*3.3333*	*5*

References

[1] N. Ellison *et al.*, ”Social network sites: Definition, history, and scholarship,” *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[2] Communities and L. Government, ”Online Social Networks.” [Online]. Available: http://www.umic.pt/images/stories/publicacoes2/Online_Social_Networks.pdf

[3] [Online]. Available: http://onlineschools.org/

[4] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, ”Measurement and analysis of online social networks,” in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 29–42.

[5] M. E. J. Newman and J. Park, ”Why social networks are different from other types of networks.” *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, vol. 68, no. 3 Pt 2, p. 036122, 2003. [Online]. Available: http://arxiv.org/abs/cond-mat/0305612

[6] M. Granovetter, ”The strength of weak ties: A network theory revisited,” *Sociological theory*, vol. 1, no. 1, pp. 201–233, 1983.

[7] M. Sosa, ”1. title: Where do creative interactions come from? the role of tie content and social networks,” *Organization Science*, vol. 22, no. 1, 2011.

[8] [Online]. Available: http://mashable.com/2012/04/23/facebook-now-has-901-million-users/

[9] [Online]. Available: http://www.nytimes.com/2009/10/25/fashion/25facebook.html?_r=2&pagewanted=all

[10] T. Stein, E. Chen, and K. Mangla, ”Facebook immune system.” [Online]. Available: http://research.microsoft.com/en-us/projects/ldg/a10-stein.pdf

[11] J. Goldenberg, B. Libai, and E. Muller, ”Talk of the network: A complex systems look at the underlying process of word-of-mouth,” *Marketing letters*, vol. 12, no. 3, pp. 211–223, 2001.

[12] P. Domingos and M. Richardson, ”Mining the network value of customers,” in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 57–66.

[13] L. Zhao, K. Park, Y. Lai, and T. Cupertino, ”Attack induced cascading breakdown in complex networks,” *Journal of the Brazilian Computer Society*, vol. 13, no. 3, pp. 67–76, 2008.

[14] M. Granovetter, ”Threshold models of collective behavior,” *American journal of sociology*, pp. 1420–1443, 1978.

[15] [Online]. Available: http://edition.cnn.com/2011/10/03/tech/social-media/facebook-phishing-scams/index.html?iref=allsearch

[16] T. Paek, M. Gamon, S. Counts, D. Chickering, and A. Dhesi, ”Predicting the importance of newsfeed posts and social network friends,” in *Proc. AAAI*, vol. 10, 2010.

[17] E. Gilbert and K. Karahalios, ”Predicting tie strength with social media,” in *Proceedings of the 27th international conference on Human factors in computing systems*. ACM, 2009, pp. 211–220.

[18] F. Kivran-Swaine, P. Govindan, and M. Naaman, ”The impact of network structure on breaking ties in online social networks: Unfollowing on twitter,” in *Proceedings of the*

2011 annual conference on Human factors in computing systems. ACM, 2011, pp. 1101–1104.

[19] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," *Arxiv preprint arXiv:1201.4145*, 2012.

[20] R. Xiang, J. Neville, and M. Rogati, "Modeling relationship strength in online social networks," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 981–990.

[21] M. McPherson, L. Smith-Lovin, and J. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, pp. 415–444, 2001.

[22] P. Marsden and K. Campbell, "Measuring tie strength," *Social forces*, vol. 63, no. 2, pp. 482–501, 1984.

[23] [Online]. Available: http://newsroom.fb.com

[24] [Online]. Available: http://developers.facebook.com

[25] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[26] R. Albert, H. Jeong, and A. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[27] [Online]. Available: http://seattletimes.nwsource.com/html/businesstechnology/2018066822_apustectechbitfacebooksecurity.html