IDENTIFYING INFLUENTIAL USERS BY THEIR POSTINGS IN SOCIAL NETWORKS

SUN BEIMING

M. Phil

The Hong Kong

Polytechnic University

2014

The Hong Kong Polytechnic University

Department of Computing

Identifying Influential Users by Their Postings in Social Networks

Sun Beiming

A thesis submitted in partial fulfillment of the requirements

for the degree of Master of Philosophy

August 2013

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

SUN BEIMING (Name of student)

Abstract

With the rapid development and increased popularity of social networks, much research effort has been conducted to analyze information of social networks, such as finding the influential users. Our research is focusing on identifying the influential social network users; as it can help to increase the marketing efficiency, and can also be utilized to gather opinions and information on particular topics as well as to predict the trends. Different from previous work, our aim is to identify the most influential users based on the interactions in their posts on a given topic.

We first propose a graph model of online posts, which represents the relationships between online posts of one topic. Three measurement methods have been developed to assess the influences of posts, so as to find the influential posts on the topic. In our work, there are two types of influences based on the different roles: starter and connecter. A starter is followed by many others, similar to a hub in a network, so it should have certain influence. A connecter is also regarded to be influential when it links starters together.

After we can measure the influence of online posts and find the influential posts, their authors can be considered as potential influential users. With the consideration of some users would have several influential posts, we develop a user graph model to refine the influence measures to find influential users. Based on the authors of influential posts found, we convert a post graph to the corresponding user graph, and then measure the influence of users which are starter and connecter respectively. Finally, the most influential users can be determined in the user graph. Also, our proposed model can be extended and used to find the sentimental influence of posts and users.

We have conducted two case studies in order to verify our proposed graph models and influence measurement methods. In the first study we applied the graph model of online posts and visualized the result of starter and connecter identifications. The experiment is performed on Twitter, and it shows that the influential starters and connecters in the post graph can be identified after integrating the results from three measurement methods. In order to validate our model, we compared the results of our methods with three centrality metrics and the PageRank algorithm. The experiment result shows that our proposed methods outperformed the others in the ability of identifying both starters and connecters. Next, the influential users identified by the post graph model and the user graph model are compared in the second case study. The results show that users with more influential posts may not be truly influential, when the followers are always the same group of people. In a user graph obtained from a given post graph, the connecters already identified in the original post graph can be refined and some new ones found are considered to be potential connecters.

List of Publications

[1] **B. Sun** and V. TY Ng. Lifespan and Popularity Measurement of Online Content on Social Networks. In *Social Computing Workshop of IEEE ISI Conference*, pp.379-383, 2011.

[2] **B. Sun** and V. TY Ng. Identifying Influential Users by Their Postings in Social Networks. In *Proceedings of the 3rd International Workshop on Modeling Social Media*, pp. 1-8, 2012 (Notable Paper in Computing Reviews Best of 2012).

[3] **B. Sun** and V. TY Ng. Identifying Influential Users by Their Postings in Social Networks. Ubiquitous Social Media Analysis, LNCS Vol 8329, pp. 128-151, Springer Berlin Heidelberg, 2013.

[4] **B. Sun** and V. TY Ng. Analyzing Sentimental Influence of Posts on Social Networks. In *Proceedings of the IEEE 18th International Conference on Computer Supported Cooperative Work in Design,* pp. 546-551, 2014.

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Dr. Vincent Ng from Department of Computing, the Hong Kong Polytechnic University, for his valuable suggestions and patient guidance during my research work, his supervision and support in my study progress, and his time and efforts spent in reviewing my papers and helping me to complete the thesis. I also appreciate the study opportunity provided by Dr. Vincent Ng.

Besides my chief supervisor, I would like to thank Prof. Kam Fai Wong for being my thesis examiner, and giving the valuable suggestions and corrections to this work. Thanks to Dr. Reynold Cheng for being my thesis examiner and providing valuable suggestions and insightful comments.

I would also like to thank Li Ho Leung and Victor Liang, the research students in our group, for giving me valuable advices and supports on my research work. I would like to extend my gratitude to Andy He, for his support and encouragement during this period.

Finally, I offer my regards and blessings to my family and friends who supported me during the development of this thesis, and to all the people who helped me to complete this work.

.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

Social networking websites have been around since the mid-90s, but over the past few years the popularity of social networking sites has rapidly increased. Recently the wide usage of smart phones further facilitated the growth of social networking services, especially for the micro-blogging sites. A social networking service is a platform to build social networks or social relations among people who, for example, share common interests, activities, backgrounds, or real-life connections. Social networking sites are like online communities which allow users to join and communicate with others.

Facebook is the most popular social network website with more than 1 billion active users. As announce by Facebook in March 2013, it has a 23% growth from one year before, and 751 million people using Facebook from their mobile devices each month [40]. Twitter is another popular site that provides micro-blogging service. It has over 500 million registered users by 2012, generates 400 million tweets daily, and 43% of users access Twitter through mobile phones [15]. Besides, there are many other social networking sites with increasing popularity, such as LinkedIn, Google Plus and MySpace. Social network sites have huge amount of data online, which has triggered a lot of interests from researchers. Academic researchers proposed models trying to analyze the information diffusion pattern and people's behavior online. Most marketers now use social media for business, especially in product promotion and review collection.

1.1 Motivation

The influential users on social networks obtained great interest from business parties. In

fact, a piece of information can be spread from one individual to another through the social network in the form of “word-of-mouth” communication. For example, the news of good and free service such as Gmail could be fast and largely spread among people through social networks. Therefore, targeting a small group of influential individuals for the product promotion can increase the marketing efficiency. Google realized the importance of identifying influential users when trying to improve advertising at MySpace [18]. New algorithms based on PageRank have been developed for ranking the most influential people on social networking sites such as MySpace and Facebook [33]. The approach changed to target advertisements on users who have more influence, not simply those with certain characteristics in their profiles. Besides, finding influential users can also be utilized to gather opinions and information on particular topics, such as collecting reviews of a new product. Another benefit is that tracing the activities of influential users can help to predict the trends.

In order to find these influential users, the first problem is to measure a user’s influence on social networks. In the past, there has been a lot of work on judging the influence of users on a social networking site. For example, many measurement metrics have made use to measure the relationships between users (i.e. follower / followee) in Twitter. However, they mostly ignore the interactions of users in their online posts. Moreover, without the consideration of the contents posted by users, they are not able to tell the influence of users on different topics. Our objective is to develop a model to find the influential users through their postings and interactions online. This model could be applied in all social networks with textual posts and authors and should be able to identify the influential users on specified topics.

According to the statistical report from the Central Registry of Drug Abuse in Hong Kong and the survey conducted by the HKSAR Government in recent years, the age range for students abusing psychotropic substance is getting younger emphatically. "The percentage of lifetime drug-taking secondary students increased from 3.3% in 2004/05

to 4.3% in 2008/09, and that of 30-day drug-taking secondary students increased from 0.8% in 2004/05 to 1.5% in 2008/09." As the problem of youth drug abuse is getting more and more serious, it is pressing to take measures against it. Meanwhile, it is found that writing blogs is a popular activity for drug abusers in Hong Kong, and they will post drug related information there. There are several popular social networking platforms with different functionalities, such as Facebook, Twitter and HK Discussion (a local famous forum). Many of those people have accounts on each social networking site. It motivated our initial work to develop a Web miner (D-Miner) which can automatically collect data across different sources, parse the posting contents and detect information related to potential drug abusers for social workers or teachers to follow [17]. After many drug related posts have been retrieved, we are interested in finding those with greater influence. This leads our research into the direction of measuring the influence of posts within a specified topic at the beginning. The topic can be extended to others and not limited to drug abuse.

1.2 Problem Statement

In order to identify the influential users within a topic, we should first solve the problem of how to measure the influence of online posts on that topic, and then find the authors of the most influential posts. The influence may have different definitions and accordingly be measured with different metrics. However, the influence is always from one to another; therefore we should first figure out the relations of the online posts before we can measure their influences.

Usually, posts are considered to be related within a thread or a chain. However, their relationships can be more complicated in certain cases. For example, a post is replying to a previous post while its content refers to a different one. Other than the direct responses as explicit relationships, there are also implicit relationships between online

posts. For example, a user has read a post online. Instead of directly replying to it, he writes a new post on this topic. In this scenario, the two posts are considered to be implicitly related, because the action of later posting is influenced by the earlier one [13, 14]. Considering these situations, we build a graph model to represent the explicit and implicit relationships between online posts on a topic. Based on the model we try to identify the influential posts and users based on their direct interactions as well as the relationships due to similar contents on the same topic.

Not all the posts on the same topic can have implicit relationship. If a friend posted on the same topic a long time apart after the user did, the chance of the latter post being influenced by the earlier one is slim. We can say that every post has its lifespan in a social network, and it dies when there is no response to it for a long time (including replies, shares, and new posts on the same topic). Hence, we develop a way to define and determine the lifespan of a post.

After we can measure the influence of online posts and find the influential posts, their authors can be considered as potential influential users. Furthermore, we need to judge who the influential users are. Considering the cases that influential users may have several influential posts, we would like to develop a user graph model to refine the influence measures for potential influential users.

Overall, my research is aimed to find the most influential posts and users on social networks which deserve most attention. Altogether, the problems can be summarized as follows:

i. Determining the lifespan of online posts on social networks
ii. Measuring the influence of online posts on a topic
iii. Measuring the influence of users and identifying the influential users

1.3 Contribution

In order to determine the lifespan of online posts, we should be able to know the death time of a post. We propose the Comment Arrival Model [6] to simulate the process of comments arriving and to determine the death time of a post. In this model the expiry time of posts is computed so that a post can be determined as dead if there is no more comment or reply received within the time. The expiry time of posts is also used in judging the relevance of two posts in the post graph model construction. Their relevance is considered to be weakened if the time interval between the two posts exceeds the expiry time.

We propose a graph model to represent the relationship between posts on a topic, and three measurement methods to assess the influences of posts [5]. The results from the three measurements are integrated to determine the influential posts. In our work, there are two types of influences based on the two roles: starter and connecter. A starter is followed by many others, similar to a hub in a network, so it should have certain influence. The connecter is also regarded to be influential when it links starters together.

Based on the authors of influential posts found, we convert the post graph to the user graph, and then refined the influence measures of users acting as starter and connecter. Finally, the most influential users can be identified from the user graph which is converted from the post graph. Furthermore, our proposed model can be extended and used to find the sentimental influence of posts and users.

1.4 Outline of the Thesis

The thesis is organized as follows. Chapter 2 reviews some related works and a preliminary study on the lifespan of online posts is presented at the beginning of Chapter 3. Then a graph model is defined in this chapter to represent the relationships between online posts, and three different methods of influence measure are proposed based on the graph model. After that, the sentimental influence analysis on influential posts is introduced. Chapter 4 defines the user graph model, and presents the conversion from the post graph to user graph, and the measurements of user influences. Chapter 5 discusses the tests with different cases to verify our models. Finally, we summarize the paper and suggest for future work in the last chapter.

Chapter 2 Literature Review

In the past, research mainly focused on using graph to analyze the influence or popularity of users or topic terms [3, 9, 28, 34], but we aim to propose a graph model to represent the relationship of online posts within a topic in order to measure the influence of posts. Then, we try to identify the most influential posts and users on this topic. Different from former work, we first formally define the lifespan of online posts, and the adjustment factors are taken into consideration that depends on the special features of a particular social networking media in practice. With the information of the explicit and implicit relationship between posts, our model tries to find the real influential posts, and to identify the most influential users based on their direct interaction as well as the underlying relation in posting online. In the next few sections, related works are presented.

2.1 Influence Measurements in Social Networks

Many methods have been proposed to measure users' influence on Twitter. A popular metric of influence is the number of a user's follower [2]. It assumes that all followers will read the contents published by that user. Yet, this method ignores the different ways of users to interact with the online contents. There are also many online tools to measure a user's influence on social network, such as Klout Score [26] and Twinfluence [39]. However, they cannot tell the influences of users on different topics. In [25], the TwitterRank algorithm, which is an extension of PageRank, was proposed to measure the user influence on Twitter taking both the topical similarity between users and the link structure into account. TunkRank [10] is another adaptation of PageRank. It makes the assumption that if a user reads a tweet from his friend he will retweet it with a

constant probability. The influence is calculated recursively considering the attention a user can give to his friends, and that their followers could attribute to their influence as well. These methods do not consider users' interaction in posts. Yet, it is interesting to judge their influences not by their friendship relations in static structure, but based on the dynamic interaction in online contents.

In the Merriam-Webster dictionary, influence has been defined as "the power or capacity of causing an effect in indirect or intangible ways". Actually, there is no standard method to measure the influence of users online [30]. In our scope, we define the influence as the ability of a user with an action to initiate a further action by other users. We want to judge their influences not by their friendship relations in static structure, but based on the dynamic interaction in online contents. Similar to our approach, Tang et al proposed the Topical Factor Graph Model [24] to analyze social influence between users. They also made assumption that users with similar interests or whose actions frequently correlated would have a stronger influence on each other [21]. Here "similar interests" can be detected by the similarity of textual contents posted by users, and "correlate actions" may include replies and retweets which depend on the social media functions. Besides, they considered the factor of topic popularity when measuring the influence strength. To illustrate, if a user posts a tweet on "Obama" on Twitter, and his friend retweets it or also posts on this topic, then this friend is influenced by the user, and his influence strength is estimated by an increased probability ($p_1 - p_2$), where p_1 is the probability of the user's friends talking on this topic after the user and p_2 is the average probability of all users in the network to talk on this topic. The reason to consider the topic popularity is that if it is already a very hot topic that many people are discussing, social users may be mainly influenced by the global trend, instead of one or two friends.

In their model, a hidden vector $\boldsymbol{y}_i \in \{1, \ldots, N\}^T$ is defined to model the topic-level influences from other nodes to node v_i in a graph of N nodes. Each element y_i^z represents

the node that has the highest probability to influence node v_i on topic z. Then they defined node, edge and global feature functions, and joined them as the objective likelihood function which is to be maximized. Node feature function is defined by the intuition that if node v_i has a high similarity with node v_{yi}, then v_{yi} may have a high influence on node v_i. The edge feature function is defined with a binary value to indicate if there is an edge between the two nodes. As for the global feature function, it is just a constraint to avoid finding the most influential node to v_i as itself. In [21], the definition of influence on social networks is similar to ours. If a user posts on a topic and his friend responses to it or also posts on this topic, then the friend is considered to be influenced by the user. Their work also considered the interaction between users on a topic. However, their model can only catch the information of conversations between two users, without the consideration of discussion threads or chains (as defined in Section 4.1).

In [43] and [44], the influence is happened when a user makes a similar action following another. The influence first comes from one user to another, and can be propagated to the third one through the network. There is no consideration of the relationship between posts within or out of the discussion thread. In their assumption, the influence from *a* to *c* exists with a probability as long as *a* has influenced *b*, and *b* has influenced *c*. Our model defines explicit and implicit relationship between posts, and can be applied in social media platforms that contain textual posts. In our post graph model, it is assumed that *a* can have influence on *c*, only when *b* follows *a* (e.g. reply, share), and *c* follows *b* on the same topic within a thread.

Although they considered the interaction of users, they did not consider the factor of time interval between related posts. The relevance of posts should get weaker along with the time, so that the influence between them also declines. Moreover, their model cannot identify different types of influence, such as different influential roles or sentimental influences.

2.2 Important Nodes / Roles Detection

Hansen, Shneiderman and Smith defined two primary types of threaded conversation networks: reply networks and affiliation networks, where the reply network can be further divided into direct reply network and top level reply network [12]. The direct reply network connects a replier to the person who is directly replied, which means he can be connected to the original author or another replier, while the top level reply network connects all repliers within a thread to the original author. Thus a top level reply network will emphasize on those who post the top-level messages and start the threads, but de-emphasize conversations that occur midway within a thread. It is suitable for some communities with short threads. However, in discussion forums with longer threads, a direct reply network is often used since people are often replying to each other later. In order to identify important people within a community, Hansen et al. also defined three social roles based on graph metrics. They are question people, answer people and discussion starters [12].

Question People: post a question and receive a reply by one or two individuals who are likely to be answer people. In a network, they can be easily identified by low indegree (receive messages from others) and low outdegree (send messages to others).

Answer People: mostly send messages to individuals who are not well connected themselves [19]. In the network they have high degrees as well as a high percentage of outdegree, meanwhile they have low clustering coefficient, which is defined as the percentage of neighbours who are connected.

Discussion Starters: mostly receive messages often from people who are well-connected to each other. In the network they have high indegrees but a low percentage of outdegree,

as well as high clustering coefficient. Answer people score is created to distinguish between the answer people and discussion starters:

$$S = \frac{outdegree}{indegree+outdegree} \times \frac{1}{clustering\ coefficient} \quad (2.1)$$

According to their definition, those with high scores are decided to be answer people and those who score low are discussion starters. This metric does not suit our model, because the clustering coefficient is better for dealing with an undirected graph or a directed graph with loops.

There is similar work done by other researchers. Mathioudakis and Koudas [32] formalized the notions of 'starters' and 'followers' for bloggers on social media. The starter does not mean the first one to open the discussion but the one who triggers an intense discussion. They expected that a blogger, who primarily generates posts that others link (inlinks) over a significant period of time, could be a starter, and the bloggers who primarily generate posts that links to other blog posts (outlinks) would be followers. They compared which bloggers behave more as 'starters' by computing the difference between the number of inlinks and outlinks of their blogs:

$$d(b) = \#inlinks(b) - \#outlinks(b) \quad (2.2)$$

A blog *b* is distinguished as 'starter' if the number of inlinks related to it is high (many other posts linking to it) and the number of outlinks is low (seldom linking to other posts), which result in high *d*(*b*). Those with low *d*(*b*) are considered to be 'followers'. They compared which bloggers behave more as 'starters' by computing the difference between the number of inlinks and outlinks of their blogs. Their experiments showed that it is possible to identify the top starters for a given query of topic words in BlogScope. In our work, we adopt the definition of the role of starter. In addition, we also propose connecters that link starters together as they are influential too. With this understanding, we try to find influential posts through the identification of starters and connecters.

Specially, Shetty and Adibi proposed the Entropy model to identify the most important nodes in a graph [23]. They dealt with the problem of finding leaders in a network. They built the graph so that nodes are representing persons or organizations and edges are representing actions they are involved in. They determined the important nodes by those who have the most effect of the graph entropy when they are removed from the graph. They used the event based entropy that has been similarly defined in [8]. Let $G =<V, E>$ be a graph. P is the probability distribution on the vertex set $V(G)$. They treated $V(G)$ as a finite alphabet. Then the graph entropy can be defined as:

$$H(G,P) = \sum_{i=1}^{|V|} p(v_i) \log(1/\ p(v_i)) \tag{2.3}$$

Taking the domain of emails for example, $P(AemailB)$ can be calculated as the number of occurrences of *AemailB* in the graph, divided by the size of the graph. If the link *AemailB*and *BemailC* are dependent to each other, this means B may forward A's email to C. For this reason, they varied the probability space from length = 1 to length = 2 and more. For instance, choose length of 2 and count sequences such as *AemailBemailC* and *BemailDemailE*. Then $P(AemailBemailC)$ would be the number of occurrences of such sequence over all possible sequences with length equal to 2 in the graph. They used the event based entropy that has been similarly defined in [23]. Their experiment showed that comparing to conventional techniques such as betweenness centrality, this method leads to a better result. More important nodes can be discovered based on their effect on graph entropy in the ordered network. However, the graph entropy model claims its results on certain assumptions, like the evidence data is complete and with no noise.

Inspired by their ideas, we propose a method of measuring the influence of online posts through a refined graph entropy approach. In addition, the methods of Degree Measure and Shortest-path Cost Measure are exploited and integrated their results to identify the most influential posts. The details are discussed in *Section 3.4*.

2.3 Sentiment Analysis and Emotion Detection

In [35], Asur and Huberman constructed a sentiment analysis classifier to label each article as positive, negative or neutral. They considered subjectivity and polarity as two factors for sentiment analysis. To capture the subjectivity on Twitter, they defined a measure as:

$$Subjectivity = \frac{|Positive\ and\ Negative\ Tweets|}{|Neutral\ Tweets|} \tag{2.4}$$

They also measured the ratio of positive to negative tweets:

$$PNratio = \frac{|Tweets\ with\ Positive\ Sentiment|}{|Tweets\ with\ Negative\ Sentiment|} \tag{2.5}$$

Similar to their work, O'Connor et al. did opinion estimation on Twitter and also used PN ratio measure as in (2.5) to find the daily sentiment on a topic [7].

However, many research works are focusing on classifying and summarizing sentiment by topics, but seldom by users. Analyzing sentimental influence of users is valuable. After we find the most influential posts and users, we can further determine their influence types. It is interesting to know whether the followers comply or oppose the influential user.

Chapter 3 Identification of Influential Posts

On social networks, if a post does not receive any response for a long time, we can say that it is dead and has no more influence on other posts. However, we can never ensure that a post will not receive any responses in future. Here, we try to define the lifespan of online posts and determine when the posts died.

3.1 Preliminary Study

On social networks, if a post does not receive any response for a long time, we can say that it is dead and has no more influence on other posts. However, we can never ensure that a post will not receive any responses in future. Here, we try to define the lifespan of online posts and determine when the posts died.

Besides, the time interval between two posts can affect their relevance. If there are two posts talking on the same topic but the latter post B is posted a long time after the earlier one A, then the relevance between the two posts should be less, and the influence of A made on B is also weakened. Therefore, before we measure the influence of a post online, we should be concerned about the lifespan of a post. As a preliminary study, we proposed the Comment Arrival Model (CAM) to simulate the process of comments arriving and to determine the death time of a post.

3.1.1 Lifespan of Online Posts

We can find the lifespan of a post when we know the time it is born and dead. The time of birth is just when the post is created. But it is hard to find the time of death for a post,

as we can never be sure that the post will not receive comments any more from that moment. In order to define the death time of a post, we propose to set an expiration time interval when there is no new comment received (arriving) within during this time interval.

Assumption: The event of comment arrival for an online post is stochastic. It means that the arrivals of comments occur independently of the time since the last comment of the post arrived.

Comment frequency distribution: The probability of the number of comments k received in a fixed time interval fulfills Poisson distribution under the above assumption. Then, the probability function of comment frequency distribution is defined as:

$$f(k;\ \lambda)\ = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3.1}$$

where λ is the expected comment frequency of a post. It can be estimated by calculating the average frequency of comment arrival from the data samples.

Average comment frequency: In order to work out the comment frequency of a post, there are 2 parameters, time t from when the article is posted until its last comment received, and the number of comments n received during the time. Initially, a threshold $\xi_0$is set as the post expiration time so that we consider a comment to be the last one if there is no other comment received after it within time ξ_0.Suppose there are N posts $\{P_1, P_2, \ldots, P_N\}$, with their tentative lifespan $\{t_1, t_2, \ldots, t_N\}$ using the initial expiration time ξ_0, and the numbers of comments received within their lifespan are $\{n_1, n_2, \ldots, n_N\}$ respectively. The comment frequency f_i of each post P_i is calculated as: $f_i = n_i / t_i$. The average comment frequency λ is then:

$$\lambda = \frac{\sum_1^N f_i}{N} \tag{3.2}$$

***Inter-comment time distribution*:** Suppose the comment frequencies follow the Poisson distribution, the lengths of the inter-comment time interval would follow the exponential distribution. The probability that the value of inter-comment time is less than or equal to x is:

$$F(x) = 1 - e^{-\lambda x} \tag{3.3}$$

Lifespan of posts: the lifespan of post is dependent on the setting of expiration time interval. We can get the basic expiration time value ξ_b by solving the function with a probability threshold α.

$$\xi_b = -\frac{1}{\lambda} ln(1 - \alpha) \tag{3.4}$$

For example, let α = 0.99, which means 99% chance that the inter-comment time is less than or equal to ξ_b using this model. It is our initial assumption that the event of comment arrival for a post is stochastic. Actually the arrivals of comments on a post are not independent in real life. Therefore we introduced an adjustment parameter to represent external factors when determining the expiration time of posts. After obtaining the basic expiration time interval, factor μ is used to adjust its value before it can finally be applied to determine the lifespan of a post.

$$\xi = \mu * \xi_b \tag{3.5}$$

The adjustment is needed because the proper setting of expiration time interval should be dependent on the actual conditions. The experiments in the next section show good accuracy of the results with the adjustment factor.

It is observed that the usage patterns are in different social networks due to different characteristics. For example, the updating frequency on a local forum may vary a lot in the daytime and night. As for the micro-blogging sites, their posts usually can have responses in a short time, but they may also quickly fade out. We will discuss more about this in the next section when we present our experimental results. After the

expiration time interval ξ of a post is decided, the lifespan of it can be determined as stated before.

3.1.2 Experiment

Sources of data

We collected the data from HK Discussion and Twitter for the experiments. HK Discussion is the most popular forum in Hong Kong, and Twitter, proving the new micro-blogging services, has gained a huge rising popularity recently. HK Discussion is a local forum, and the users are mainly in Hong Kong. On the other hand, Twitter allows its users to build relationships and share information globally. From the data collected, it is found that most of the Twitter users are from China, Hong Kong and US. The date range and location of data collected are given in Table I.

We collected drug-related posts from the two platforms, in either English or Chinese. There are 15 posts with 460 comments in total from HK Discussion and 531 Tweets (Twitter status) with their replies and retweets were retrieved for the experiments. From our observation, a popular Tweet may be retweeted by more than 100 times, but only 100 retweets can be retrieved due to the Twitter API limitation.

Table 1 Source Data Description (CAM)

Platform	**Date Range**	**Location**
HK Discussion	06/2010 – 02/2011	Hong Kong
Twitter	02/2011– 03/2011	China, HK and US

Validity of Comment Arrival Model

Initially, we set the expiration interval ξ_0 long to 30 days, because we want the basic data set to be as complete as possible. The timestamps of each post and their comments

were examined and the average comment frequencies were calculated for the two platforms. After that, we used the comment arrival model to get the basic expiration time intervals. The results are shown in Table 2.

Table 2 Result from Comment Arrival Model

Platform	**Average Comment Frequency (/hour)**	**Basic Expiration Time (hour)**
HK Discussion	$\lambda = 1.20$	$\xi_b = 7.68$
Twitter	$\lambda = 3.57$	$\xi_b = 4.33$

We applied the basic expiration time obtained on the data to test the model performance. Assuming the replies and comments for each post are complete in our dataset, the model is regarded to determine the lifespan of a post correctly if the post has no reply or comment after the basic expiration time ξ_b in the real dataset. The accuracy is then measured by the proportion of number of posts P_{nc} with no more comment received after this interval over the total number of replying posts in this dataset.

$$Accuracy = \frac{C(P_{nc})}{C(P_{all})} \tag{3.6}$$

We found that the model performance is better on Twitter with the accuracy of 90.8%, but only 73.3% on HK Discussion. It is because HK Discussion is a local forum, and its users are living in the same place and the same time zone. The local users browse and post on the forum less frequently during the night. Thus the time period of users' different activeness should be considered to adjust the value of expiration time.

Further analysis has been done on the posting time to get the users' activeness in different hours within a day. The number of posts accumulated for each hour is collected from the HK Discussion forum and is shown in Figure 1. From the numbers and their percentages we can see that users were less active after 02:00 until 10:00 in the morning. Hence, the expiration time should be longer if the post or last comment is made within

this period. The model performance can be improved with the accuracy up to 86.7% when the factor μ is set to 2.86 for the inactive period, and keeps as 1 for other time periods. The value of μ for the inactive period is obtained by the ratio of the average time interval between posts in the inactive period to the average time interval in all times. For those posts of which some comments have been missed, the average percentage of missing comments is only 5.2% when the new factors are used.

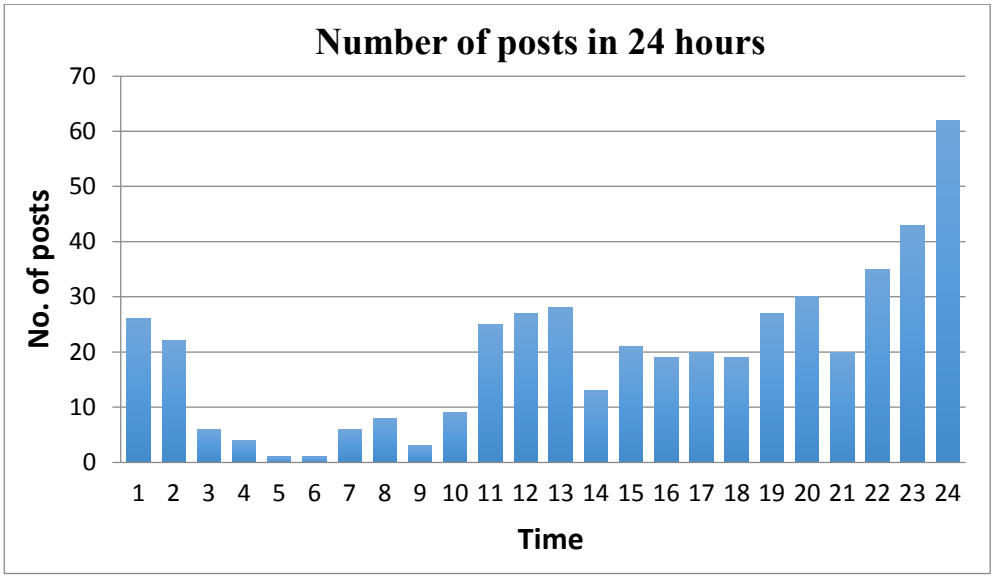

Figure 1 Posting Pattern on HK Discussion

We did a similar experiment on Twitter, and found that the situation is quite different. The boundary of active and inactive periods of users is not clear. It is observed that Twitter users update posts much more frequently, and they seem to be active all the time. It is believed that users are international and most of them use cell phone to post. Consequently, time and location have not been the attributing factors for their posting behavior.

In order to verify the model assumption, a second set of experiments is conducted using a larger data set. We collected more than 1700 tweets from 915 users on Twitter, which are talking on the two topics: “Sichuan Lushan earthquake” and “H7N9 influenza”. The users are mainly from China, Hong Kong and Japan, and the data time is from March to April in 2013.

In the beginning, we set the expiration interval ξ_0 to 30 days as in the first experiment. After applying the comment arrival model, we get the basic expiration time interval ξ_b as 4.90 (h). The model performance is tested by using the basic expiration time to determine the lifespan of posts. The accuracy is 65.5% which is calculated by the equation (3.6).

As the users are from adjacent time zones in this data set, there is also inactive period which is from 01:00 to 08:00 (HKT). This situation is similar to that of HK Discuss in the previous experiment. Hence, we set the value of μ by the ratio of the average time interval in the inactive period to that of all time: $\mu = 1.75$. The value of expiration time ξ is then adjusted to 8.57 (h).The model accuracy is increased to 87.0% with the adjusted expiration time interval. These experiments show that our lifespan model can achieve good accuracy in real cases.

3.2 Graph Model of Online Posts

In order to identify influential posts on social networks and further find their authors as influential users, we propose several methods to measure the influence of online posts. Before we can measure their influence, we first need to figure out the relationship between posts. Usually, relationships between posts are considered as a chain. However, they are more complicated in some cases, such as when a post is replying to this one but its content refers to another. For this purpose, we propose a graph model to analyze those online posts and their relations on the same topic, where nodes represent posts and directed edges represent the relationships between posts.

A graph is defined as $G(V, E)$, where V is the set of posts and E is the set of directed edges which represent the relationships between those posts. Each post $v \in V$ can be described as a tuple of the form (n, t, u, c) where n is the node type, t is the timestamp, u is the author of the post and c is its content. Each directed edge $e \in E$ is represented as $(v_i, v_j, p, w_{i,j})$ where v_i, v_j are nodes and e is an edge directed from v_i to v_j which means v_i is related to v_j, p specifies the type of relationship (either explicit or implicit), and $w_{i,j}$ is the weight of edge in range (0, 1] that measures the strength of their relationship. The relationship is directional and irreciprocal. It is defined that each post can only be related to (point to) earlier posts. Therefore, it is a directed acyclic graph. Also, the graph should be of single edge connection between any two nodes.

3.2.1 Types of Relationship

Explicit relationship: The relationship is given explicitly by the information in data collected from the social media platform, including the relationship of direct reply and some other forms (depending on the functions provided by the social network media,

such as “share” on Facebook, “retweet” on Twitter and “citation” on forums). An relevance score $r_{i,j}$ is assigned to each edge from v_i to v_j, in order to calculate the edge weight (as shown in *Section3.2.4*). The score is set to 1 for all explicit relationships, which means full relevance. For example, $r_{i,j} = 1$ if v_i is a reply or retweet to v_j on Twitter.

Implicit relationship: The relationship connects2 posts that are not directly related but similar and talking on the same topic. In order to identify the content relevancy of two posts, we adopt a method for text similarity measurement. For the implicit relationship, $r_{i,j}$ indicates the degree of content relevancy from v_i to v_j that can be determined by measuring the content similarity score. The score should be in the range (0, 1]. The value of 1 means the two posts are highly similar, so they can be judged as highly relevant or the same as a direct relationship. The conditions of building an implicit relationship can be different depending on the features of the social networks applied on. In general, it is restricted by the time interval between two posts, as their relationship would weaken or dissolve when the time interval exceeds a certain time (called expiration time). For some forums in which only members within a group can see the posts of each other, the user’s identity is also a restriction. For the blogging sites such as Facebook and Twitter, where one’s posts can only be seen by friends or followers, the building of implicit relationships of posts is limited to their authors’ friendship network.

Similar to the concept of reply network defined by Hansen et al [12], we study online posts instead of users. There are two types of reply networks. The direct reply network connects a reply to the post it is directly responding to, while a top level reply network groups all replies (including replies of the reply) within one thread to the starting post. Building implicit relationship among the posts can fill up the missing connections in the two types of reply networks. In a direct reply network, the implicit relationship can indicate the relevancy from the indirect replies to the original post if they are on the same topic; while in a top level reply network they may reveal the conversations that happened between replies. For example, if the situation is B replies to A, C and D reply to B, and

they are all talking on the same topic, their connections in the two networks with the implicit relationship added (in dashed arrow) are shown in Figure 2.

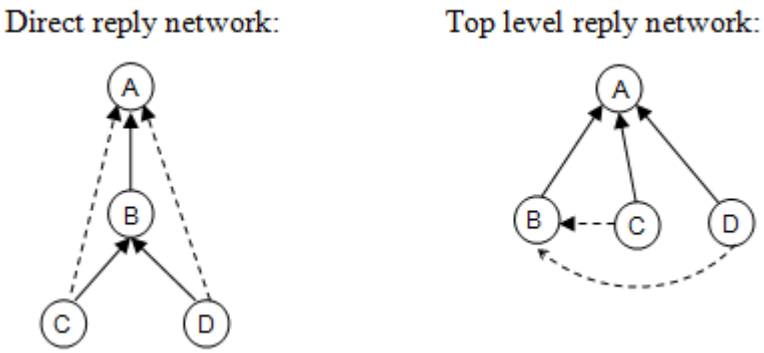

Figure 2 Types of Reply Network

3.2.2 Types of Posts

The type of a post is determined by the role it plays. In our work, each post would be characterized in four types: root, follower, starter and connecter. Among them, starters are certainly considered to be influential. Many researchers have tried to identify starters in a network. As for connecters, they are considered as bridges that connect two or more peaks in centrality analysis [22]. Also, a bridge node is also important in a network if it connects starters. Therefore, a connecter is considered influential. It should be noticed that in our definitions, follower, starter and connecter are referring to the type of posts, not of users.

Root: It is the first post discussing a topic or a subtopic within a certain period, so it is not related to any others. In the graph, roots are the nodes who are not pointing to others (with no out-degree).

Follower: It is a response (e.g. reply, comment or share) of a post or a new post talking on the same topic as another post before, which means it is explicitly or implicitly related to others. In the graph, followers are the nodes who are pointing to others (with some out-degree).

Starter: It is identified when it received a large number of explicit or implicit responses (followers); meanwhile a starter is not following many others. In a graph, conversation starters are the nodes who point to a few but be pointed by many others, i.e., they are of high in-degree and low out-degree. Moreover, it is better for a starter to have followers also being followed by many others. In a graph, this situation can be observed as having a sub-tree rooted at a follower with its descendant nodes of high in-degree also.

Connecter: It connects two or more starters as a bridge. It will have a big impact on the message transmission if without this node.

It should be noticed that a post may play multi-roles at the same time. It is also possible that the roles of posts can change over time. For example, a follower may become a starter after a period of time, and later may also be a connecter; meanwhile it is still a follower of others. The details of the identification of the node types will be discussed in *Section 3.4*.

3.2.3 Content Similarity

It is well known that some popular social networking sites would have their posting contents usually of short text, such as Facebook and Twitter (the trend of micro-blogging). Here, we adopted the method of keyword matching using Tanimoto coefficient to measure the content similarity of two posts. The Tanimoto similarity measure of weighted vector [1] is described next.

Consider a set of vectors of the form $X_i = (x_{i1}, x_{i2}, \ldots, x_{iN})$, where x_{ik} is either 0 or w_k, a positive weight assigned to the kth entry. The Tanimoto coefficient for a pair of such vectors, X_m and X_n, is:

$$T_{mn} = \frac{X_{mn}}{X_{mm} + X_{nn} - X_{mn}} \tag{3.7}$$

Where $X_{ij} = X_i \cdot X_j$. The value of T_{mn} ranges from 0 to 1.

In order to apply it in our model, a glossary of keywords should be built to define the topic. Consider K is the set of keywords, and each keyword $k \in K$ is defined with a set of synonyms $Z = \{z_1, z_2, \ldots, z_n\}$ and their association scores $S(z_j)$ to k in the range (0, 1). For each post v_i, it can be represented by:

$$\vec{v}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,N})$$

where the vector entries describe the presence of those keywords. $x_{i,k} = 1$ indicates the presence of the kth keyword, and if the keyword is absent but some of its synonyms are found, the entry value is calculated by $\text{Max}(S(z_j))$. Let $x_{i,k} = 0$ when neither the keyword nor its synonyms are found in the text. For example, let "Japan" "Earthquake" and "Tsunami" together define a topic. "Fukushima" can be considered as a synonym of "Japan" here while "disaster" is the synonym of both "Earthquake" and "Tsunami" as it is their hypernym. Therefore the content relevance of posts v_i and v_j can be measured by the Tanimoto coefficient:

$$r_{i,j} = T(\vec{v}_i, \vec{v}_j) = \frac{\vec{v}_i \cdot \vec{v}_j}{\left|\vec{v}_i\right|^2 + \left|\vec{v}_j\right|^2 - \vec{v}_i \cdot \vec{v}_j} \tag{3.8}$$

The keyword dictionary for content analysis cannot handle the out-of-vocabulary words as there is no dynamic updating in the current dictionary. On the other hand, micro-blogging social networks become highly popular in recent years. Their limitation of text length for each post makes users to create new words for short, which induces a

lot of out-of-vocabulary words. To address this problem, we propose to develop a dynamic dictionary where new words can be learned from the posts retrieved. Initially, a user defines a basic dictionary of keywords for the topic. A post is considered to be relevant to this topic if it contains at least one of the defined keywords. Then a new word could be recognized by counting its document frequency, which is the number of posts retrieved on the topic containing this word. Meanwhile, a word cannot be used to describe the topic if it is also frequently mentioned in other topics. Therefore, a new word could be detected and added into the dictionary when it has high document frequency on the specified topic, and relatively low document frequency on all topics. In order to determine the new words, a method is proposed to suggest the high-ranking words according to their Importance Score. Based on the new word suggestion, it will be manually decided which one could be inserted to the dictionary finally. The algorithm of the Importance Score calculation (DICT_IS_Calculation) is presented below.

1. Define a set of keywords K relevant to the topic T_i as the basic dictionary.
2. Retrieve the posts $V(T_i)$ that contain any keywords in K. These posts are considered to be relevant to T_i.
3. For each word w in the post $v \in V(T_i)$, the number of posts in $V(T_i)$ that contains w is counted as $df_i(w)$.
4. Suppose the set of posts on all topics is V, the total number of posts is $|V|$. For each word w in the post $v \in V(T_i)$, the number of posts in V that contains w is counted as $df(w)$.
5. The Importance Score of w on topic T_i can be calculated as:

$$IS_i(w) = df_i(w) \cdot \log \frac{|V|}{df(w)} \tag{3.9}$$

3.2.4 Edge Weight

The weight assigned to an edge is the degree of relevance between two posts and high weight edges indicate strong relationships. Edge weight is measured by two factors: the content relevance and the time interval between posts.

$$w_{i,j} = \alpha_T * r_{i,j} \tag{3.10}$$

$\alpha_T$is a factor used to diminish the relevance degree. Suppose the expiry time is ET and the maximum time is MAX. $\alpha_T = 1$ within [0,ET] and linearly decreases until it finally drops to 0 at MAX as shown in Figure 3. The method for finding expiry time ET has been introduced in *Section 3.1*. MAX can be the longest time period between a post and its replies in the dataset.

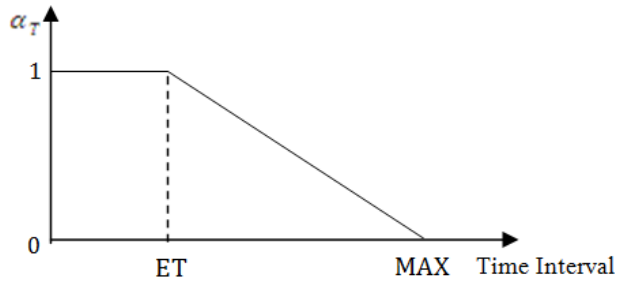

Figure 3 α_T Value Changes Over the Time

An example is shown in Figure 4, where *T* is the time interval between the root post and its first reply.

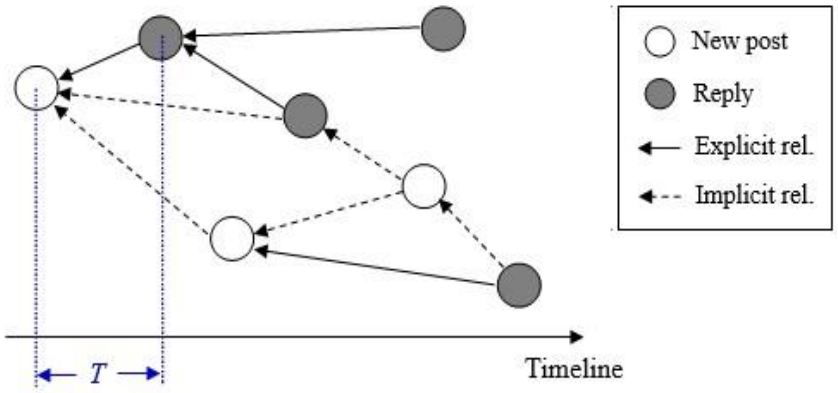

Figure 4 A Post Graph with the Timeline

3.3 Graph Transformation for Data Cleaning

It is commonly found that some responses of a post are just written by the posting author. For example, users often add their own comments after sharing a post on Facebook. On Twitter, users may continually talk about the same topic in several messages due to the word length limit of a post. Those follow-up posts are represented as child nodes of the first post in our graph model, while they are actually just an extension or supplement to the original one. Besides, some edges with very small values of weight indicate that the relationship between the two connected nodes is very weak. In order to increase the computation efficiency of measures on the graph, the edges with its weight below a threshold can be removed. Also, in some social networks such as Facebook and Twitter, it is quite common that replies become a conversation between 2 or 3 friends. It may induce a long linear branch in the graph. These nodes are considered as a kind of noisy data. Therefore, after an initial graph is constructed and before it can be analyzed, we need to perform some transformation for cleaning purpose.

3.3.1 Merge Consecutive Nodes of Same User

For the followers of a post that are just written by the original post author within a short time, they are identified and merged with their parents to form a single node so that the influence measurement of nodes can reflect a more accurate situation. Their contents will be added together as the content of the new merged node. The graph transformation is performed based on the algorithm below.

Suppose there are two nodes, $v_a = (n_a, t_a, u_a, c_a)$, $v_b = (n_b, t_b, u_b, c_b)$, and the edge from v_a to v_b is $e = (v_a, v_b, p, w_{a,b})$.

- If $u_a = u_b$, and $w_{a,b} > \delta$ then
  - Replace v_a and v_b with the new node $v_{new} = (n_{new}, t_b, u_b, c_a + c_b)$ where the new content is the combination of c_a and c_b.The new type n_{new} will be determined after all the transformations are done.
  - Remove the edge between v_a and v_b: $(v_a, v_b, p, w_{a,b})$
  - Update the edges pointing to v_a or v_b: $(v_i, v_a, p, w_{i,a})$, $(v_i, v_b, p, w_{i,b})$ change to $(v_i, v_{new}, p, w_{i,new})$
  - Update the edges pointed by v_a or v_b: $(v_a, v_k, p, w_{a,k})$, $(v_b, v_k, p, w_{b,k})$ change to $(v_{new}, v_k, p, w_{new,k})$

For the transformation of related edges, with attributes of relationship type p, the weight of implicit relationship should be re-calculated as the vectors of merged nodes are also changed, which will affect their similarity scores for implicit relationships. The construction of merged node vector is shown as follows.

Assume the nodes to be merged are v_a and v_b with their vectors:

$$\vec{v}_a = (x_{a,1}, x_{a,2}, \ldots, x_{a,N}), \quad \vec{v}_b = (x_{b,1}, x_{b,2}, \ldots, x_{b,N})$$

The new vector $\vec{v}_{new}$ should take their maximum value as its entries:

$$\vec{v}_{new} = (\max(x_{a,1}, x_{b,1}), \max(x_{a,2}, x_{b,2}), \ldots, \max(x_{a,N}, x_{b,N}))$$

The value of each entry in the new vector is calculated using the max() function, because the new text is merged as the union of words from the two posts. The word frequency is not considered here.

3.3.2 Remove Low-weight Edges

For any edge with a small weight value in a graph, it represents a weak relationship between the two nodes. It may be due to the time interval between the two posts is very long, or the similarity of their contents is very low. We set a threshold to determine which relationships are too weak, and remove those edges between nodes, that is, remove $e = (v_a, v_b, p, w_{a,b})$ if $w_{a,b} < \varepsilon$, where ε is a very small value. The threshold value can be set in an empirical way. For example, a user can collect a set of posts and judge their relevance manually. Initially, the threshold value can be set as 0.1, then its value can be adjusted based on 2 simple rules. Based on post relevance judged by the user, the value should be reduced if some edges between relevant posts are removed after applying it. On the other hand, the value should be increased if many edges are still linking non-relevant posts. The final threshold value should be the one with best precision in judging the relevance between nodes. The other way is to obtain the value by some standard statistical measures, such as the lower bound of 95% confidence interval in the edge weight distribution.

After the removal of low-weight edges, some nodes may not point to any others, and become new roots. The possible reason for them being a root is that it has been a long time since the topic was talked about last time, or it is the first post to talk on this topic. Some nodes may become isolated when their edges are all removed. Those isolated nodes can also be removed as they are already disconnected from the graph.

3.3.3 Group Nodes in Linear Patterns

For the cases that the replies become a conversation between several close friends, there would be long discussion chains that only involve a few people but will reduce linear branches in the post graph. It is suggested to group these consecutive nodes in the

repeating patterns in order to simplify the graph. Spam posts may also induce nodes in linear repeating patterns. They are a kind of noisy data. Grouping them can also help to reduce the computation complexity of the graph and increase the efficiency of node influence measures.

In our work, this merging operation is performed only on a linear branch, which are the nodes connected in a chain with no other branch. Once the nodes have been labeled with author identifications, we can detect the frequent linear patterns. Taking an example as shown in Figure 5, the pattern of "B replies to A" happens two times. Therefore we group the two nodes and generate a new node to represent them.

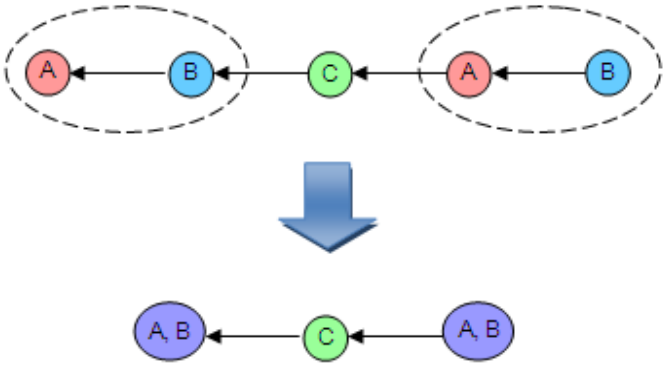

Figure 5 Merging Nodes of Linear Patterns

3.4 Influence Measurements

3.4.1 Degree Measure

As mentioned in previous sections, the degree of a node can be used to identify starters. Since a starter is supposed to have many followers and it is not a follower of many others, we first compute the difference between the in-degree and out-degree of each node. The in-degree of a node v is denoted as $deg^{+}(v)$ which is the sum of weight of the

incoming edges incident to the node v, and the out-degree $deg^-(v)$ is the sum of weight of its outgoing edges. The difference $d(v)$ is measured as one factor [32]:

$$d(v) = deg^+(v) - deg^-(v) \tag{3.11}$$

Another factor is the weighted average of the in-degrees of its follower $v_j \in Fol(v_i)$ to reflect the popularity of its followers:

$$\begin{aligned} s(v_i) &= \frac{\sum_{v_j \in Fol(v_i)} w_{i,j} \cdot deg^+(v_j)}{\sum_{v_j \in Fol(v_i)} w_{i,j}} \\ &= \frac{\sum_{v_j \in Fol(v_i)} w_{i,j} \cdot deg^+(v_j)}{deg^+(v_i)} \end{aligned} \tag{3.12}$$

Then we can identify a node $v_i \in V$ as a starter when both $d(v_i)$ and $s(v_i)$ reach the preset thresholds (σ_1 and σ_2):

$$d(v_i) \geq \sigma_1 \wedge s(v_i) \geq \sigma_2$$

3.4.2 Shortest-Path Cost Measure

The basic idea of this method is to judge a node's influence by measuring how many other nodes would be affected and how much impact would be if the target node is removed from the graph. It should be noted that in a graph the relationship edges are built from later posts to earlier ones; conversely the influences traverse in reverse directions from earlier posts to later ones.

In our definition, a post should have influence on its followers, as the followers are responses (e.g. replies, citation and share) that are somehow activated by the original post (followee). These followers may also have influence on their own followers. As a result, a post may have indirect influences on its followers' followers, and so on. In a

graph $G(V, E)$, the descendant set $Des(v)$ of a node $v \in V$ includes its followers directly pointing to it and other descendants that can reach it through paths. For every $v_d \in Des(v)$, there is at least one directed path from v_d to v in the graph.

If the path from node v_d to v_n is $(v_d, v_{d+1}, v_{d+2}, \ldots, v_n)$, the *relationship strength* from v_d to v_n can be measured as the accumulative weight:

$$W(v_d, v_n) = \prod_{i=d}^{n-1} w_{i,i+1} \tag{3.13}$$

where v_i is pointing to v_{i+1} and $w_{i,i+1}$ is their edge weight. If more than one path from v_d to v_n exist, the maximum accumulative weight is taken as their relationship strength value. By doing this, the value of weight between any two nodes can be constrained in the range (0, 1). The reason not to do summation and normalization of $w_{i,i+1}$ is that it will induce new weights with too small variance, which is difficult to differentiate afterwards. On the other hand, the ancestor set $Anc(v_d)$ of a node v_d is defined accordingly: $v_a \in Anc(v_d)$ when $v_d \in Des(v_a)$.

The algorithm of finding ancestors is similar to finding the shortest path with respect to the cost between nodes in a graph, except that we calculate the path cost as the product of the weights instead of the sum. It is assumed that each node would have influence on its descendants in the graph. To measure the influence of a node, we remove it from the graph and capture the change of path cost between these descendant nodes and their ancestors. The path cost $c(v_d)$ of a node v_d to its ancestors $v_a \in Anc(v_d)$ is the average of their relationship strength value:

$$c(v_d) = \frac{1}{\left|Anc(v_d)\right|} \sum_{v_a \in Anc(v_d)} W(v_d, v_a) \tag{3.14}$$

Here, we take the average in order to do normalization for the nodes in later time, because posts afterwards may have more ancestors. When a node v_i is removed from a

graph, its adjacent edges are also removed. Its descendants $v_d \in Des(v_i)$ may be disconnected from some of their original ancestors. Even if they can reach their ancestors through other paths, their relationship strength may be weakened if the removed node is on their shortest path. Suppose v_i is on the path with the shortest cost between v_d and its ancestor $v_a \in Anc(v_d)$. After v_i is removed, a new path should be found with the new relationship strength value W' and $W'(v_d, v_a) \leq W(v_d, v_a)$. If no path can be traced between v_d to v_a, it means v_d is disconnected from v_a, and their relationship strength will be set to 0 ($W'(v_d, v_a) = 0$). If v_i is not on that path, the relationship strength between v_d andv_a will not be changed: $W'(v_d, v_a) = W(v_d, v_a)$.

Let $C(v_d, G, v_i)$ be the average shortest-path cost between the node v_d and its ancestors after removing v_i from the graph G. The influence of a node $v_i \in V$, $Inf_c(v_i)$, in the graph is then:

$$Inf_c(v_i) = \sum_{v_d \in Des(v_i)} (C(v_d, G, \varnothing) - C(v_d, G, v_i)) \tag{3.15}$$

For the nodes who do not have descendants ($Des(v_i)=\emptyset$), their influence will be set to 0 as they have no one to be influenced. The pseudo code of SP_Influence_Measure is shown Figure 6.

```
SP_Influence_Measure(G, v_i)
Input: a graph G(V,E), a node v_i ∈ V
Output: the influence score Inf_c(v_i) of node v_i
for each v_i's descendant v_d in G
    for each v_d's ancestor v_a in G
        c(v_d) := c(v_d) + W(v_d, v_a)
    c(v_d) := c(v_d)/|v_a|     /*avg path cost of v_d to
                                 its ancestors*/
    G' := remove node v_i and related edges from G
    for each v_d's ancestor v_a' in G'
        c_new(v_d) := c_new(v_d) + W(v_d, v_a')
    c_new(v_d) := c_new(v_d)/|v_a'|   /*new avg path cost of v_d
                                        after removing v_i*/
    d := c_new(v_d) - c(v_d)      /*the diff between new and
                                    original path cost*/
    Inf_c(v_i) := Inf_c(v_i) + d   /*sum up the diff*/
Obtain the influence score Inf_c(v_i) of node v_i
```

Figure 6 SP_Influence_Measure Algorithm

Comparing to the degree measurement in *Section 3.4.1*, our method considers multi-level relationship between posts, even if they are not on the same path. For example, as shown in Figure 7 (explicit relationship denoted by solid arrow and implicit relationship denoted by virtual arrow). Suppose node A is removed to see its influence on B and C. Then, B will be disconnected from any other nodes, while C can be still connected to D. Hence, A has a larger influence to B than to C. In this case, the calculation of the influence measure of node A also includes the relationship between C and D, which is not considered in the degree measurement. Another advantage is the avoidance of duplicate counting on node E when measuring the influence of node A in multi-levels.

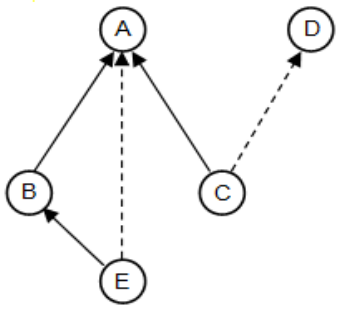

Figure 7 An Example Graph of Related Posts

3.4.3 Graph Entropy Measure

Based on the graph model proposed, a graph can be considered as an ordered network with the node types of root, follower, starter and connecter defined. Shetty and Adibi [23] showed their success in finding important nodes through graph entropy in an ordered network. The graph entropy can be defined differently for various problems and we adopted a similar approach as in Dehme [31]. In their work, the entropy of a network is defined by using the local information graph, where metrical graph properties are used for defining information functional of each vertex. The graph entropy measure is better choice because it considers the factors of the entropy of a node, as well as the remnant graph entropy when the node is removed, which captures both the local influence for identifying starters, and the impact to the whole graph for discovering connecters.

Consider a graph with arbitrary node labels. In order to determine the probability value for each node for calculating the graph entropy, we first define the local vertex functional. Generally, the information functional is used to quantify structural information based on a given probability distribution. In our case, information functional is defined as the centrality of nodes.

For the graph $G = (V, E)$ where $v_i \in V$, graph entropy is defined by:

$$E(G,P)=\sum_{i=1}^{|V|} p(v_i)\log(1/\,p(v_i)) \tag{3.16}$$

The probability for each node is defined as:

$$p(v_i)=\frac{f(v_i)}{\sum_{j=1}^{|V|} f(v_j)} \tag{3.17}$$

f represents an arbitrary information functional. Unlike traditional centrality measurement, such as closeness centrality, betweenness centrality and eigenvector centrality, in our model the centrality of a node only looks at the nodes that point to it or can be reached through paths. Recalling the term of "descendant" that is defined in last section, a node's descendants is used to measure its distance-weighted centrality.

$$f(v_i)=\sum_{v_d\in Des(v_i)}\frac{1}{d(v_d,v_i)} \tag{3.18}$$

$d(v_d, v_i)$ is the distance between the node v_i and its descendant v_d. If there is an edge that directly links to them, their distance can be calculated as the reciprocal of the edge weight.

$$d(v_d,v_i)=\frac{1}{w(v_d,v_i)} \tag{3.19}$$

Otherwise, if v_d can reach v_i through a path (v_d,v_{d+1}, v_{d+2},…,v_i), then the distance between v_d to v_i will be the sum of edge distance along the path. For the case of more than one path exists, the shortest path distance will be taken.

The steps of measuring node influence through graph entropy are shown below.

1. Compute the entropy of each node v_i as:

$$E(i)=-p(v_i)\cdot\log(p(v_i)) \tag{3.20}$$

2. Remove v_i and its edges from the graph

3. Calculate the entropy of remaining graph as $EN(i)$
4. Calculate the influence of node v_i as:

$$Inf_e(v_i) = \frac{EN(i)}{\log(EN(i) / E(i))} \tag{3.21}$$

The formula (3.21) is referred from [23], which proved to be able to identify important nodes in the network built of Enron (company) emails. We adopt it to measure the influence of node v_i by $E(i)$ and $EN(i)$, and try to find nodes which have higher centrality and caused more difference in the graph after they are removed from the graph..

3.5 Identify Influential Posts

In order to find the influential nodes, we ranked the nodes based on their influence scores from different measurements. Starters and connecters can be identified first as the preliminary result. Starters are determined by degree measure, and connecters are identified by using the other two methods. In our work, a connecter should fulfill two conditions: (i) Have a higher rank in the measurements of shortest-path cost or graph entropy; and (ii) Connect two starters by different authors.

As we have defined influence from the aspects of starter and connecter, the influential nodes are either starters or connecters. Based on the results from the three measurements in the previous sections, we are able to determine the most influential posts. The starters or connecters are determined not influential if they are ranked low in all the measurements. The others are considered as influential posts, and their authors are considered as potential influential users.

3.6 Analyze Sentimental Influence

After the influential posts are identified, we are further interested in analyzing their sentimental influences. We can judge the sentimental influence of a post through the sentiment of the post and its followers. For example, the sentiment of influencer A is positive, and that of influencee (the one being influenced) B is also positive, as they are talking on the same topic, it can be considered that B complies with A. In other words, the sentimental influence of A on B is compliance. Otherwise, if the sentiment of B is negative while A is positive, we can say that B opposes A, and the sentimental influence is opposition.

The sentiment and sentimental influence of posts can be used to find interesting patterns. Barbagallo [11] and his colleagues did an empirical study on tweets in the tourism domain and found that negative tweets are more retweeted. However, they did not research on the sentimental influence type, so we don't know whether the negative tweets mainly get compliance or opposition. The differentiation of compliance and opposition influence would be useful in marketing. If a marketer wants to promote a new product, it could be efficient if he starts from someone more influential first. More importantly, the main influence of this person should not be negative. Of course, the more complied audience he or she has, the better it would be for the product promotion. Besides, revealing sentimental influence of a user can help to detect the Internet abuse. With the popularity of social networks in youngsters, there are increasingly incidents of Internet abuse. Some people were maliciously attacked or humiliated on social networks and suffered a lot from this. By analyzing the sentimental influence towards a user can help to detect the abuse earlier so that actions can be taken before it causes bigger harm to the user.

Based on our proposed graph model of online posts, the attributes of post sentiment and the sentimental influence can be added into the node and edge respectively. Besides, the sentimental influence is estimated between the influencer and the influencees that are not connected through edges.

3.6.1 Post Sentiment

The sentiment analysis on a post is different from the traditional analysis on articles, because some of the wordings used in social networks are quite different from normal text. People often like to post short text online, using abbreviations and special combination of symbols and letters to express their emotion. The contents often do not follow proper grammar. Hence, we would like to propose a new method to estimate the sentiment of posts.

The attribute of post sentiment is defined as PS_i for the post v_i. Its value can be positive (POS), negative (NEG) or neutral (NEU), which will be judged from the post content c_i. Emoticons are frequently used in online posts. Many social media platforms have their own set of emoticons with symbols, such as **:)** and >_<, which are highly used. Emoticons can express the users' feelings directly with less ambiguity. Therefore we preferred to make use of emoticons to determine the post sentiment at first. K.M. Ip [27] summarized the emoticons from Sina Weibo, Facebook and Yahoo blog. Based on his work, the list of emoticons has been extended and the emoticons are classified into positive and negative categories (see appendix). For those posts that do not contain emoticons, and those with the same number of positive and negative emoticons, we can utilize some existing sentiment word dictionary, such as SentiWordNet [37] and SCWS [38], and apply the keyword-based method to judge the sentiment. The steps to determine the sentiment of posts is shown as follows.

1. Calculate the difference between the number of positive and negative emoticons in the post content:

$$PN = |Pos\ emoticons\ in\ c_i| - |Neg\ emoticons\ in\ c_i|$$

2. Judge the post sentiment PS_i by:

PS_i = POS if $PN > 0$;

PS_i = NEG if $PN < 0$.

3. For the posts with no sentiment assigned, calculate the difference between the number of matched positive and negative words in the post:

$$PN = |Pos\ words\ in\ c_i| - |Neg\ words\ in\ c_i|$$

And judge the sentiment the same as Step 2.

4. Other posts are considered with neutral sentiment:

PS_i = NEU if $PN = 0$.

3.6.2 Sentimental Influence

After the post sentiment is determined, we are able to find the sentimental influence between posts. The attribute of sentimental influence is defined as $SI_{i,j}$ and can be added to the edge $e_{i,j}$ from post v_i to v_j. Its value can be compliance (CPL) or opposition (OPP), which means v_i complies with or opposes v_j.

Suppose post v_i is an influencer, its descendants in the post graph can be considered to be v_i's influencees. Besides the nodes that are connected to the influencer through edges, we also need to estimate its sentimental influence on other influencees. In Heider's Balance Theory [16], it is said: "my friend's friend is my friend; my friend's enemy is my enemy; my enemy's friend is my enemy; my enemy's enemy is my friend." Based on this theory we can estimate the sentimental influence of an influencer on its influencees following the principles below:

i. If A complies with B, B complies with C, then A complies with C.
ii. If A complies with B, B opposes C, then A opposes C.
iii. If A opposes B, B complies with C, then A opposes C.
iv. If A opposes B, B opposes C, then A complies with C.

Specially, as influencees, the posts with neutral sentiment are defined to be neutral (neither complying nor opposing) to others, but neutral influencers are considered to be complied by followers with positive sentiment and to be opposed by negative followers. Considering these cases, the principles are modified as follows. The new principles are suitable for neutral sentiment as well.

i. If A complies with B, B complies with C, then A complies with C.
ii. If A complies with B, B opposes C, then A opposes C.
iii. If A opposes B, B complies with C, then A opposes C.
iv. If A opposes B, B opposes C, then A complies with C.
v. If A complies with or opposes B, B is neutral to C, then A complies with C if they have the same sentiment, and A opposes C if their sentiment is different.
vi. If A is neutral to B, then A is also neutral to C whatever B is to C.

After applying these principles, the sentimental influence $SI_{i,j}$ from influencer v_i to influencee v_j can be judged as compliance if they have the same sentiment (except neutral), or v_i is neutral and v_j is positive. The sentimental influence is considered to be opposition when the sentiment is opposite, or v_i is neutral and v_j is negative. The conditional equation is shown as follows:

$$SI_{i,j} = \text{CPL if} \begin{cases} PS_i = PS_j, PS_i \neq \text{NEU}, PS_j \neq \text{NEU}; \\ PS_i = \text{NEU}, PS_j = \text{POS}. \end{cases}$$

$$SI_{i,j} = \text{OPP if } \begin{cases} PS_i \neq PS_j, PS_i \neq \text{NEU}, PS_j \neq \text{NEU}; \\ PS_i = \text{NEU}, PS_j = \text{NEG}. \end{cases}$$

$$SI_{i,j} = \text{NEU if } PS_j = \text{NEU}. \tag{3.22}$$

Chapter 4 User Graph Model

Although the influential starters and connecters are identified from the post graph, we still have the problem on determining the influential users. Consider the cases that (i) for a starter many of its followers are actually from a small group of users (one user can reply several times); or (ii) a connecter links with two starters who have a large set of common follower users. In these cases the influence may be wrongly judged in the post graph model. Therefore we proposed the user graph model to refine the influence measures of potential influential users. The comparison of the results from two models is discussed in the case study in *Section 5.2*.

A user graph can be converted from the post graph. The reason we do not build the user graph directly is that we would like to keep the information of relationships among a group of posts, such as discussion threads and discussion chains (as illustrated in *Section 4.2.1*), rather than just the replying relationship between two people. However, we are not going to build a complete user graph from the post graph due to high computational complexity. As it is natural to consider users who have made influential posts, we select the starters and connecters in the post graph as seeds, then look at their neighbors and finally find possible connections between distant starters. The definition of user graph model is given in *Section 4.1* and the process of converting a post graph to the corresponding user graph will be discussed afterwards.

4.1 Definition

The user graph is defined as $G_u(U, E_u)$, where U is the set of users, E_u is the set of directed edges which represent the relationship between users. Each node $u_k \in U$ is the

author of post v_i^k in the Post Graph G_v.

Node types: There are three node types defined in a user graph: starter, connecter and follower. Each node can belong to one or more types. At first, the type of a user is the summation of types of his posts. For example, starter users are the authors of starter posts identified in the post graph. Yet the author of a connecter in a post graph may no longer play the same role in the user graph. On the contrary, some new nodes could be detected as connecters in a user graph, even though none of their posts connect two starters in the corresponding post graph. Therefore the type of connecter will be determined after the graph conversion and measurement.

Edge types: $e(u_k, u_j) \in E_u$ is the edge directed from u_k to u_j representing that u_k is related to u_j, which means u_k has replied or responded to u_j either explicitly or implicitly (as defined in the post graph model). Besides, there are another type of *virtual edge e'*(s_k, s_j) defined between two starters, to represent the path from s_k to s_j. The virtual edges are built when there is at least one directed path between two starters, and their distance is very long. In this case, we will keep the shortest path length as the weight of virtual edge. The nodes on the paths are not important so it is not necessary to show them in the user graph.

Edge weight: $w(u_k, u_j)$ is the weight of edge $e(u_k, u_j)$ that measures the strength of their relationship. It is affected by the number of interactions and the relevancy of their dialogs. For virtual edges linking two starters, the edge weight $w'(s_k, s_j)$ is calculated as the shortest path length from s_k to s_j in user graph as described before.

4.2 Graph Conversion and Measures

In order to capture the influence of users, we convert a post graph to a user graph. Considering the complexity and cost effectiveness, instead of processing the complete graph, we use a biased sampling method starting with potential influential users, who have posts as starters or connecters identified in the post graph. Then we propose several measurements to capture the influences of u-starters and u-connecters in different respects. The terms "**u-starter**" and "**u-connecter**" are used to refer to the starters and connecters in user graph model. For the rest of the section, we discuss the details of how to convert a post graph to a user graph and measure the influences of u-starters and u-connecters. Figure 8 shows the overall flow of the operations and measurements on user graph.

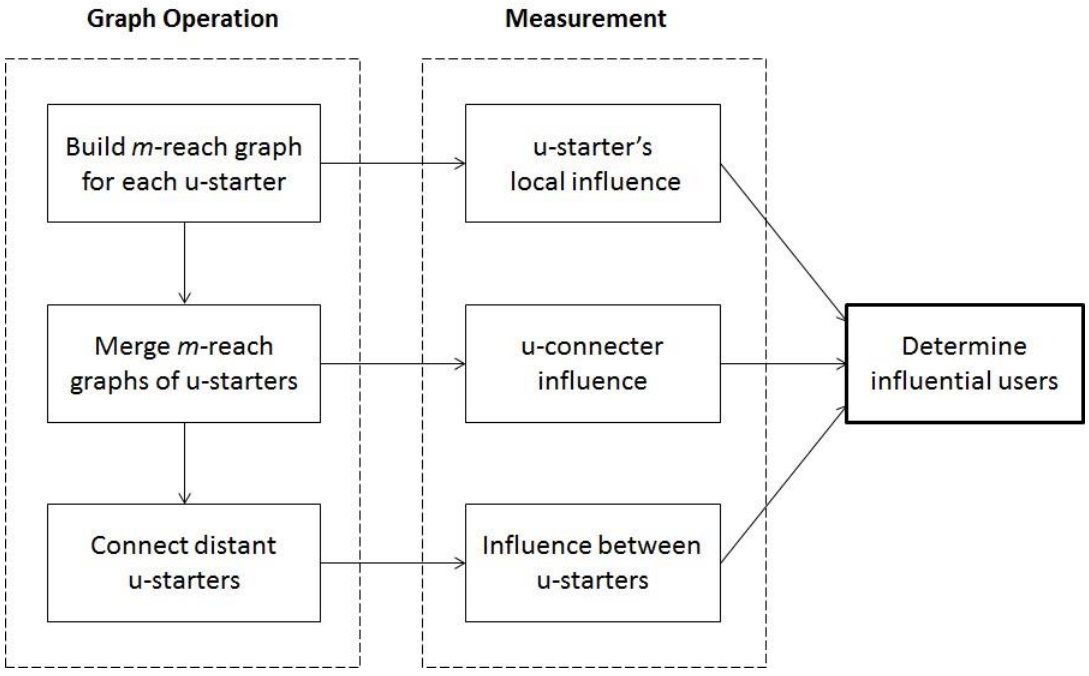

Figure 8 Workflow of Graph Operation and Measurement

4.2.1 Building *m*-reach graph for each u-starter

In the post graph model, a starter is observed when it has obtained a large amount of followers and descendants. However, it is hard to measure its influence on users, because one user may write a number of posts, or reply several times within a discussion. Moreover, if a user has several posts as starters, it is necessary to consolidate all the followers and descendants in terms of users. For this reason, we need the conversion from the post graph to a user graph where each user is represented as one node. When the user graph is directly built for all discussions from different u-starters, some of their descendants will be merged and their influence may not be accurately judged. Therefore, we first built an m-reach user graph for each u-starter in order to capture its local influence.

M-reach graph

"*M*-reach" is a measure defined by Borgatti [36] that counts the number of unique nodes reached by a given node in *m* links or less. In our user graph, $g^m(u_k)$ is u_k's *m*-reach graph which consists of nodes that can reach u_k via a path of length *m* or less. Here the path length is defined as the number of hops to go through without consideration of edge weights.

Discussion thread and discussion chain

In a post graph, a starter together with its descendants forms a discussion thread. In the Post-reply Opinion Graph by Stavrianou et al. [4], they defined the discussion chain which is different from discussion thread: "The discussion chains consist of the paths in the graph whose starting node is a root and ending node is a leaf when we inverse the direction of the edges." In Figure 8, a post-reply graph shows the difference between discussion chains and threads. These definitions are used in the following paragraphs.

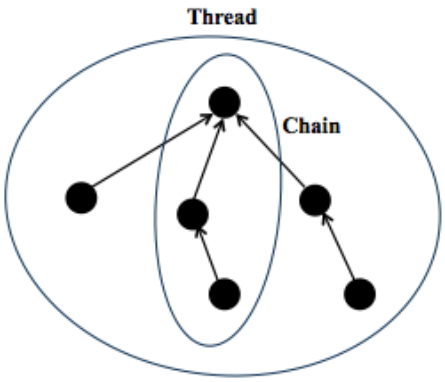

Figure 9 Discussion Threads and Discussion Chains

Algorithm of building m-reach graph

Suppose the set of starters found in post graph $G_v(V, E)$ is $S \subset V$, each starter (post) $s_i^k \in S$ has an author u_k, then u_k is a u-starter. An m-reach user graph $g^m(u_k)$ will be built for each u-starter u_k. For each starter s_i^k whose author is user u_k, the discussion thread in post graph will be converted to user graph $g^m(u_k)$. Here, the value of m will be determined during the experiment.

In order to keep the distance information (as defined in *Section 3.4.3*) from the starter to its descendants in a discussion chain, depth-first search (DFS) starting from s_i^k is conducted in the post graph G_v. For each descendant v_a^x of s_i^k (with authors u_x and u_k respectively), the shortest distance between v_a^x and s_i^k is notated as $d_a^{(x, k)}$.

The path length is defined differently for "m-reach". That is, the path length from v_a^x to s_i^k is the minimum number of distinct users on the path for v_a^x to reach the starter s_i^k. It is represented as $m_a^{(x, k)}$, and used to control the depth of searching. Suppose the value of m is given as m_0, the MR_DFS algorithm is shown in Figure 10.

```
1 for each starter s_i^k by user u_k
2     label s_i^k as visited, set m_i^(k,k) to 0
3     let S be a stack
4     S.push(s_i^k)
5     while S is not empty
6         v_a^x := S.top()
7         for each v_a^x 's unvisited follower v_b^y in G_v
8             label v_b^y as visited
9             if there is a visited node v_o^y with author u_y
10                m_b^(y,k) := m_o^(y,k)
11            else
12                m_b^(y,k) := m_a^(x,k) + 1
13            if m_b^(y,k) <= m_0
14                update g^m(u_k) with node v_b^y and edge e(v_b^y, v_a^x)
15                S.push(v_b^y)
16            continue at 5
          /*Reset the node v_a^x as unvisited after all its followers
            are visited, so that it can be visited in other path*/
17        delete m_a^(x,k) and label v_a^x as unvisited
18        S.pop()
```

Figure 10 Algorithm of MR_DFS

```
1 add node u_k with type (starter) in g^m(u_k)
2 for each node v_b^y and edge e(v_b^y, v_a^x) obtained from DFS in G_v
3     if v_b^y is not visited
4         if there is no user node u_y in g^m(u_k)
5             add a new node u_y in g^m(u_k)
6         add v_b^y 's type in u_y's type
7     if e(v_b^y, v_a^x) is not visited
8         if there is no edge from u_y to u_x in g^m(u_k)
9             build the edge e(u_y, u_x)
10            w(u_y, u_x) := w(v_b^y, v_a^x)   /*initialize edge weight*/
11        if there is an edge e(u_y, u_x) in g^m(u_k)
12            w(u_y, u_x) := w(u_y, u_x) + w(v_b^y, v_a^x) /*update edge weight*/
```

Figure 11 Algorithm of MR_Build_Graph

The *m*-reach user graph $g^m(u_k)$ is built and updated during the process of DFS in the post graph (Step 14 in MR_DFS). In our user graph model, there are two basic attributes: node type and edge weight. The process of updating *m*-reach graph actually refers to changing the values of these attributes. The MR_Build_Graph algorithm in Figure 11

shows how to build and update for $g^m(u_k)$.

An example of building *m*-reach graph is illustrated in Figure 12: (a) is a post graph showing the relationship between six posts (node 1 to 6) with four authors A, B, C and D; (b) is the *m*-reach user graph converted from (a), and each node represents a user with its post IDs labeled in the bracket. In (a), node 1 is a starter with author A while the other nodes are its descendants. The edge weights are labeled beside the edges. Suppose we want to convert this post graph to an *m*-reach user graph with $m = 2$, node 2 and 4 are in 1-reach, and they belong to the same user B, so the two nodes are merged into one node B in (b), while the weight of the edge B→A is the sum of the weights for 2→1 and 4→1. If we look at the chain of nodes 1, 3, 5 and 6, the path lengths for the descendants to reach the starter are: $m_3^{(C,A)} = 1$; $m_5^{(D,A)} = 2$; $m_6^{(C,A)} = 1$. It should be noticed that the author of node 6 is C, which is the same as node 3, so the path length for node 6 is reduced to 1. If node 6 has followers of other users, those followers will have the path length equal to 2, therefore they will also be considered within *m*-reach.

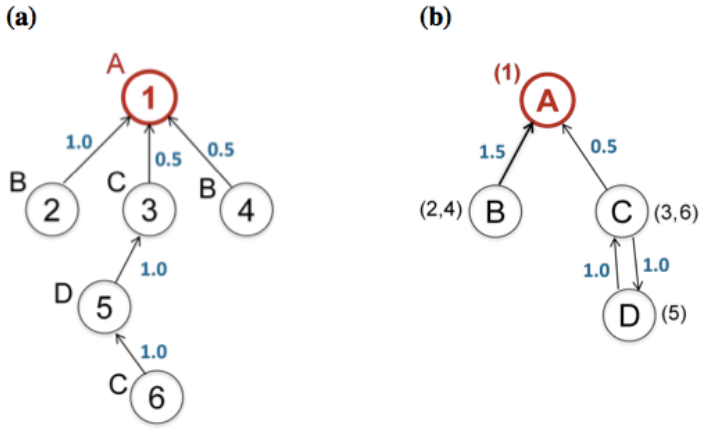

Figure 12 Example of building m-reach user graph from post graph

As for the connecters, because they are defined as bridges to link with starters, they are certainly in 1-reach to a starter. This means all the connecters will be included some

starter's *m*-reach graphs as long as $m \geq 1$.

4.2.2 Measuring the local influence of u-starter

The m-reach graph can be used to measure the local influence of u-starters. We proposed three measures to calculate a u-starter's influence in its m-reach graph from three aspects.

The distance-weighted centrality of a node has been defined in equation (8). It is a measurement that counts the number of its descendants with the weight reciprocal to their distances. The distance information can be obtained from the post graph and used for calculating the influence score of a u-starter on its descendants. It is defined that for the u-starter u_k, the maximum value of its influence on each user is 1. Suppose $d_a^{(x,\,k)}$ is the shortest distance between v_a^x and s_i^k in G_v, then the influence score of u_k on u_x is:

$$I_k(u_x) = \mathrm{Min}\left(\sum_{v_a^x \in Des(s_i^k)} \frac{1}{d_a^{(x,k)}},\ 1 \right) \tag{4.1}$$

The centrality influence of u_k is the sum of influences on all its descendants in the *m*-reach graph $g^m(u_k)$. Let $C(u_k)$ be the centrality influence score of u_k.

$$C(u_k) = \sum_{u_x \in g^m(u_k)} I_k(u_x) \tag{4.2}$$

The users in an *m*-reach graph actually form a community. Graph density is used to measure how many of the users within the community have interactions with others. Suppose $|E_k^m|$ is the number of edges in the *m*-reach graph $g^m(u_k)$, $|V_k^m|$ is the number of nodes, and $|V_k^m|(|V_k^m| - 1)$ is the maximum possible number of edges in a directed graph. Then, $D(u_k)$, the graph density of u_k's *m*-reach graph would be:

$$D(u_k) = \frac{\left|E_k^m\right|}{\left|V_k^m\right|\left(\left|V_k^m\right| - 1\right)} \tag{4.3}$$

The third measure considers how strong the interactions are in the u-starter's community. It is measured by the sum of weights of all the edges in the *m*-reach graph.

$$N(u_k) = \sum_{g^m(u_k)} w(u_y, u_x) \tag{4.4}$$

The three measures are combined into an influence score of the u-starter u_k using the formula below:

$$M_S(u_k) = \alpha \cdot C(u_k) + \beta \cdot D(u_k) \cdot (\left|V_k^m\right| - 1) + (1 - \alpha - \beta) \cdot N(u_k) / \left|V_k^m\right| \tag{4.5}$$

where α and β are positive, and $0 < \alpha + \beta < 1$. Each factor is weighted depending on user's need and the feature of real data. Besides, there are normalization factors associated with $D(u_k)$ and $N(u_k)$.

4.2.3 Merging *m*-reach Graphs

Since an *m*-reach graph is built for each u-starter separately, it is possible that one user results with several *m*-reach graphs. Here, we would like to merge the common user nodes as well as their edges in different *m*-reach graphs. Figure 13 shows an example of merging two u-starters' *m*-reach graphs ($m = 2$).

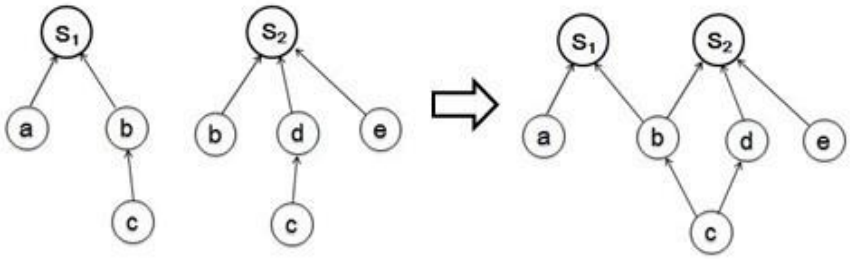

Figure 13 Merge 2-reach Graphs of Two U-starters

The process of merging nodes includes the combination of node types for the same user. The edge weight will remain the same if there is only one edge from one user to another.

In the cases that more than one edge exist between two users with the same direction, the maximum edge weight among them will be taken as the merged edge weight. The reason of not taking summation and normalization of the edge weights is that the user graph is not a complete graph, so it will be over weighted if the graph happens to capture more interactions between the two users. These operations are all associative. So the overall action of merging is associative, which means the result is unique no matter what the merging sequence is.

4.2.4 Measuring the influence of u-connecter

As u-connecters are 1-reach from their connected u-starters in the graph, after merging the *m*-reach graphs of different u-starters, the u-connecters should be in a graph joining all *m*-reach graphs from u-starters they connect.

First, if a u-connecter links two u-starters which are connected directly after their m-reach graphs are merged, there is no need to have a connecter here. For the existing u-connecters, there should be a way to measure and compare their influences. We adopt the method of Shortest-path Cost Measurement used in the post graph model to identify connecters. The basic idea is to remove the u-connecter from the user graph and measure the impact on the influence propagation from the u-starters. The same formula is used to calculate the influence of a u-connecter u_k:

$$M_C(u_k) = \sum_{u_d \in Des(u_k)} (C(u_d, G_u, \varnothing) - C(u_d, G_u, u_k)) \tag{4.6}$$

However, the above formula has a different meaning, as the ancestors are replaced with a u-starter here. Let $C(u_d, G_v, u_k)$ be the sum of the *relationship strength* (as defined in *Section 3.4.2*) from u_d to the u-starter after removing u_k from the graph G_v. This u-starter should be the parent of the u-connecter u_k. In the case that u_k has several u-starters as parents, the sum of measuring results for several u-starters will be taken as

the final influence score of u-connecter u_k.

Besides the existing ones, some new u-connecters may be found as a broker to link two u-starters (one is his parent and the other is his child in the user graph). We can also use the above method to measure their influences. When a new u-connecter does not have a post as a connecter in the post graph, it means they have not behaved as connecter within a discussion and they are only considered as potential connecters which should have the ability but not yet functioned.

4.2.5 Connecting distant u-starters

After merging the *m*-reach graphs, there may still be disconnected subgraphs, or some isolated *m*-reach graphs of u-starters. In order to connect them and discover inter-starter influences, we built virtual edges between distant u-starters.

The u-starters are defined as distant when they do not present in each other's *m*-reach graphs (e.g. S_1 and S_3 shown in Figure 13). For finding possible connections between distant u-starters, we looked up in the post graph initially and determined the existence of directed path between them. For example, if u_j and u_k are not in the *m*-reach graph of each other, first we want to check in the post graph if there is a directed path from u_j's post to u_k's. Let v_i^k (i = 1,2,3,…) be u_k's posts in G_v. For each v_i^k, it has a descendant set $Des(v_i^k)$. We need to find out whether u_j has a post v_d^j in $Des(v_i^k)$.

Once the condition is met, it means at least one path exists from u_j to u_k, then we will build an virtual edge from u_j to u_k. The edge weight is calculated as the shortest path length (number of distinct users) between them. Similarly, we can check if the inverse path from u_k to u_j exists. The edges are considered as different in opposite directions.

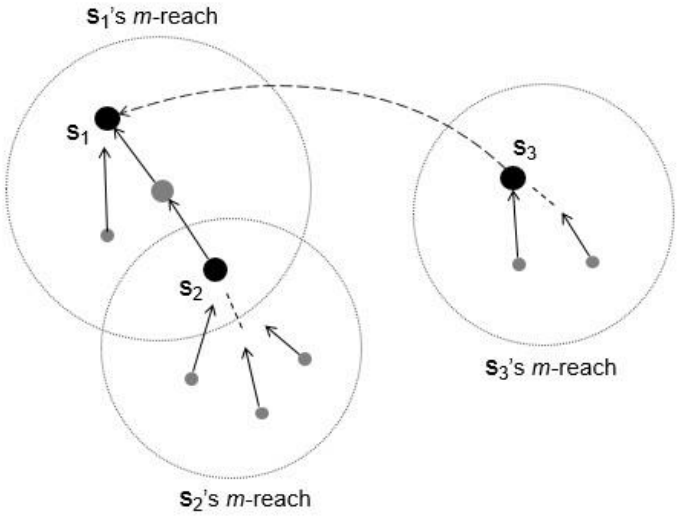

Figure 14 Connect Distant U-starters (from S_3 to S_1)

4.2.6 Measuring the influence between u-starters

A u-starter can be more influential if it influences other u-starters. The influence between u-starters can be easily captured if a u-starter is in another's *m*-reach graph. However, it is possible that they have some influence on others through paths longer than *m*-reach. In this case, the influence is considered inversely related with the distance between the u-starters. As defined before, $w'(s_k, s_j)$ is the shortest path length from u-starter s_k to s_j. After all the possible virtual edges are built, the influence of one u-starter s_k on another one s_j can be calculated by:

$$I_k'(s_j) = \begin{cases} 1 & \text{if } s_j \text{ is in } g^m(s_k) \\ \min(\dfrac{m}{w'(s_j, s_k)}, 1) & \text{if } e'(s_j, s_k) \text{ exists} \end{cases} \tag{4.7}$$

where the value of *m* will be determined in the experiment. This measure counts s_j's influence on s_k as 1 if $s_k$is in s_j's *m*-reach graph. Otherwise the influence would be at most 1. An example is shown in Figure 13 that S_1, S_2 and S_3 are u-starters, and S_1 has

influence on S_2 and S_3. The influence of S_1 on S_2 counts as 1 as S_2 is in S_1's *m*-reach graph, while S_1's influence on S_3 is measured by the second formula. Finally, the influence of the u-starter s_k on other starters is the summation of influences on each one:

$$M_I(s_k) = \sum I_k'(s_j) \tag{4.8}$$

4.2.7 Efficiency improvement by sampling

If a complete post graph is to be converted, in order to keep the information of relationships between posts within the same discussion thread, the *m*-reach graph for each user starter (u-starter) is required for its local influence measure. The value of *m* will be set to infinite, which means all nodes directly or indirectly connected to the starter via paths should be visited. The complexity of finding shortest path for each starter is $O(|V|^2)$. Therefore building *m*-reach graphs for *n* starters will cost $O(n|V|^2)$. When a post graph is converted to the corresponding user graph, and the *m*-reach graphs will be merged or connected. Each node and edge in the post graph should be visited for the conversion. The complexity of converting the whole post graph is $O(|V|+|E|)$ and that of finding the shortest path between starter is $O(n^2)$.

The complexity is largely reduced by using our sampling method. Building an *m*-reach graph will cost c^{m+1} (*c* is the fan-out value of the graph) which should be a constant. Also, it will cost $O(n^2)$ to connect distant starters with shortest paths. Therefore, the efficiency can be greatly improved by this sampling method when dealing with graphs in large scale.

However, the information of those descendants which are far away from starters and connecters will be lost after sampling. This may affect the accuracy in finding influential users. The experiment is conducted to compare the results of graph conversion with and

without sampling, which is discussed in Section 5.2.4. It shows that the model using sampling method can achieve similar results as the one with whole graph conversion. Therefore, the loss of information of distant descendants in the sampling method is considered to have little influence on the accuracy in identifying influential users.

Chapter 5 Experiments

5.1 Case study for post graph model

A case study is conducted to verify the post graph model. We applied the three influence measurement methods to see whether the influential posts as starters and connecters can be detected.

5.1.1 Data Description

Our proposed model can be applied for different social media. Both explicit and implicit relationships can be identified between text-based posts. For our experiments, we chose Twitter to conduct the experiments as it has many users and its data are easy to collect.

In order to find the most influential posts and their respective authors during the information diffusion within a topic, we selected a general user (neither famous people nor public media) who has written some posts on a topic, found the user's friends who have responded to the posts or also talked on this topic, then found out the friends of friends and so on. Completely random sampling method is not suitable here, because we need data from users with more connections between them so that the graph can be well formed. Tweet data are collected on the topic of "Steven Jobs and iPhone 4s". The keyword set is defined as {"iPhone 4", "iPhone 4s", "iPhone 5", "iPhone Mini", "Steve Jobs", "Apple", "ios 5", "Siri"}. As they are specialized terms, their synonym sets are empty. For illustration of the implicit relationship between posts, we showed how the similarity score is calculated for the two posts below in the next few paragraphs.

Post A: “Just heard someone succeed in running **Siri** on **iPhone4**, although I already changed to **iPhone 4s**...”

Post B: “I'm not going to buy an **iPhone 4s**, it’s nothing special but **Siri**. I believe the last work by **Steve Jobs** is **iPhone 5**, I'll wait for it!”

Here post B is not a reply of A, but it is published later than A. And B’s author is a friend of A’s author. In this case, their implicit relationship will be judged by the content similarity. There are 3 keywords found in A, and 4 in B, with 2 of them in common. With Tanimoto coefficient method, the similarity score should be calculated as the number of keywords in intersection over the union, which is 2/5 = 0.4. In addition, the interval between the two posts does not exceed the expiry time. Therefore the weight of the edge representing their implicit relationship is 0.4. Table 3 gives the data description for the experiment. In this case study we use a small data set, because we intend to visualize the influential nodes in a graph in order to compare the results of different measurement methods.

Table 3 Description of Data (Case Study 1)

Platform	Twitter
Topic	Steven Jobs and iPhone 4s
Time	11/10/2011 - 31/10/2011
Location	Hong Kong
No. of users	158
No. of tweets	211

5.1.2 Preliminary Results

Starters and connecters can be found after the three influence measurement methods are applied. As mentioned before, degree measure can be used to identify starters. Two factors are calculated: (i) the degree of each node $d(v)$; (ii) the weighted average of its follower in-degrees $s(v)$. The top nodes that $d(v) + s(v) > 2$ are selected. The results are plotted in the diagram shown in Figure 15.

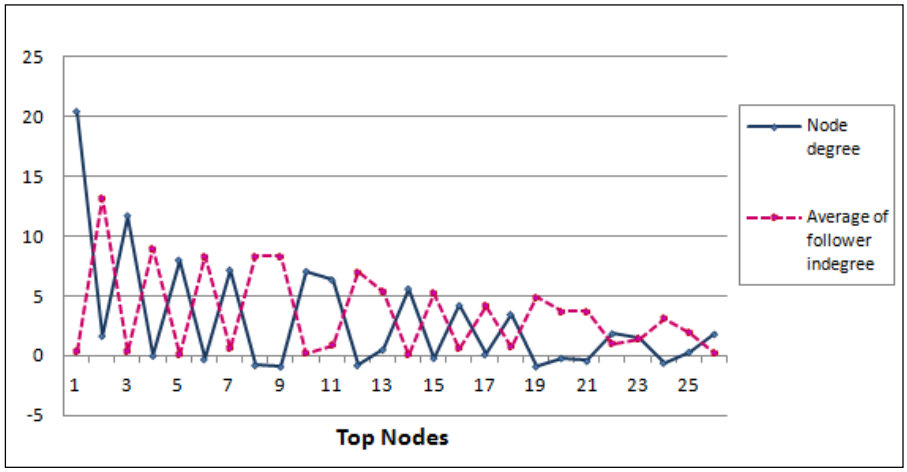

Figure 15 Degree Measures

It is observed that the results of the two factors are not aligned most of the time. The reason is that a node with higher degree should have more followers, and it becomes difficult for all its followers to have a high in-degree. On the contrary, there exist some nodes with only a few followers, but most of the followers have high in-degree. These nodes can be detected by high score of $s(v)$. For our work, we finally selected the 10 nodes with $d(v) > 3$ and $s(v) > 0.1$ as starters (Node 1 – 10 labeled in Figure 16).

As for the connecters, we integrate the results from Shortest-path Cost Measure (SCM) and Graph Entropy Measure (GEM). After calculating the influential scores $Inf_c(v_i)$ and

$Inf_e(v_i)$, all the nodes are ranked. After examining the top ranking nodes, besides the found starters, other nodes which connect starters are considered as connecters (Node 11 – 20 in Figure 16). The connecters discovered by each method are listed below in ranking order.

– SCM: 11, 14, 13, 12, 15, 16, 20, 17 (nodes labeled in Figure 16)
– GEM: 11, 14, 12, 15, 13, 16, 20, 17, 18, 19 (labeled in Figure 16)

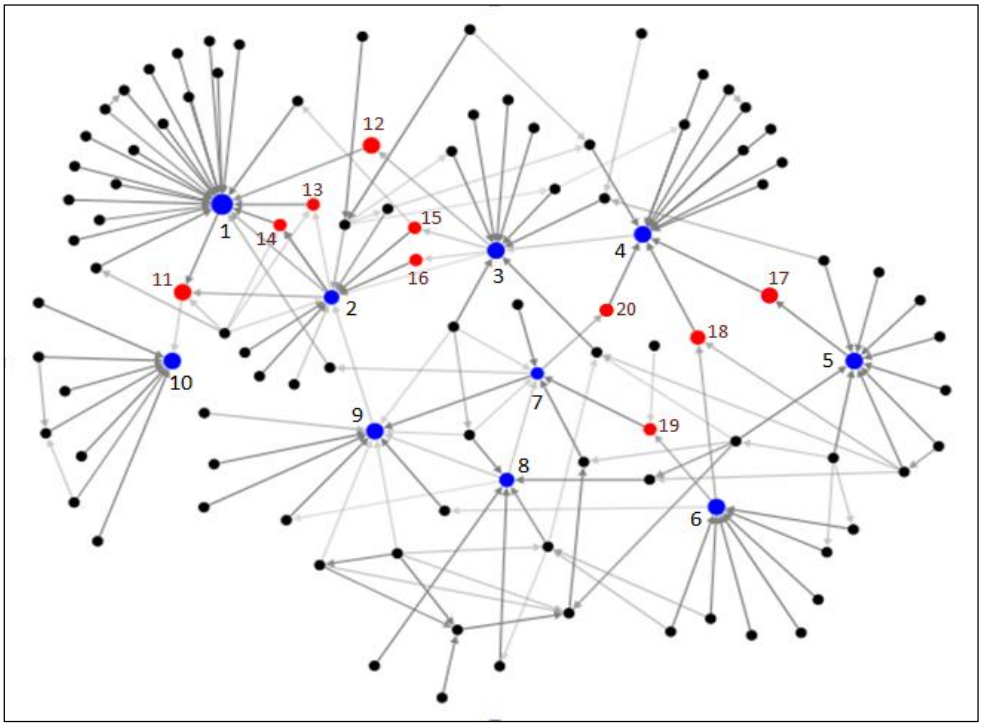

Figure 16 Graph of Starters and Connecters

Figure 16 only shows the starters, connecters and the nodes related to them. Strong relationship is represented in full line. For the nodes, their types are labeled in color, and the node size is proportional to their degrees.

5.1.3 Discussion and Final Results

In comparison, SCM only identified 4 starters in its top 10 ranking nodes, and was able to find all starters in top 21; while GEM could find 7 starters in top 10 and all starters in top 14. It is because GEM looked into both node entropy $E(i)$ and remnant graph entropy $EN(i)$ in calculating the influence score, which was aimed to achieve high node centrality as well as large effect in the graph after removal. As for the SCM algorithm, we can see that its influence score is in the range from 0 to the number of the node's descendants. There is no difference between its close followers and distant descendants when measuring a node if the weights are all 1. As a result, it is more likely to find the nodes with more descendants, whereas GEM can find the nodes with more ancestors or descendants.

Finally, we can find the most influential posts considering the results of all measurement. For the starters, node 7 and node 8 by different authors are ranked low by SCM and GEM, so they are not considered to be influential in the final result. Since node 7 is not influential any more, we look at the connecters 19 and 20 that connect node 7. It is found that they also have relatively low rankings. Therefore they are also removed from the influential list.

Noted that not every node connecting two starters can be a connecter. The connecters are detected by the two measurements, which means their removal from the graph will have a certain impact on the information transmission, and they should have some followers to make them more influential. In Figure 16, we can see that nodes 13 and 14 are actually connecting the same starters 1 and 2, and so are the nodes 15 and 16 which connect starters 2 and 3. These connecters are less influential than those who connect starters as the only bridge.

5.1.4 Performance Comparison

We compared the results of our methods with some centrality metrics and PageRank algorithm in the same graph. From Figure 16, we can identify the following 8 nodes as influential starters: 1, 2, 3, 4, 5, 6, 9, 10. Besides, there are 8 influential connecters determined: 11, 12, 13, 14, 15, 16, 17, 18. The top 20 ranking nodes are retrieved as influential posts for each measurement. Their performance is compared through the numbers of influential starters/connecters that can be found in top 20. The comparison details are shown in Table 4.

Table 4 Comparison of results by centrality, PageRank and our model

Measurement Method	No. of starters identified	No. of connecters identified	Remark
Betweenness Centrality	8	7	Connecter 17 missed
Closeness Centrality	5	6	Starter 5, 6, 10 missed Connecter 17, 18 missed
Eigenvector Centrality	4	6	Starter 4, 5, 6, 10 missed Connecter 17, 18 missed
PageRank	8	1	Only connecter 11 is identified.
Our Model	8	8	All the starters and connecters are identified.

It is found that, besides our model, the Betweenness Centrality and PageRank are also able to identify all the influential starters. It is noted that the starters exactly rank top 8 by PageRank algorithm. However, PageRank can just identify 1 connecter, which is the least one compared with the others. Closeness Centrality and Eigenvector Centrality find

the same set of 6 connecters. Betweenness Centrality performs comparatively well as it just missed 1 connecter. But our model is able to identify all the influential starters and connecters. In conclusion, our model outperforms the other 4 methods in this case.

5.2 Case study for user graph model

The data set used in the previous case study is not large enough to show the difference. In order to compare the results in post graph and user graph, in this section, we reported an experiment with a larger data set. We collected more than 1700 tweets from 915 users, on the topic of "Sichuan Lushan earthquake" (an earthquake happened in China on April 20, 2013) and "H7N9 influenza". More description on the data set is shown in Table 5.

Table 5 Description of Data (Case Study 2)

Platform	Twitter
Topic	"Sichuan Lushan earthquake", "H7N9 influenza"
Time	31/03/2013 - 30/04/2013
Location	China, Hong Kong, Japan

5.2.1 Influential users in post graph

Similar steps have been done as in previous case study to identify starters and connecters in the post graph. Finally, 49 starters and 5 connecters are found in the posts. Some starters or connecters are actually written by the same authors, so we identified 29 users as influential in total. If we rank the influential users found in post graph according to the number of starters/connecters they have, the top 5 results are listed in Table 6. For those with the same number of starters and connecters, they are ranked based on the

highest ranking of their posts.

Table 6 Information of Top Influential Users in the Post Graph

Rank	1	2	3	4	5
No. of starters and connecters	6	6	5	3	2

In order to justify our user graph model, we converted the post graph into user graph, and then measured the users' influences in the user graph of two types: u-starter and u-connecter.

5.2.2 The influence of users as starters

First, we converted the post graph into user graph, and then measured the users' influences in the user graph of two types: u-starter and u-connecter. The local influence of a u-starter is measured by three factors as shown in (4.5). In this experiment, we put more weight on the centrality measure, set $\alpha = 0.5$, $\beta = 0.25$, so the last factor is 0.25. The value of m is decided by the distance between close starters in the post graph. We tried to make more m-reach graphs contain only one starter, meanwhile have common descendant nodes so that they are connected after the merging operation. For the cases that the starters are far away from each other, we suggested the m value not larger than 5. The influence between u-starters is also taken into consideration. When some u-starters have similar local influence, their ranking will be judged by the inter influence measure. After all, the ranking of influential starters is a little different from that in post graph. In top 5 influential users found

- Top 2 users keep the same.
- A new influential user is identified on rank 3 in user graph.
- The user on rank 4 in post graph does not rank on top in user graph.

The new influential user found in the user graph only has one post as starter. However, this starter has a large number of followers, and these followers have interactions with each other, which makes its local influence score higher. Figure 17 and 18 shows the post graph and corresponding user graph of this node and its descendants within 5-reach. For the user falling off the top 5 list, the main reason is that his followers or friends are from a small community, and there are no connecter to propagate their discussion to another community.

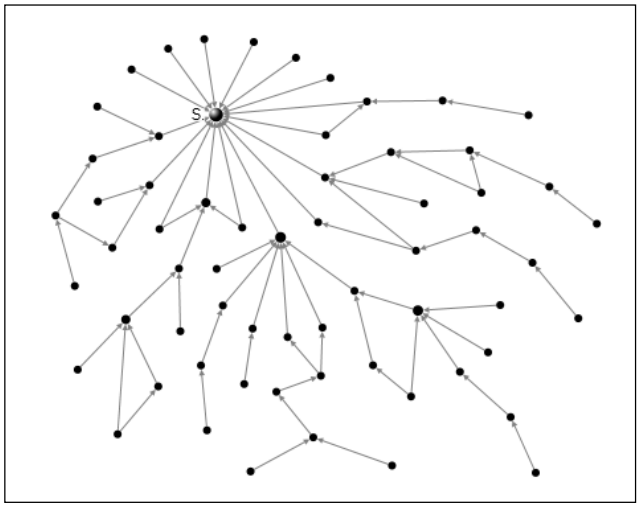

Figure 17 Post Graph of an Influential Starter

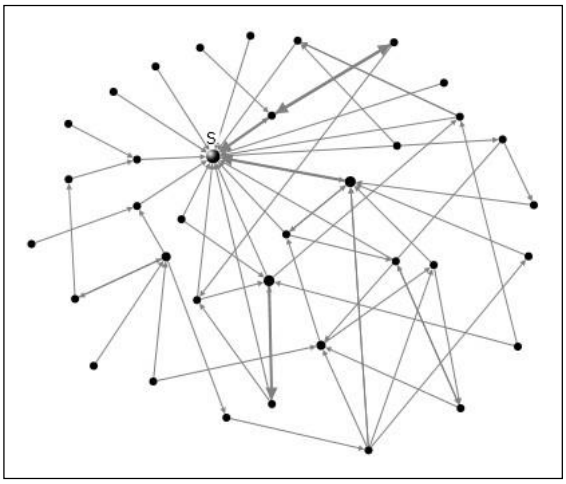

Figure 18 User Graph of the Starter

5.2.3 The influence of users as connecters

For the same dataset, in the post graph, there are 5 connecters identified. Yet, But after the graph is converted into user graph, it is found that 2 of them are not connecters anymore, because the u-starters they link with are directly connected. The remaining 3 u-connecters are determined to be influential users.

However, 1 new u-connecter is found in the user graph, who which links with 2 different u-starters. As stated above, it is only considered as a potential connecter. The result proves that in the post graph the connecters already identified can be refined and some new connecters may be found. The new connecters are not that influential as they are just supposed to have the ability but have not acted as a connecter before. Therefore the identification of influential connecters will be more accurate and complete if the data set is large enough.

5.2.4 Experimental results comparison

In order to compare the results after graph conversion with sampling and the whole graph, we set the value of *m* to infinity and completed graph conversion with the same data set. The difference in the number of starters and connecters found in the post graph, and the user graph converted with or without sampling is given in Table 7.

Table 7 Comparison of results with/without sampling

	No. of starters	**No. of connecters**	**No. of new connecters found in user graph**
Post graph	49	5	-
User graph converted by sampling	29	3	1
User graph converted from the whole graph	29	3	2

The results show that the sampling method can achieve the same result as the whole graph conversion in identifying starters. The top 5 influential starters have the same rank in both cases. It shows no big difference even though only one case of the inter-starter influence is captured by sampling. As for connecters, it can be seen that 1 more potential connecter is found when the whole post graph is converted, while the remaining number of connecters obtained from the post graph keeps the same.

5.3 Experiment on Sentimental Influence

5.3.1 Data Description

Data is collected from 4 popular social networking sites in Hong Kong: Twitter, HK Discuss forum, Yahoo Blog and Yahoo News, and there are influential posts found on 10 topics. There are 5564 posts and replies on the 10 topics from Twitter, 1600 from HK Discuss and 863 from Yahoo News/Blog. The topics are categorized into public topics and personal topics, as users are likely to be affected by external sources on public topics, such as current affairs and news. If a topic is generated by a user who first posted it, the sentiment of its replies should be mainly influenced by the user's post. The description of topics is given in Table 8.

Table 8 Description of Topics for Sentimental Influence

	Content	Category
Topic 1	Donation for Philippine typhoon disaster	Public
Topic 2	Snowden seeks Hong Kong's help	Public
Topic 3	English Premier League Football: Manchester United v.s. Arsenal	Public
Topic 4	Milk powder restrictions by Government	Public
Topic 5	Striking for HKTV's license rejection	Public
Topic 6-10	Family/life, social problem, knowledge sharing, personal experiences	Personal

In order to compare the results and findings on different types of social media platforms,

we selected a microblogging site Twitter, a local forum HK Discuss and Yahoo Blog to find the influencers on personal topics. Twitter and HK Discussion are also used for public topics. It is a general assumption that the posts by famous people, such as celebrities and idols, should get much more positive responses due to fans' favor, but not influenced by the post content. In our experiment the posts are collected from general users. As there are seldom blogs talking on public topics and having many responses on Yahoo Blog, we changed to use Yahoo News to conduct experiment on public topics.

5.3.2 Results and Findings

1) *Do the posts on the same topic have similar sentiment trend on different social media platforms?*

The average sentiment scores for the 10 topics on different social media platforms are measured and plotted in Figure 19 for comparison. Figure 19(a) shows the results for public topics, where it is found that HK Discuss and Yahoo News have similar sentimental trend except for Topic 2. There is a slightly bigger difference of the average sentiment on Twitter. But for the personal topics as shown in Figure 19(b), there are much more differences on the three platforms.

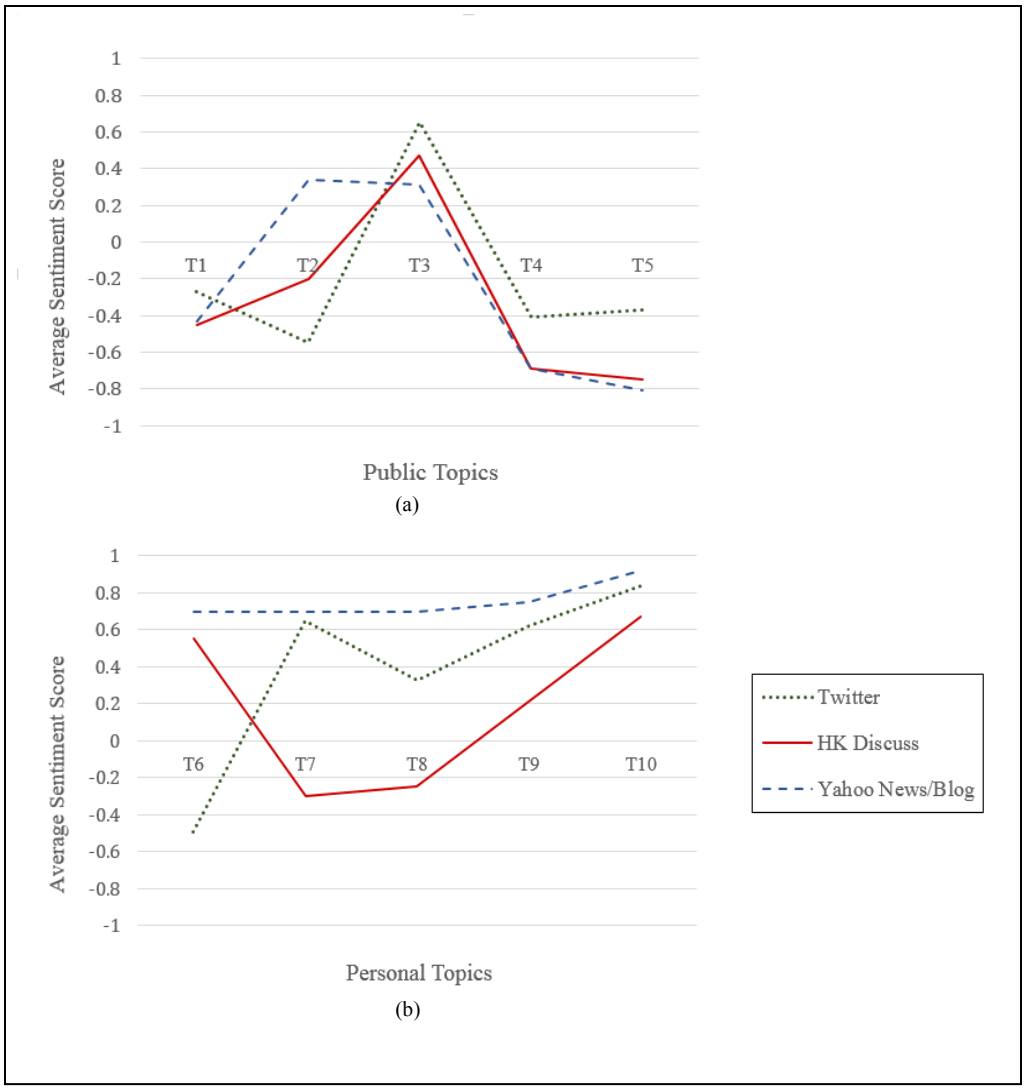

Figure 19 Comparison of average sentiment on different social network

This result is reasonable because the details of post contents from different social networks can be very different from each other, although they belong to the same topic category.

2) How much does the sentiment of influencers affect the receptors' sentiment?

The sentimental influence has been defined in two types: compliance and opposition. In

order to know whether the types of the influencers' sentiment have different effect on their sentimental influence, the average complying rate and average opposing rate are calculated for the influencers with positive and negative sentiment separately. The results are shown in Table 9.

Table 9 Sentimental Influence (Complying / Opposing Rate)

Influencer's Sentiment	Public Topic		Personal Topic	
	Complying Rate	Opposing Rate	Complying Rate	Opposing Rate
Positive	56.5%	34.0%	72.5%	11.0%
Negative	71.1%	19.6%	35.3%	38.3%

It can be seen from Table 9 that the complying rate is very high for the influencers with negative sentiment on public topics, and for those with positive sentiment on personal topics. As for the opposing rate, it is relatively high for the influencers with negative sentiment on personal topics. It should be noted that some of the influencers have no obvious sentiment on public topics. In our definition, if the influencer's sentiment is neutral, the sentimental influence cannot be judged. So the influencers with neutral sentiment on some public topics are not included in this analysis result. But we also calculated the averages of the proportion of positive and negative responses for them, which are 37.2% and 48.7% respectively. This result shows no obvious sentiment preference of the receptors.

5.3.3 Discussions

We compared the results of average sentiment score on different social media platforms, and found that on public topics HK Discuss and Yahoo News have similar sentiment

trends. It may be because the users on the two social media are mainly local people in Hong Kong, while on Twitter the users are more international. Our method can be used to find and compare the sentiment trend on different social media platforms. There are many factors that can cause the difference, such as the user groups, the popularity of the social media and their different features. The emoticon set provided by the platform is an important indicator of the users' sentiment. Therefore, before analyzing the sentiment on other social media, their special emoticons should be collected.

In addition, the experimental result shows that the influencers with negative sentiment on public topics, and those with positive sentiment on personal topics, can have a greater sentimental influence of compliance on their receptors. In our observation, the negative sentiment on public topics can get more people's attention, and their sentiment trends are easier to be induced. Meanwhile, on the personal topics, the sentiment of their responses is more likely to be positive. When the influencer talks positively, the repliers are tend to be positive to support his or her opinion. On personal topics their chance with positive sentiment is not that low as on public, even when the influencer expresses a negative view. The main reason is that many repliers would like to help and comfort the user. This phenomenon is highly common on Yahoo Blog.

Chapter 6 Conclusion

In conclusion, we dealt with the problem of finding influential users based on their interactions in social networks. Different from others' work, we tried to identify the most influential users in different roles through their posts on the same topic initially, then discover the influential users. Additional contributions are the following:

1. We proposed the Comment Arrival Model in order to determine the lifespan of online posts, and a graph model showing the explicit and implicit relationship between posts. The models could be applied in different social media platforms that generate textual posts.
2. We presented three measurements to reflect the influences of online posts so as to distinguish starters and connecters in the graph. We extended previous definitions of the node centrality and graph entropy to apply them in our model.
3. We proposed the user graph model and different measurements to clarify the influences of starters and connecters. Instead of building the interaction graph directly, we converted a post graph to the corresponding user graph using biased sampling which selects potential influential users identified in the post graph as seeds. The cost of building user graph is reduced, and the information of discussion chains from the post graph is also considered.
4. We proposed a method to determine the sentiment of posts based on the emoticon list and sentimental word dictionary. We further defined and analyzed the sentimental influence of the posts.

Our graph model has its advantage in dealing with online posts and users with more

interactions. We have carried out case studies and visualized the graphs to validate the models. It is found that the Shortest-Path Cost Measure and Graph Entropy Measure are able to detect both starters and connecters in the post graph. We also compared the performance of our methods with the three centrality metrics, Betweenness centrality, Closeness centrality and Eigenvector centrality, as well as the PageRank algorithm. The experiment result shows that our proposed methods in the post graph model can identify all the influential starters and connecters in the respective datasets, which outperforms the other four. Besides, the influential users identified by the post graph model and the user graph model with or without sampling are compared. It shows similar results when the graph is converted using the sampling method; in addition, the users who have more influential posts are not certain to be more influential, when the followers are always the same group of people.

Although the data cleaning and sampling methods are adopted to reduce the complexity, post graph still very high, especially for Facebook and Twitter, where millions of posts generated hourly. Furthermore, there are some computational models we used that need further improvement:

i. Lifespan of a post: currently the parameter of external influence is added into the model as a ratio to adjust the expiration time. This model could be improved by considering more factors, and applying other functions to set the parameters, such as sigmoid function.

ii. Content similarity between posts: the post graph model can be more effective if it is integrated with advanced text mining techniques, so that the relevance between posts can be judged more accurately.

iii. Edge weight function: a simple model (linear function) is used for the edge weight calculation. Actually there are other options such as exponential function, which drops smoothly. The edge weight model should be improved and need more testing in future.

Besides, the future work could include the analysis of sentimental influence of users. The sentimental influence of posts has been analyzed. However, it can be more complicated if we want to find the sentimental influence between users. Intuitively, the sentimental influence of user A on user B could be judged through the calculation of the probability that B's post complies with or opposes A's post. One problem encountered is that Heider's Balance Theory cannot be applied in the user graph, because it can be a directed graph with circles. Therefore a solution is needed to determine sentimental influence between people who are not directly linked by edges.

Appendix

List of Emoticons:

Positive Sentiment

Emoticon	Meaning
:-) :) :] =) :3 :>	Smile/Happy
:-D :D 8D =D =3 XD XDD	Laugh
:-P :P =P	Naughty
;-) ;) ;D ;P	Wink
:-* :* ^3^	Kiss
^_^ ^o^ ^^	Smile
8-) 8) B-) B)	Smile
<3 <33	Love
d(>w<)b	Thumbs up (Like)

Negative Sentiment

Emoticon	Meaning
:-(:(:[=(	Frown/Sad
:-O :O =O= -o-	Shocked
>:-(>:(	Grumpy
:-/ :/ :-\ :\	Unsure/Doubt
:'(;*(T_T T.T TT Y_Y	Cry/Weep
-_- -.- - - =_= =.= = =	Squint
o.O O.o	Confuse
>_< >.< ><	Upset/Painful
:S :-S =S :$:-$	Embarrassed/Hesitated

e_e 9_9 zzz	Sleepy
@_@ 3_3 +_+	Dizzy
x_x	Dead/Unconscious
9_6	Crazy
c.c C_C	Thinking/Disagree
/.\ /_\	Disappointed
=3=	Pout/Unhappy
=_=" = =" = =b -.-' - -'	Sweat
=_=# = =#	Angry
q(;^;)p	Thumbs down (Dislike)

References

[1] A. H.Lipkus. A proof of the triangle inequality for the Tanimoto distance, J Math Chem 26 (1-3): pp.263–265, 1999.

[2] A. Leavitt, E. Burchard, D. Fisher, and S. Gilbert.The Influentials: New Approaches for Analyzing Influence on Twitter. *Web Ecology Project,* http://tinyurl.com/lzjlzq, 2009.

[3] A. Sala, L, Cao, C. Wilson, R. Zablit, H. Zheng, and B. Y. Zhao. Measurement-calibrated Graph Models for Social Network Experiments. In *WWW'10 Proceedings of the 19th International Conference on World Wide Web*, pp.861-870, 2010.

[4] A. Stavrianou, J. Velcin, and J. Chauchat. PROG: A Complementary Model to the Social Networks for Mining Forums. From Sociology to Computing in Social Networks, pp.59-79, Springer, 2010.

[5] B. Sun, and V. TY Ng. Identifying Influential Users by Their Postings in Social Networks. In *Proceedings of the 3rd International Workshop on Modeling Social Media*, pp. 1-8, 201.

[6] B. Sun, and V. TY Ng. Lifespan and Popularity Measurement of Online Content on Social Networks. In *Social Computing Workshop of IEEE ISI Conference*, pp.379-383, 2011.

[7] B. O'Connory, R. Balasubramanyany, B. R. Routledgex, and N. A. Smithy. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of AAAI Conference on Weblogs and Social Media*, 2010.

[8] C. Nobel, and D. J. Cook. Graph-based anomaly detection. In*Proceedings of the 9thACM SIGKDD InternationalConference on Knowledge Discovery and Data Mining*,pp.631-636, 2003.

[9] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, andB. Y. Zhao. User interactions in social networks and theirimplications. In *Proceedings*. *of EuroSys*, pp.205-218, April 2009.

[10] D. A. Tunkelang. Twitter Analog to PageRank, http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/, 2009.

[11] D. Barbagallo, L. Bruni, C. Francalanci, and P. Giacomazzi. An Empirical Study on the Relationship between Twitter Sentiment and Influence in the Tourism Domain. Information and Communication Technologies in Tourism 2012, pp.506-516, 2012.

[12] D.L. Hansen, B. Shneiderman, and M.A. Smith. Visualizing Threaded Conversation Networks: Mining Message Boards and Email Lists for Actionable Insights. In *Proceedings of AMT '10*, pp.47-62, 2010.

[13] E. Bakshy, B. Karrer, and A. Adamic, Lada. Social Influence and the Diffusion of User-Created Content. In *10th ACM Conference on Electronic Commerce*, pp.325-334, Stanford, California, Association of Computing Machinery, 2009.

[14] E. Bakshy, W. A. Mason, J. M. Hofman, and D. J. Watts. Everyone's an Influencer: Quantifying Influence on Twitter. In *WSDM'11 Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pp.65-74, Hong Kong, China, 2011.

[15] eMarketer: Mobile User Bases at Facebook and Twitter Keep Growing. http://www.emarketer.com/Article/Mobile-User-Bases-Facebook-Twitter-Keep-Growing/1010135, 2013.

[16] Fritz Heider´s Balance Theory: http://wilsede-alive.com/?p=29.

[17] H. L. Li, and Vincent T. Y. Ng.Discovering Potential Drug Abuse with Fuzzy Sets. *In Proceedings of the 2010 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp.2656-2662, 2010.

[18] H. Green. Google: Harnessing the Power of Cliques, BusinessWeek, 2008.

[19] H. T. Welser, H. Gleave, D. Fisher, and M. Smith. Visualizing the signatures of social rolesin online discussion groups. *Journal of Social Structure* 8(2), pp.1-31, 2007.

[20] J. C. Zhao, L. Dong, J. J. Wu, and K. Xu. MoodLens: An Emoticon-Based Sentiment Analysis System for Chinese Tweets in Weibo. In *Proceedings of the 18th ACM SIGKDD*, pp.1528-1531, 2012.

[21] J. Han, M. Jiang, and S. Yang. Mining Topic-level Influence in Heterogeneous Networks. In *Proceeding of CIKM'10*, pp.199-208, 2010.

[22] J. Scott. Centrality and Centralization. In Social Network Analysis: a handbook, pp.82-99, London: SAGE Publications, 2000.

[23] J. Shetty, and J. Adibi. Discovering Important Nodes through Graph Entropy. In *the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.74-81, 2005.

[24] J. Tang, J. Sun, C. Wang, and Z. Yang. Social Influence Analysis in Large-scale Networks. In *Proceeding of KDD'09*, pp.807-816, 2009.

[25] J. Weng, E. Lim, J. Jiang, and Q. He. TwitterRank: Finding Topic-sensitive Influential Twitterers. In *WSDM '10 Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pp.261-270, 2010.

[26] Klout Score: http://klout.com/home.

[27] K. M. Ip. Social Website Emotion Sentiment Analysis. Final Year Project Report, BSc in Coputing and BBA in Management, HK Polytechnic University, 2013.

[28] L. Tang, and H. Liu. Graph Mining Applications to Social Network Analysis. Managing and Mining Graph Data. In *Managing and Mining Graph Data*, pp.487-513, 2010.

[29] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging Topic Detection on Twitter based on Temporl and Social Terms Evaluation. In *Multimedia Data Mining Workshop at KDD*, pp.4:1-10, 2010.

[30] M. Cha, H. Haddadi, F. Benevenuto, and K. P.Gummad. Measuring User Influence in Twitter: The Million Follower Fallacy. In *4th Int'l AAAI Conferenceon Weblogs and Social Media*, pp.325-334, Washington, DC, 2010.

[31] M. Dehme.Information processing in complex networks: Graph entropy and information functionals. In *Applied Mathematics and Computation*,vol 201, pp. 82–94, 2008.

[32] M. Mathioudakis, and N. Koudas. Efficient Identification of Starters and Followers in Social Media. In *EDBT '09 Proceedings ofthe 12th International Conference on Extending Database Technology: Advances in Database Technology*, pp.708-719, 2009.

[33] M. Trusov, A. V.Bodapati, and R. E.Bucklin. Determining Influential Users in Internet Social Networks. Journal of Marketing Research, pp.643-658, 2010.

[34] M. U. Ilyas, and H. Radha. A KLT-inspired Node Centrality for IdentifyingInfluential Neighborhoods in Graphs. In *Conference on Information Sciences and Systems*, pp.1-7, 2010.

[35] S. Asur,and B. A. Huberman. Predicting the Future with Social Media. In *Proceedings of 2010 Web Intelligence*, pp.492-499, 2010.

[36] S. P. Borgatti. Identifying Sets of Key Players in a Social Network. Computational and Mathematical Organization Theory, pp.21–34, Springer, 2006.

[37] SentiWordNet: *http://sentiwordnet.isti.cnr.it/*.

[38] SCWS 中文分詞: *http://www.xunsearch.com/scws/index.php*.

[39] Twinfluence: *http://twitterfacts.blogspot.com/2008/10/twinfluence.html*.

[40] Yahoo News: How Facebook has grown: Number of active users at Facebook over the years, May 2013.
http://news.yahoo.com/number-active-users-facebook-over-230449748.html.

[41] Y. J. Tang, and H. H. Chen. Emoticon Modeling from Writer/Reader Perspectives Using a Microblog Dataset. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology, the 5th International Joint Conference on Natural Language Processing*, pp.11-19, 2011.

[42] Z. Yang, and M. Purver. Predicting Emotion Labels for Chinese Microblog Texts. Unpublished manuscripts, School of Electronic Engineering and Computer Science, Queen May University of London, 2012.

[43] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137-146, 2003.

[44] Goyal, F. Bonchi, and L. Lakshmanan. A data-based approach to social influence maximization. In *Proceedings of the VLDB Endowment*, 5(1): pp.73-84, 2011.