A SIMULATION-BASED APPROACH TO ANALYZE THE INFORMATION DIFFUSION ON TWITTER :

Seeds Setting Strategy for Spreading Control

Yimei Zhu[1]

Email: ucabhuc@ucl.ac.uk

MSc in Web Science and Big Data Analytics

University College London

Project supervisor: Dr. Shi Zhou

Submission date: September 4, 2018

Abstract

Information dissemination in complex networks such as online social networks is always triggered through the activation of a small number of initial nodes as the seeds. Meanwhile, the seeds settings could determine the speed and coverage of spreading to a great extent. This research established the topological simulation environment by part of the Twitter network and constructed an iterative simulator with an infection rate predictive model and the susceptible-infected epidemic model to intuitively reproduce the message propagation process. It simulated and compared the spread under ten different seeding strategies, mainly analyzing three aspects which may promote influence maximization: the criteria for seeds selection, the seeds appearance sequence and the seeds activity level.

The results show that seeds selection based on higher degree performs better than higher betweenness or eigencentrality. Meanwhile, sequential seeding strategies can better trade-off the speed and coverage than the single-stage one while the dynamic one dominates the others. Moreover, higher active seeds can finally reach a more extensive influential range. Finally, a combined seeding strategy taking advantage of the best conditions above was generated in this research, contributing to the best performance in information diffusion till now. Together with all the simulation results, this strategy would provide more effective theoretical bases and practice support for the prediction and control of the messages propagation regarding the seeds settings.

Acknowledgements

In the first place, I would like to deliver my sincere gratitude to Dr. Shi Zhou for his supervision and suggestions on this research topic. I am very grateful for helping me determine the main exciting research direction and for continually stimulating me to develop further ideas into this dissertation.

Secondly, I would like to thank the project partner, Liyi Zhou, for providing the accurate infection rate predictive model for my simulation process and for his patient explanation on the confusion I encountered related to it.

Furthermore, I want to thank my boyfriend, Jiaqi Ye, for his wholehearted support and companionship during my studies, especially for his encouragement when I feel deep stressed. Here I also would like to deliver my gratitude for his opinions about how this topic could be developed in the future.

Finally, I would like to offer my heartfelt gratefulness to all family and friends of mine for their sincere care and assistance provided.

Contents

List of Figures

Chapter 1

INTRODUCTION

1.1 Background

With the maturity of Web 2.0 concepts and technologies, the social media, represented by Facebook, Twitter, and Sina Weibo, has been gradually recognized and applied by people as an essential platform for sharing opinions, interests, and experiences, changing their traditional lifestyle [37]. The internal structure of social media is called social network, which is famous for its large number of users, a wide range of data types and exponential information dissemination speed [13].

Twitter, proved to be both the online social network (OSN) and the news media [21], plays an important role in the hybrid spreading of information between society and cyberspace, especially for the breaking news [15]. For instance, on the day of the 2016 U.S. presidential election, more than 40 million election-related tweets were sent by 10 p.m. (Eastern Time) [38]. The Twitter web interface also displays a list of trending topics on a sidebar on the home page, helping the users understand what is happening among the whole world and what people′s opinions are about it, which could be illustrated in figure 1.1 below. At the time when people take advantage of it to transmit the messages, how to control the direction, speed, and range of propagation becomes the most worthwhile topic. When analyzing the information diffusion phenomenon on Twitter, it is always represented as one complex network graph of interactions within a group of individuals.

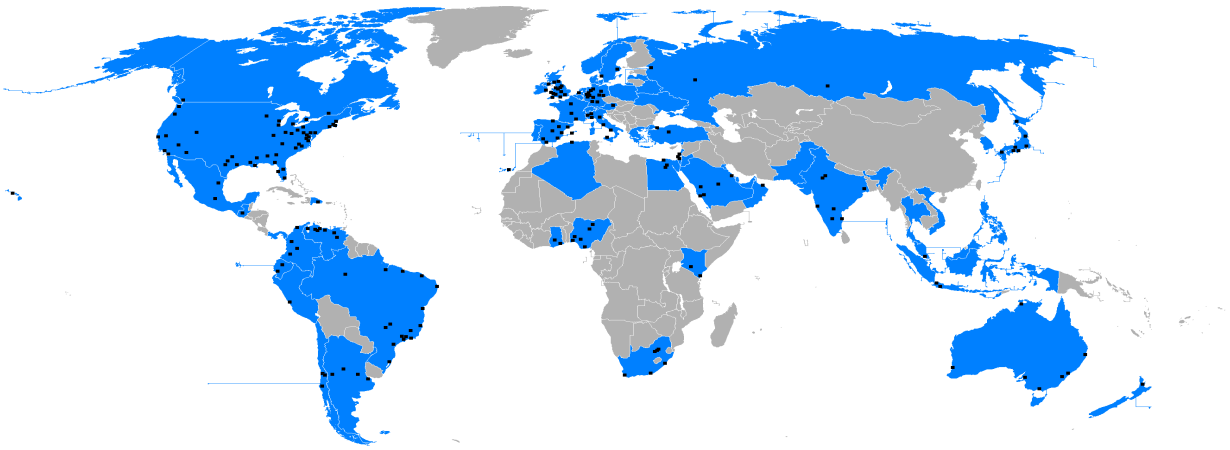

Figure 1.1: Countries and cities with local trending topics on Twitter [41]

Complex networks theories are now widely utilized as one medium to study the various propagation processes from the disease transmission among the humanity [33], message and rumour dissemination on OSN websites [5], to universal advertising behaviour within contemporary enormous marketing [26]. The changes of the topological diagram of networks resulted from the different propagation situation can intuitively illustrate the spreading speed and coverage, which makes a significant contribution to analyze how several factors influence the spreading process in each event, such as the initial seeds setting, the integrated topology properties and the top influential nodes [12]. Through this, some general knowledge can be generated and may provide practical instructions on the measures to control the disease spread into a smaller range [35] or oppositely help determine marketing strategies for the merchants to enhance advertising influence on more customers [19].

The most popular research topic related to information spread control on social networks recently is the influence maximization problem in complex networks [19], referring to how to choose the set of initial Twitter users as propagation seeds to make final information coverage the widest [16]. Meanwhile, several other research topics in this area are also significantly important, such as finding highly influential users [8] and how the properties of the Twitter network topology affect the propagation situation [25]. Compared with conventional methods of controlling information dissemination like shielding keywords based on the event itself [24], user-based network factors control are universal and more effective. After understanding the reciprocal effects of all the factors on information spreading, Twitter can become better utilized for product promotion by enterprises [28] and even for announcement release by political figures.

1.2 Purpose and Research Questions

As a start, this report will establish a directed complex network topology based on more than 70-thousand real existing users of Twitter and their interactive relationship, which will be identified by whether they follow the other users or not. Compared with some topological graphs created by software [39] or some incomplete network topology intercepted by snowball sampling method with very few users [10][22] in previous researches, the 70-thousand-user network can provide an integrated and realistic simulation platform, better representing the nature of Twitter social circles.

Meanwhile, Liyi [45] generates an infection rate prediction model through training XGBoost, with a relatively high accuracy of above 70%. When conducted an information spread simulation on Twitter, nearly all the predecessors just merely set an infection probability of whether the users posting the message would infect their followers, which may result in random and inaccurate results [25]. Compared with this above method, the model trained according to the propagation process of the real event on Twitter can better reflect the information transmission mechanism within Twitter itself, separate from any other networks.

Then, with the considerable part of the existing Twitter network topology and the Twitter-specific predictive model, the simulation process on information spreading can be conducted. Based on the simulation results in different time periods after the information starts to spread, the dissemination speed and coverage under various context can be analyzed and compared, for example, the different seeds with disparate features and the different chronological sequence in which the seeds appear.

This research aims at providing the theoretical bases and data support for the prediction and effective control of the messages propagation results on Twitter based on the seeds settings. Therefore, three central following research questions are raised to study the information spread speed and coverage results, including the criteria of seeds selection, the different order that seeds appear and whether the seed users active or not despite the same determined 20 seeds.

1.2.1 Research Question 1: What features should base on to select seeds for influence maximization?

When social networks become the primary way for people to disseminate information contemporarily, within the business field, merchants always choose several users to help them publish promotional information initially, achieving the purpose of publicity and sales expansion. The same situation considered in complex networks is how to choose a set of initial nodes to begin the spreading can make the infected coverage maximum. The influential level of each node in the entire network may be determined by in- or out-degree, betweenness centrality, and eigenvector centrality, which are defined in section 2.1.4.

All three properties regarding seeds stated above will be evaluated in this report to determine better seeds selection guidance. A straightforward way is to choose five seeds respectively with the best value of these evaluation criteria and then simulate the diffusion initialized by these seeds respectively. Then, in specific time periods, the results can be compared by the final coverage and the widest one would indicate the best seeds determination manner.

1.2.2 Research Question 2: How single-stage and sequential seeding strategy affect the information diffusion process?

The wide-spread communication events on OSNs originated only from the information release of several independent individuals, but can eventually become known to millions. Usually, these seeds, although appearing at early stages, always have an inherent order. Then people will think that if seeds release the news together just after affairs happen, what changes will bring about.

In this research question, many investigations can be conducted. Firstly, the sequential seeding strategies will be divided into two types: in a small-to-large or large-to-small order according to their followers′ amount for the same seeds. Meanwhile, their appearance will be set in a consistent time interval. Then, the relatively best sequential strategy will be resolved after comparing their performance.

Secondly, the best static sequential manner will be compared with the dynamic sequential one. The static one means ranking the followers′ amount of all users within the whole net-

work in the beginning while dynamic one means ranking this value only in the spreading network topology only considering the remaining uninfected nodes.

For the third step, the single-stage seeding pattern will be compared with the sequential ones mentioned above to generate a better strategy. Also, after these 3 procedures, one can make a conclusion about the best seeding strategy.

1.2.3 Research Question 3: How does the activity level of seeds affect the spreading results?

Taking the realistic factors into account, the influence of seeds selection on communication depends not only on the topological characteristics of users in the network but also on the degree of activity of themselves. In this research, this will be tested by changing the activity level for the same seeds which is reflected in the model′s features. Through this, although their sequence of appearance is the same, the final performance will vary from each other, contributing to useful guidance for seeds selection in information diffusion context regarding the users′ different range of activity degree.

1.2.4 Research Hypotheses

There are also three hypotheses derived from the above questions listed to be examined.

Hypothesis 1 Seeds selection based on betweenness centrality will result in a broader final coverage than on degree or eigenvector centrality.

Based on the definitions of the centrality of the nodes stated in section 2.1.4, one node with a higher eigencentrality will be connected to many nodes with high scores themselves [34]. Such users, together with the users with high degrees, can better access information from other nodes and have a faster and more direct impact on others in the same community [14]. Furthermore, one can imagine that a node with the higher betweenness may have a considerable impact on the network due to its significant advantage in controlling the information passed between individuals or different communities. Due to their locations on the most paths occupied by the message, they are also important factors disrupting communications between other users when removing them from the network [6]. Therefore, compared with degree and eigencentrality features, seeds with a higher betweenness centrality may finally active more nodes throughout different communities.

Hypothesis 2 ***Sequential seeding strategy can better trade-off between diffusion speed and coverage than single-stage one; meanwhile, the dynamic sequential strategy will generate the best performance.***

In this paper, the judgment model of infection is trained by the propagation network in which the seeds appear in different orders. Therefore, there will be better performance in the sequential seeding strategy. Another assumption made on this topic is the seeds appearing in large-to-small degree order will perform better than in the opposite order, which may result from more nodes infected by the high-degree seed at the early stage.

When talked about the dynamic seeding strategy, delayed appearance of several seeds can better reflect the natural propagation processes. Meanwhile, generating the new seeds depending on the topological structure changes of the network can somehow avoid selection of seeds with high potential to be infected at early stages and may lead to the information spreading to several local communities [17].

Hypothesis 3 ***When the seeds become inactive, their spread influence around the topics will drop considerably.***

In the information dissemination network, the activity level is always be defined as the degree to which the user responds to related messages, measured by the ratio between the number of messages he/she spreads and the number of messages received. In the egocentric social networks, the seeds play a key role in spreading, and their activity degree can reflect the level of communication with other netizens at ordinary times and the level of conviction of their published messages. Therefore, high activity degree can lead to better information dissemination impact finally [10].

While these assumptions may seem reasonable, the real-world behaviour is difficult to predict and control, especially when dealing with thousands of agents. Their behavioural prediction model is analyzed and generated from historical data, and their behavioural feedback may loop in reality, which may bring about a significant influence on the possible candidates. Therefore, all of the above hypotheses will be tested respectively through multiple simulation analysis.

1.3 Contributions

1.3.1 New network topology

As is mentioned in section 1.2, when the predecessors study the propagation properties of information on OSNs like Twitter, the general methods to generate the network topology for simulation were node sampling, link sampling, and snowball sampling [22]. The former two methods could not even guarantee the relatively strong connectivity of the topology by just simply choosing a fraction of nodes or links, which was considered unfavourable for studying the propagation mechanism. While the latter one could continually seek its several later generations from one determined node, which does not take the descendants of other points into account, especially those in the final generations. This method will also lead to incomplete network, which seems the direction of communication is limited to some extent when studying information dissemination [10]. Moreover, what is even less reliable is that some people use software to simulate such social networks, which lacks much practical significance and theoretical support when analyzing the influence of the topological properties on information spreading [39].

Unlike the above unconvincing methods, the topology network utilized in this research is generated from various complete topic circles. This method can not only ensure the network connectivity to a large extent but also guarantee the local community networks which start with any individual user. Therefore, this report utilizes a total of more than 70,000 users within various social circles on Twitter to establish a relatively complete network topology. Moreover, after the topological properties analysis for both the Twitter and the generated one, the similar results stated in section 3.1 prove it can well reflect the epitome of information dissemination across the entire Twitter network.

1.3.2 New infection rate predictive model

In traditional research algorithms, one would set a fixed percentage like 5% as the infection rate within each spreading iteration when utilizing the epidemic models [16]. Such an approach can not reflect the unique patterns of information dissemination on Twitter, on which the number of infected users will significantly vary in different situations. Therefore, when studying some topics such as the influence of the order in which seeds appear on message diffusion, the results will also have a specific range of deviation from reality.

In order to better reflect the spread characteristics on Twitter, a predictive model specified in whether any unique user is infected or not at a certain moment is generated and trained by Liyi [45]. In this way, the actual spread situation of the event on Twitter, including when it breaks out and when it fades, can be readily observed and tracked.

1.3.3 New Simulator

Generally, the information dissemination simulator contains the following parts: an integrated spreading topology, the dynamic propagation model, seeds settings, the infection rate and the time interval for iterations. Within the unique simulator utilized in this research, the topology and the infection rate setting have been explained in the former two sections. Meanwhile, the dynamic model is the susceptible-infected epidemic model, which indicates that once a user is infected with relevant information, he/she will never recover to the uninfected state. Moreover, this user will continue to keep in touch with susceptible people and may transmit information to them.

When it comes to the seeds selection, some researchers just initially set a percentage to choose a fraction of nodes as the seeds randomly [16]. However, because this research will further explore the diffusion influence resulted from the properties of the seeds both in topological and personal, the specific seeds will be carefully chosen according to different requirements and their parameters of related characteristics will also be adjusted to achieve the purpose of comparative analysis.

One last point worth mentioning about the unique simulator is the time interval for iterations, which is also a significant difference from other previous simulators. Traditionally, one can freely choose the time interval during each iteration to adapt to the simulation complexity. In this research, it utilizes the model which is trained by the datasets containing the records that whether a unique user will be infected or not every 30 minutes within two days. Therefore, in order to guarantee the accuracy of the model, the time interval for spreading iterations is set to 30 minutes.

1.3.4 New research sub-topic

One part of this article has studied a new idea when analyzing the sequential seeding

strategies, which is how the appearance order associated with the degree values for the same seeds influence the information diffusion results. Starting with the same seeds for sequential strategy is divided into 2 cases, from the seeds with higher degree values to lower ones or from the lower to higher ones.

Unlike the traditional researches about the effects of the order in which seeds appear on the information transmissions, they usually focus only on whether the seeds appear at the same time performs better than their separate occurrences [18][25], with no much attention on which specific order would be better when adopting the sequential strategy.

Meanwhile, this report also analyzes seeds selection based on which characteristics could achieve the influence maximization within the network, and the conclusion is that for the OSNs such as Twitter, it is better to choose the nodes with the highest degree values. Therefore, when comparing the different cases regarding the appearance sequence for the same seeds, their order depending on their degree was then taken into account, which would also the origin and research significance of this new research topic.

1.3.5 New research method on one sub-topic

When analyzing how the spreading results would be affected by the activity level of the seed or other specific nodes, the predecessors usually directly remove these nodes from the network to observe the changes from the original results [10]. Through this method, it can only study the unique role of these nodes playing in the spread with no reflection on the changes brought about by their activity interval. Meanwhile, due to the removal of some edges connected with these nodes, it also changes the overall network topology to some extent, resulting in the non-negligible one-sidedness and limitations of conclusions.

In this report, benefiting from the simulator and the infection rate predictive model, the features of all nodes within the entire network required to be constantly updated throughout the propagation process. Moreover, the values that reflect the nodes activity have a significant effect on the predictive model, demonstrating its essential position in the process of information dissemination. Therefore, the spread influence caused by the activity level of the seeds can be analyzed by changing the parameters associated with this aspect. More detailed information related to how to adjust these features can be seen in section 4.3.

1.3.6 New seeding strategy generation

When the predecessors studied the influence of seeds on the information dissemination on OSNs, most of them generally analyzed one or two aspects separately. For example, they studied the seeds selection related to their centrality [25], the better one in single-stage and sequential seeding strategy [18] and the influence caused by seeds activity [10]. In this research, all the topics mentioned above were examined by a simulation approach. After generating the separate conclusions for these unique topics, how to combine these results to provide more integrated theoretical instructions on the seeding strategy has become a meaningful topic. Therefore, this research came up with a new seeding model by utilizing the best situation within above aspects.

In this model, it utilizes the dynamic sequential seeding idea with the seeds selected dynamically with the highest degree and relatively high activity level. This combined model could spread the related information to the widest coverage compared with the previously separated strategies, providing a new idea for seeding strategy in the future practice.

Chapter 2

LITERATURE REVIEW

2.1 Complex Networks Theory in Social Networks

Within the field of network research, the complex network is defined as the sufficiently complicated topological network structure formed by a large number of nodes and their intricate relationship with each other. No matter in medical science, economic science or information science, there exist such complex network structures [11]. Constructing complex networks is also an effective approach for researchers to study how OSNs like Twitter evolve through mining interactions among a substantial amount of people. Although by no means exhaustive, several primary observational attributes for OSNs are stated below [5].

2.1.1 Large-scale Property

OSNs are representatives of complex networks with the amount of the nodes often in millions and edges often 100 times more than nodes. For instances, 335 million active users communicate on Twitter, and even 2.2 billion users are active on Facebook monthly [42]. When some explosive news occurs, countless users will be continuously involved in the discussion. Moreover, some of Twitter′s nodes correspond to celebrities, including Barack Obama and Justin Bieber, owning more than 100-million followers [32].

2.1.2 Small-world Property

The small-world networks are identified by a short path length of $O(logn)$ between any two nodes within the network and a higher clustering coefficient than a binomial random

graph with the same number of nodes and the same average degree [3]. According to the Twitter statistics collected in 2017, the 90-percentile effective diameter was 5.91 edges, and the 0.59 clustering coefficient was reported [1][2].

2.1.3 Scale-free Property

The scale-free networks are characterized by a power law of degree distribution pattern [4]. This pattern has been proved for both in-degree and out-degree distributions in YouTube, LiveJournal, and Flickr [29], as well as in Twitter [2][21].

2.1.4 Nodes Centrality Property

In this report, the centrality rankings are examined. The betweenness centrality measuring the extent to which a vertex lies on paths between any other vertices within the network can be represented below [7][31]:

$$C_B(i) = \sum_{s,t} w^i_{s,t} = \sum_{s,t} \frac{n^i_{s,t}}{n_{s,t}} \tag{2.1}$$

Meanwhile, the closeness centrality Closeness centrality measures the average distance from one node to other vertices, which can be calculated by the following equation [31]:

$$C_C(i) = \frac{1}{l_i} = \frac{n}{\sum_j d_{i,j}} \tag{2.2}$$

Furthermore, the eigenvector centrality can measure the influence of each node in the network, which is similar to the Google′s PageRank algorithm. The following equation shows the relative score for this centrality: [34]:

$$x_i = \frac{1}{\lambda} \sum_{j \in M(i)} x_j = \frac{1}{\lambda} \sum_{j \in G} a_{i,j} x_j \tag{2.3}$$

where $M(i)$ is the neighbors set of i and λ is a constant. And in matrix notation with the adjacency matrix A $\mathbf{x} = x(i_1), ..., x(i_n)$ this yields:

$$A\mathbf{x} = \lambda \mathbf{x} \tag{2.4}$$

2.2 Dynamical Models for Spreading

The process of diffusion always follows unique rules in different fields, and the predecessors have concluded several spreading models.

2.2.1 Linear Threshold Models

This model type states that whether or not the node status transforms is based on the fraction of neighbours′ status compared with the threshold value [36]. A simple example with the probability of 0.5 is shown in graph 2.1.

2.2.2 Independent Cascade Models

The independent cascade models indicate that one inactive node will be activated by its neighbor who has already been active in an arbitrary order with a determined probability [30]. A simple example with the probability of 0.5 is shown in graph 2.2.

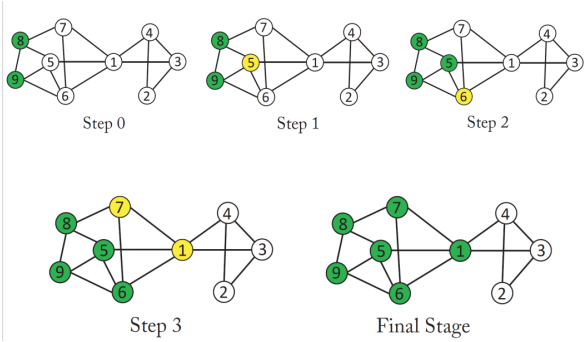

Figure 2.1: Linear threshold spread process

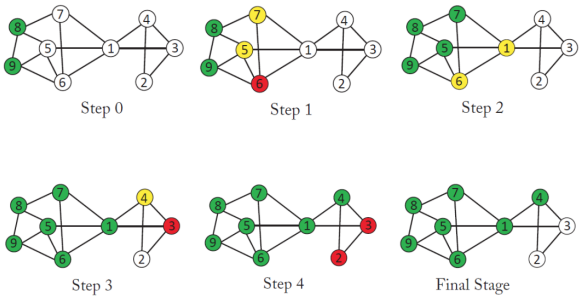

Figure 2.2: Cascade models spread process

2.2.3 Epidemic Models

The classic epidemic models like SI and SIR are the most widely used mathematical models for illustrating information diffusion. In these models, people in different compartments are assigned to the same state including the susceptible state (S: unaware of the information and would be infected later), the infected state (I: already aware of this information and may spread to others) and the recovered state (R: already aware of but not infected by this information and will not spread to others any more) [44]. Moreover, there is a certain probability for one state to convert into another and the schematic diagram of the state transition process is shown in figure 2.3 below.

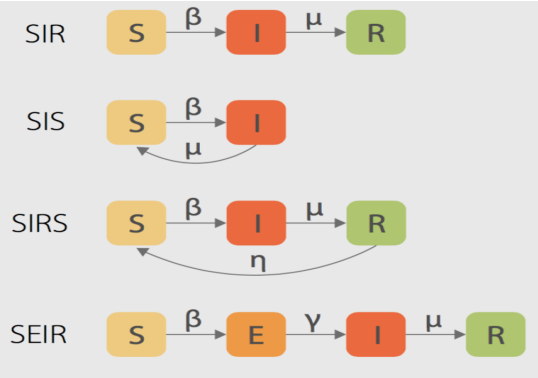

Figure 2.3: Classic epidemic models [33]

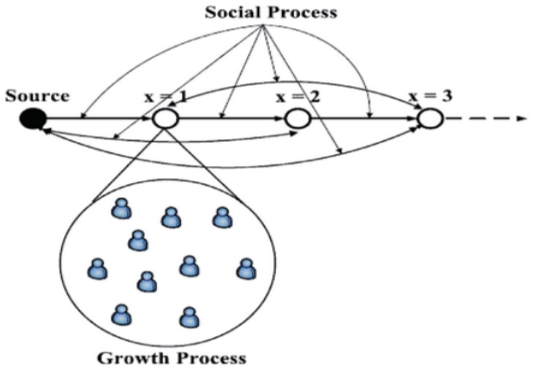

Figure 2.4: Logistic models spread

2.2.4 Other Models

There also exist several other models like linear influence model and diffusive logistic model. Firstly, the linear one is created based on many empirical analyses on Twitter. The main assumption is that the total number of newly infected nodes depends on the influential degree of nodes infected in the last time period, which can be expressed as following:

$$V(t+1) = \sum_{u \in A(t)} I_u(t - t_u) \tag{2.5}$$

where $V(t)$: the number of infected nodes at time t; $A(t)$: the overall infected nodes set; and I_u: the influential function of node u got infected at time t_u.

Moreover, the logistic model divides the diffusion process into the growth and the social stages (see figure 2.4), concerning about both the temporal and spatial dimensions. For more details, the growth stage handling the nodes at the same distance from the propagation source while the social stage deals with the nodes with different distances [40].

2.3 Seeds Influence on Information Spread

When talked about the influence of the seeds settings on the information spread, there are 3 main aspects: the selection of the seeds based on their topological properties within the overall network, the seeds′ time of appearance and whether the seeds are active or not. Searching for the optimal seeding strategy is generally an NP-hard problem; as a result, greedy algorithms and other approximate heuristic strategies are utilized to search

for effective solutions [25].

Firstly, as long as conducting the simulation experiments for approximate solutions on information influence maximization in many applications, the selection of seeds is an unavoidable factor. When handling the large-scale network, a set of seeds are chosen through centrality analysis based on their topological attributes like in-degree or out-degree, betweenness, closeness, and K-shell [25]. In the former study, the spreading process is initialized by several highly influential nodes with relatively high centrality [25].

Secondly, many researchers have compared single-stage seeding strategy that means all the seeds appear together at the beginning with sequential seeding strategy that means the seeds appear in a given order. The available results indicate that the single-stage strategy can bring about a higher spreading speed in the early stage while the sequential one can trade off between the spread speed and the coverage, finally making more nodes active and giving a broader coverage [18][25]. When implementing the sequential approach, another advanced way is to add the seed by dynamic rankings of nodes. This way is realized by recalculation of network measurements during the diffusion process, with only non-infected nodes taken into consideration, which can contribute to more coverage in 90% cases [16].

Furthermore, when studying the influence caused by whether the seeds or the top n most engaged users (and they are the direct followers of the seeds) active or not, the results show that whether making the seeds inactive or making the top n inactive, the spread coverage will decrease obviously. Moreover, the seed has more influence than the top 10 influencers but less than the top 100 influencers and random 100 users over time. Also meanwhile, the Top 100 series dominates the others [10].

2.4 Topology Influence on Information Spread

Another interesting topic concerning the influence on the spread caused by the network topological properties mainly studied three aspects: the degree distribution, the network density, and the assortativity coefficient.

Firstly, it studied the influence of the power-low exponent when thinking about the degree distribution. As the exponent increases, the spread coverage will decrease subsequently.

Secondly, to measure the network density, it took the average degree into consideration. The result shows that the coverage will increase as the density increases [39].

Furthermore, the coverage of active nodes decreases dramatically as the assortativity coefficient becomes larger. The underlying reason may be that as the assortativity coefficient increases, the selected seeds at the beginning based on degree rankings may take more steps to reach the nodes with small degree [25].

Chapter 3

METHODOLOGY

3.1 Datasets

3.1.1 Network Topology for Simulation

An integrated Twitter network is considered as the best platform for simulating the information spreading process on the Twitter social media. However, according to the statistics collected in July of 2018, Twitter has 335 million active users, far beyond the computational capabilities of standard computers when executing such an intricate link structure. Meanwhile, under normal circumstances, the number of users involved in one topic or event is always below 50 thousand, never spreading throughout the whole Twitter network [24]. Therefore, it is enough to intercept social circles formed by over 70-thousand users on Twitter as the simulation network.

Dataset statistics	
Nodes	81306
Edges	1768149
Nodes in largest WCC	81306 (1.000)
Edges in largest WCC	1768149 (1.000)
Nodes in largest SCC	68413 (0.841)
Edges in largest SCC	1685163 (0.953)
Average clustering coefficient	0.5653
Number of triangles	13082506
Fraction of closed triangles	0.06415
Diameter (longest shortest path)	7
90-percentile effective diameter	4.5

Figure 3.1: Original network from SNAP

Dataset statistics	
Nodes	76245
Edges	1768149
Average Degree	21.875
Average Clustering Coefficient	0.501
Number of triangles	11890765
Diameter (longest shortest path)	14
90-percentile effective diameter	5.4
Average Path Length	10
Modularity with resolution	0.813
Number of Communities	336

Figure 3.2: New network after updating

The original network shown in figure 3.1 above is obtained from Stanford SNAP datasets [23], with 81,306 users and 1,768,149 directed edges between them, which indicates their relationship status of whether follow or not with each other [27]. However, this information comes from 6 years ago, many users have written off, and the relationship between them may have changed a lot. In order to better obtain the certain characteristics of the users within the entire network required by the prediction model, Twitter API is utilized to conduct the data crawling [20]. Then, the result indicates there are still 76,245 users remaining, and the relationship among them has been updated to 1,667,871 edges. As the network statistics are shown in figure 3.2, the diameter of the entire network is 14, and 90-percentile effective diameter is short with 5.4 edges, together with the relatively high average clustering coefficient, indicating the small-world property. Compared with the entire Twitter social network analyzed in 2017 stated in section 2.1.2, the similar statistics indicate this network can reflect the nature of the Twitter topology.

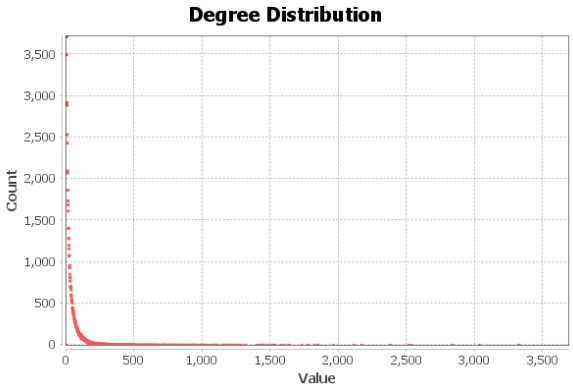

Figure 3.3: Degree distribution

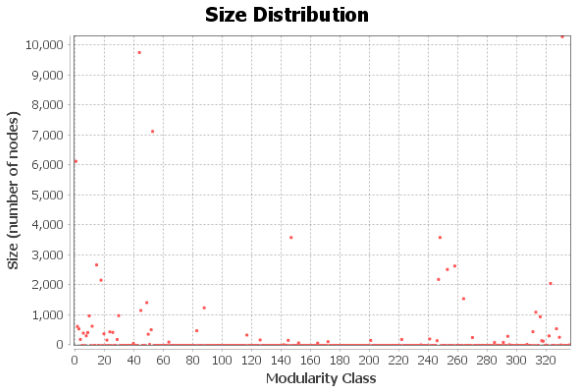

Figure 3.4: Communities size distribution

Meanwhile, this network also has the scale-free property. Its degree distribution follows the power law pattern shown in figure 3.3 with the average degree of 21.875, which indicates a few users in the network has a large number of followers. In reality, they represent the official media or the celebrities in various fields.

Another feature that should also be concentrated on is the communities distribution. After processing the whole network, there are 336 communities with the modularity value of 0.813. According to their size distribution above, most communities only have users less than 1,000 while only about three communities have nodes over 4,000. Taken information diffusion in such a network into consideration, if the seeds are selected in the same community, the information will be slower to spread to other community nodes. Therefore,

choosing the seeds scattered in various communities may accelerate the whole propagation procedure.

In conclusion, all the statistics mentioned above indicate this intercepted network from Twitter can provide a sustainable environment for the simulation stage to analyze the information spreading characteristics on Twitter or the important factors that affect the diffusion pattern.

3.1.2 Infection Rate Prediction Model

Usually, when researchers simulate the spreading process with the epidemic models, the infection probability is always set to 5%. With improvements to this stochastic method, the machine learning process is used as a tool for predicting whether the specific user would be infected by the information given the exact time. For the training datasets, the actual related information dissemination situation on Twitter within 48 hours just after Givenchy′s death has been restored to generate the relatively accurate model by Liyi [45]. The distribution of the original 55,559 records for around 6,000 nodes status of 1-infected or 0-uninfected over every 30 minutes are also illustrated in figure 3.6 below.

Generally, the XGBoost algorithm created by Tianqi Chen [9] is utilized to train the model, and the overall 88 features in figure 3.5 are selected after importance analysis. For a brief explanation, the features include the time information after the event occurred (TwM), the activity level of the user (UsM), the topological characteristics of the user in the static network (Nw) and the dynamic spreading network (SNw). Meanwhile, the features also extract the above information of those other users who may send messages to him/her (Nw and SNw ones with -1 or 0) and some statistical data like the maximum or the mean (Stat).

Moreover, this model will judge whether the user is infected or not every 30 minutes after the event occurs, which concerning more about the temporal factor with no information about the actual geographical distance. Therefore, in order to obtain more accurate predictions by this model, the features of each node would be updated every 30 minutes as well. More detailed information about features extraction during the simulation process will be given in the following sections.

UsM_deltaDays	UsM_deltaDays-1	Nw_inDegree-1	SNw_outDegreeCentrality
UsM_statusesCount	UsM_statusesCount-1	Nw_outDegree-1	SNw_outDegreeCentrality0
UsM_followersCount	UsM_followersCount-1	Nw_degreeSeed0	SNw_outDegreeCentrality-1
UsM_favouritesCount	UsM_favouritesCount-1	Nw_inDegreeSeed0	SNw_inDegreeCentralitySeed0
UsM_friendsCount	UsM_friendsCount-1	Nw_outDegreeSeed0	SNw_outDegreeCentralitySeed0
UsM_listedCount	UsM_listedCount-1	Nw_degreeSeed-1	SNw_inDegreeCentralitySeed-1
UsM_normalizedUserStatusesCount	UsM_normalizedUserStatusesCount-1	Nw_inDegreeSeed-1	SNw_outDegreeCentralitySeed-1
UsM_normalizedUserFollowersCount	UsM_normalizedUserFollowersCount-1	Nw_outDegreeSeed-1	Stat_average_kOut
UsM_normalizedUserFavouritesCount	UsM_normalizedUserFavouritesCount-1	SNw_nFriendsInfected	Stat_average_t
UsM_normalizedUserListedCount	UsM_normalizedUserListedCount-1	SNw_friendsInfectedRatio	Stat_average_deltaDays
UsM_normalizedUserFriendsCount	UsM_normalizedUserFriendsCount-1	SNw_generation	Stat_average_statusesCount
UsM_deltaDays0	TwM_t0	SNw_generation0	Stat_average_followersCount
UsM_statusesCount0	TwM_tSeed0	SNw_generation-1	Stat_average_favouritesCount
UsM_followersCount0	TwM_t-1	SNw_timeSinceSeed0	Stat_average_friendsCount
UsM_favouritesCount0	TwM_tSeed-1	SNw_timeSinceSeed-1	Stat_average_listedCount
UsM_friendsCount0	TwM_tCurrent	SNw_totalNodesInfected	Stat_average_normalizedUserStatusesCount
UsM_listedCount0	TwM_time	SNw_nodeInfectedCentrality	Stat_average_normalizedUserFollowersCount
UsM_normalizedUserStatusesCount0	Nw_degree	SNw_totalInDegree	Stat_average_normalizedUserFavouritesCount
UsM_normalizedUserFollowersCount0	Nw_degree0	SNw_totalOutDegree	Stat_average_normalizedUserListedCount
UsM_normalizedUserFavouritesCount0	Nw_inDegree0	SNw_inDegreeCentrality	Stat_average_normalizedUserFriendsCount
UsM_normalizedUserListedCount0	Nw_outDegree0	SNw_inDegreeCentrality0	Stat_max_kOut
UsM_normalizedUserFriendsCount0	Nw_degree-1	SNw_inDegreeCentrality-1	Stat_min_kOut

Figure 3.5: 88 features

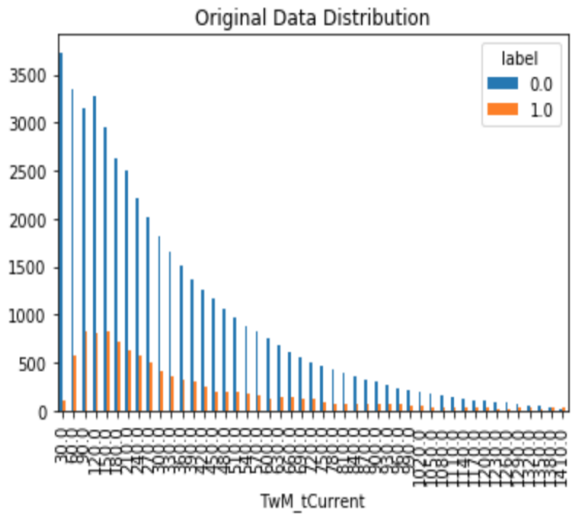

Figure 3.6: Nodes status distribution

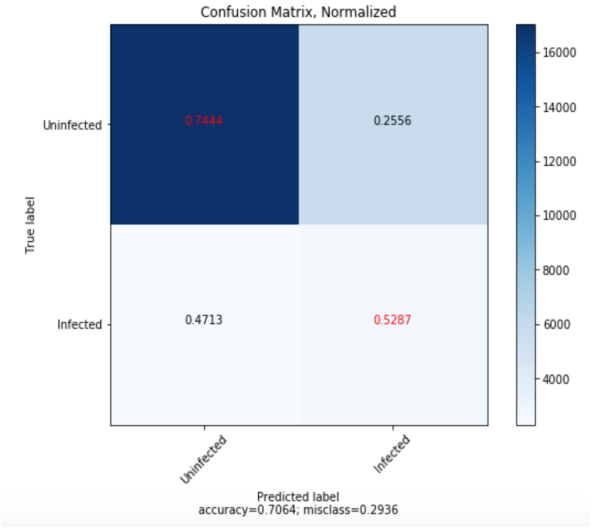

Figure 3.7: Normalized confusion matrix

Final model can give a high accuracy of over 70%, which can also be analyzed from the normalized confusion matrix in figure 3.7 above. When putting one record of node features into the model, it will give an infection rate between 0 and 1. In order to determine whether the user will get infected or not, the median 0.5 is set to be the critical point, suggesting that the information will infect the user with a predictive probability of over 0.5.

3.1.3 Givenchy's Death - Spreading on Twitter

Givenchy, one famous aristocratic French fashion designer, whose understated style represented romantic elegance, has died aged 91 on 10 March 2018 [43]. This message was

firstly posted on Twitter by his family and friends. After forwarded by several large media afterwards, the news then quickly and widely spread to more than 5-thousand users within 24 hours.

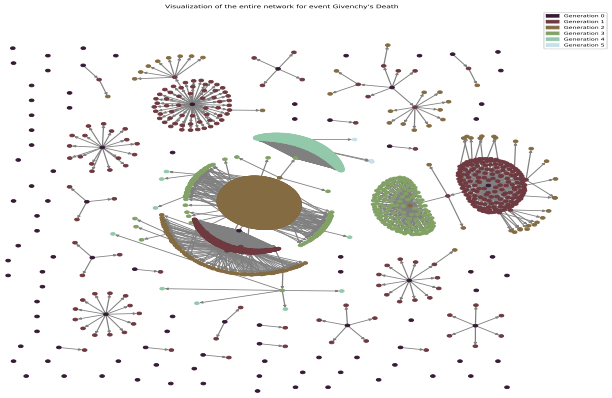

Figure 3.8: Entire spreading network

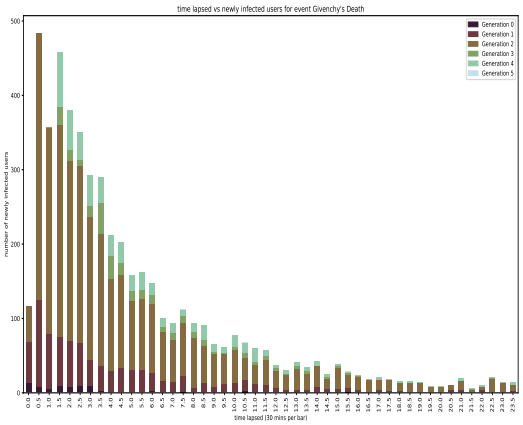

Figure 3.9: Newly infected users

According to the above spread network visualization and time-based newly infected users amount for this event, it is obvious that most infected users appeared in the third generation. This phenomenon indicates that although there were few seeds, after the retweet of several influential users in the second generation, the coverage of this message would suddenly expand since then.

Meanwhile, focusing on the time periods, it can be concluded that almost 70% of users got infected within the first 6 hours just after the occurrence of this event. Nevertheless, as time goes on, fewer and fewer users were influenced until this message was not discussed any more on the Twitter.

3.2 Procedure

3.2.1 Network Platform Establishment

Firstly, according to the dataset generated from part of Twitter containing the 76,245 users and their relationship with each other, a topological platform can be constructed for conducting the simulation. In this network, the individual users are set as nodes and their follow situation as edges $(i_1, j_1), (i_2, j_2), ..., (i_N, j_N)$, where an edge from i to j means j is a follower of i and N represents the overall edges amount of 1667871. However, the nodes′

strength and the edges′ weight are not measured in this research. Therefore, the simulation network platform is established with an unweighted directed network model.

Afterwards, based on the transforming algorithm above, Gephi helps to draw the topology graph and calculate the information including network diameter, average clustering coefficient, degree distribution and centrality distribution, which can also be sorted to select seeds subsequently. After comparing these statistics with the whole Twitter′s reported, this network is determined appropriate to utilize.

3.2.2 Simulator Construction

A simulator with strong operability should be attached most importance in this research. It is created in Python with *networkx* and some other packages, including one input system for network topology and seeds information, one time-based iterated system for updating the nodes′ dynamic features and spreading records every 30 minutes and one output system for illustration of the whole information propagation. Then, with this simulation framework, the different seeding strategies could be implemented and analyzed to make an in-depth comparison.

3.2.3 Main Simulation Methods

Although analyzing different directions of seeding strategies like properties and sequence, The experimental idea of controlling variables is always the main research method utilized. For example, when analyzing seeds selection based on what features is better, though seeds are different, the number of them and their appearance time are set to be the same by single-stage strategy. Moreover, when comparing the single-stage with the sequential strategy, the seeds are set to be the same series of users. Only through this the comparison of the spread speed and the final coverage will become meaningful. Finally, their results will be illustrated in one statistical diagram to analyze the advantages or shortcomings of different strategies.

3.2.4 Combined Seeding Model Generation

Finally, to give better theoretical instructions on the seeding strategy, this research will

generate a relation formula regarding different aspects mentioned in the research questions to reflect their intertwined impact on the information diffusion results. Meanwhile, the weights of them can be set by parameters optimization, better reflecting their different influence degree. Though this, the network problem can also be transformed into a mathematical form, handling this problem more statically and logically.

3.3 Tools and Algorithms

Throughout the whole research process, the following tools and algorithms are utilized.

Firstly, Gephi is efficiently utilized in network representation and topological statistics calculation, which is essential for handling complex networks.

Secondly, Python 3.6 is the main programming language with many convenient packages, including, for example, the $'networkx'$ dealing with the complex networks problems and the $'matplotlib'$ for drawing the diagrams.

Algorithm 1 Sequential Seeding Strategy with Static Degree Ranking (Large-to-Small)

Input: Network $G(V, E)$; PredictiveModel M; UsersWithFeatures $V(I)$; ProbabilitySetting p, SeedsSet S; SeedsAppearanceTime t^{**}
Output: InfectedNodes
/** Seeds are ranked by degree in the descending order **/
Initialize: time $t = 0$; timeInterval=t^*;

```
function SIMULATION(G(V,E), V(I))
    InfectedNodes = Seeds[t**]∈ S      /** (By changing InfectedTime t→ t**) **/
    for t = 0 → T by t* do
        for InfectedNodes[t] ∈ V(I) do
            for each follower j of InfectedNodes[t] do
                if j is uninfected then
                    V(j) = updating V(j)
                    InfectionRate r = M.predict(V(j))
                    if r > p then
                        InfectedNodes = InfectedNodes ∪ j
                    end if
                end if
            end for
        end for
    end for
    return InfectedNodes
end function
```

The simulator created almost by time-based iterations and state judgments; moreover, one algorithm example for sequential seeding strategy with static degree rankings from large to small is illustrated above.

Last but not least, the computational complexity required to process this enormous network using time-based iterations is too high. Therefore, all the code files are run on Amazon Web Service (AWS), which can help save plenty of time. Even so, it still needs 10 hours to receive the information dissemination results of the entire network in each case.

Chapter 4

RESULTS AND ANALYSIS

4.1 Influence from Seeds Selection Criteria

4.1.1 Degree, Betweenness and Eigenvector Centrality

When talked about what features should base on to select the seeds for influence maximization within a complex network, in this report, three regular measurements are examined and compared, including the degree, the betweenness, and the eigencentrality. In order to conduct this test, after generating all these above required statistics and sorting them in in the descending order, top 5 nodes with the highest values in each type are selected as the seeds to start their simulation respectively. Meanwhile, the single-stage seeding strategy is utilized and the overall evaluation time interval should be only from 0 to 150 minutes because of the limited computing capacity. Then all the results in 3 condition are clearly illustrated below.

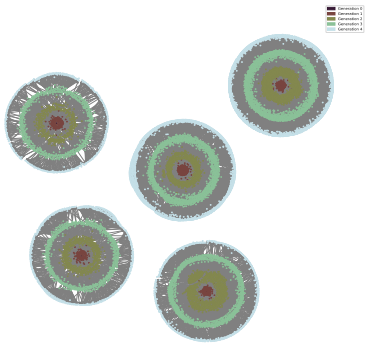

Figure 4.1: Degree_top5(a)

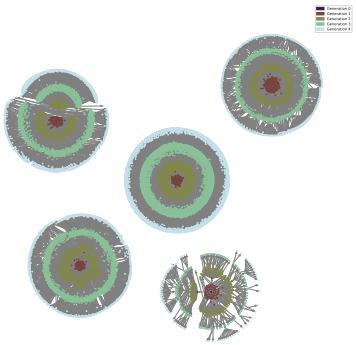

Figure 4.2: Betweenness(a)

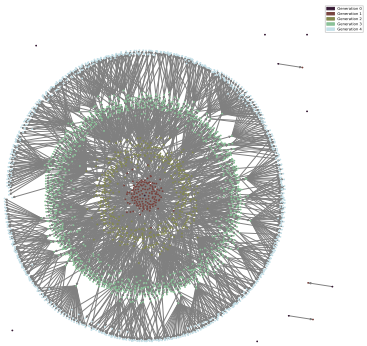

Figure 4.3: Eigencentrality(a)

Firstly, according to the different topologies shown from figure 4.1 to 4.2, the final spreading network with the highest degree seeds seems similar to the one with the highest betweenness seeds. However, it is obvious that there is a community in figure 4.3 with the spread situation inside which looks a lot sparse. Although it starts with the most connective nodes between individuals or communities, the topological result indicates it may just reach the maximum within each community rather than becoming a bridge between different communities. Such a result may be due to the structural topology itself, of which the large topic circles may have no nodes to connect between each other.

However, the spreading result with the highest eigencentrality seeds in figure 4.3 looks a significantly different pattern from the other 2 cases. The five seeds selected according to their importance within the whole network are all located at the center of the network and constantly spread outward to form an annual-ring-like topology.

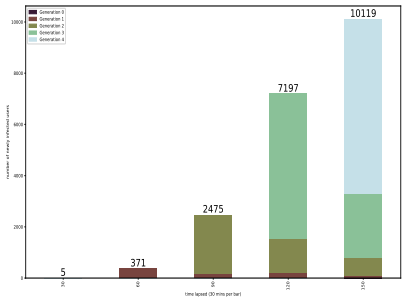

Figure 4.4: Degree_top5(b)

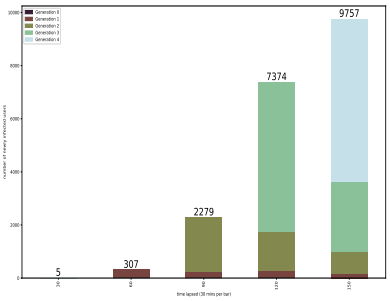

Figure 4.5: Betweenness(b)

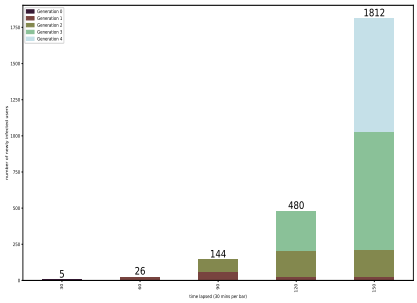

Figure 4.6: Eigencentrality(b)

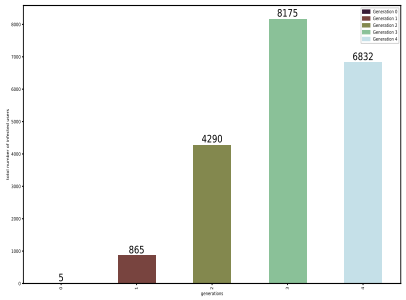

Figure 4.7: Degree_top5(c)

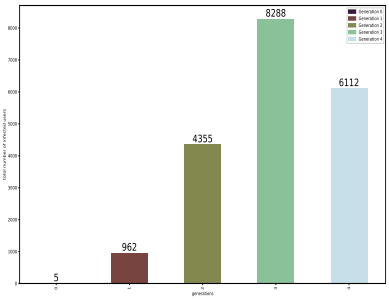

Figure 4.8: Betweenness(c)

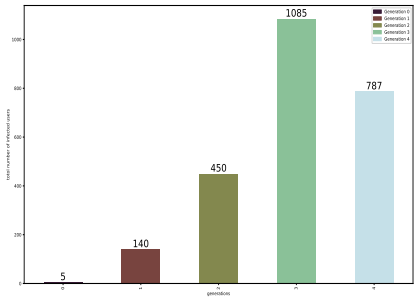

Figure 4.9: Eigencentrality(c)

Then, according to the above statistical histograms for all results of 3 cases from 4.4 to 4.9, the newly infected nodes within different time stages can be clearly displayed, followed by the total infected users of each generation. Within the same time horizon, the degree one

and the betweenness one spread the information to 20,167 and 19,722 users respectively while the eigencentrality one can only spread to overall 2,467 users. It can be clearly seen that the spread coverage of the former two types has expanded dramatically from about 90 minutes while the other one suddenly extends to a wider range just from 150 minutes. Meanwhile, from the height comparison of the histograms in the first two measurements, it can be concluded that the newly infected users are almost exponentially related to the time changes, and their growth rates in each time sub-interval are also consistent between these two cases.

Moreover, an interesting phenomenon regarding the total infected users in each generation is that all three results show the same pattern. They have been continually increasing in the first three generations and have gradually declined since the fourth generation. The reason for this manner should result from the accurate judgment of the model, which is based on the reality of the propagation fading of the related topic after a certain period of time.

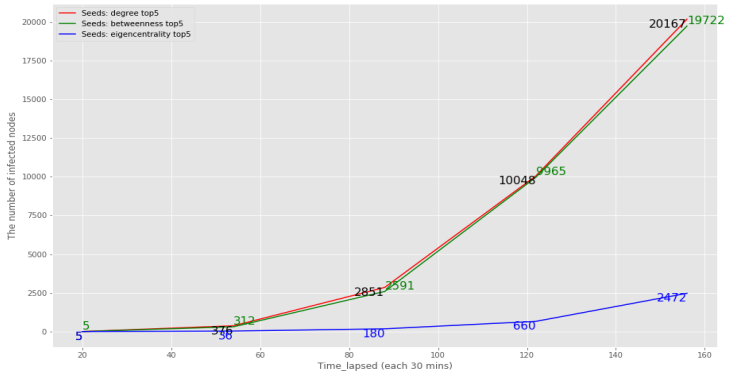

Figure 4.10: Total infected nodes based on different seeds selection

Finally, through presenting three results in the same line chart in the figure 4.10, the spreading speed and the coverage can better be compared. Whether based on the spread speed and the coverage, the information diffusion results started from the seeds with the highest degree and betweenness apparently show better performance than another. Although the spreading speed of the previous two cases was relatively consistent within the first 120 minutes, however, the degree one finally reaches the largest coverage.

Now transferring the attention to the first hypothesis stated in section 1.2.4: *seeds se-*

lection based on betweenness centrality will result in a wider final coverage than on degree or eigenvector centrality.

After the test through the above simulations, this hypothesis has been proved to be a one-sided account. In the spreading process on Twitter, although the diffusion started with the seeds with the highest betweenness centrality can reach more people than that with the highest eigenvector centrality, it finally influenced fewer users than that of the greatest degree. The underlying reason may be the unique topological structure of the data sampling from various topic circles, between which may have fewer connections to make the betweenness property play its original role.

In conclusion, in order to achieve the influence maximization within one network, the seeds selection whether based on the highest degree or based on the betweenness can achieve a balance between the speed and the coverage. Moreover, starting with the nodes with the largest out-degree, which also means the users with the most followers, can finally spread the information to more people. Therefore, selecting the seeds based on the largest out-degree value would be considered as the best strategy among them.

4.2 Influence from Seeds Appearance Sequence

4.2.1 Sequential Strategies Based on Degree Size

After having learned that utilizing the maximum out-degree value to select seeds is the best strategy, there is a further topic worthy to be explored, which is the best order for their appearance in the sequential seeding strategy. In this report, two orders of simulation on this topic are tested including seeds release in order of degree values from small to large and inversely from large to small. At the beginning of the information spread process, 20 seeds are randomly chosen from the network. For the simulation with the ascending manner, it releases 5 new seeds every half hour with degree values from small to large. Meanwhile, the other case is conducted the same way in the descending order. Moreover, the evaluation time interval for these two cases is from 0 to 210 minutes.

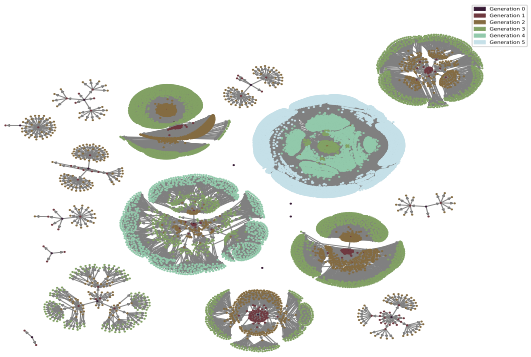

Figure 4.11: Degree low to high (a)

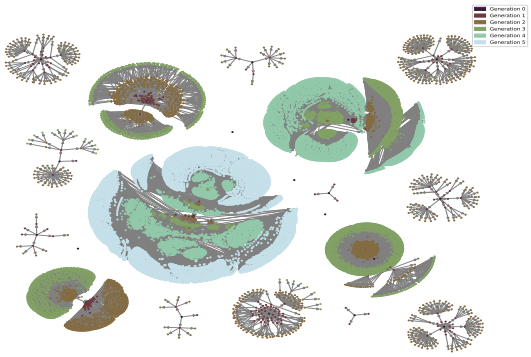

Figure 4.12: Degree high to low (a)

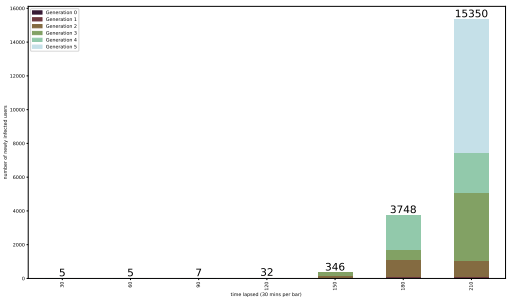

Figure 4.13: Degree low to high (b)

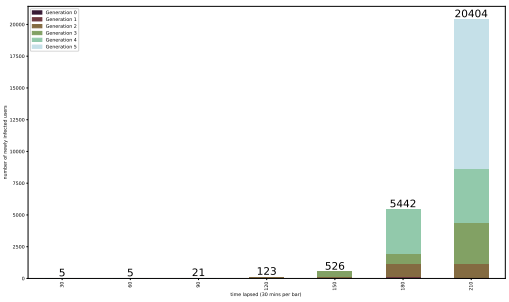

Figure 4.14: Degree high to low (b)

According to the figure 4.11 and 4.12, although the overall 20 seeds in the two simulations are the same, the final spreading topological networks resulted from different seeding sequences look a little different in communities size.

When moving to figure 4.13 and 4.14 showed the newly infected users over time, two cases display the same tendency. The information did not begin to spread on a large scale until 3 hours later. Moreover, the number of new infections began to show an exponential growth trend since then. The growth rate from 180 to 210 minutes indicates 310% in the descending order while 275% for the other situation. Within the 210 minutes after the spreading start, although the case of seeds released in order of degree values from large to small finally achieves a wider infection range of 26,526 than the ascending one of 19,493, the final results may change if the evaluation time interval is extended according to the growth rate comparison above. However, according to the small base of the ascending strategy and the current evaluation time, the descending seeding strategy performed better.

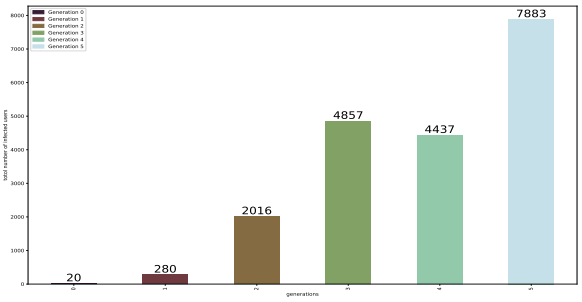

Figure 4.15: Degree low to high (c)

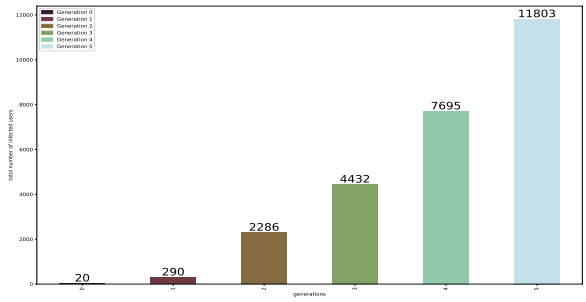

Figure 4.16: Degree high to low (c)

When taken the total number of infected users in each generation shown in figure 4.15 and 4.16 into consideration, these two types display the disparate patterns. The descending one exhibits a slight exponential rise while the other case belongs to the form of the increase with fluctuations. The number of infections in the first four generations tends to be consistent; however, just because in the fifth generation, the descending seeding one has nearly doubled the number of users infected with ascending seeding, resulting in a gap between the infections in the sixth generation and even in the final total. Therefore, only through the descending seeding strategy, the continuously expansive propagation of the information can be achieved.

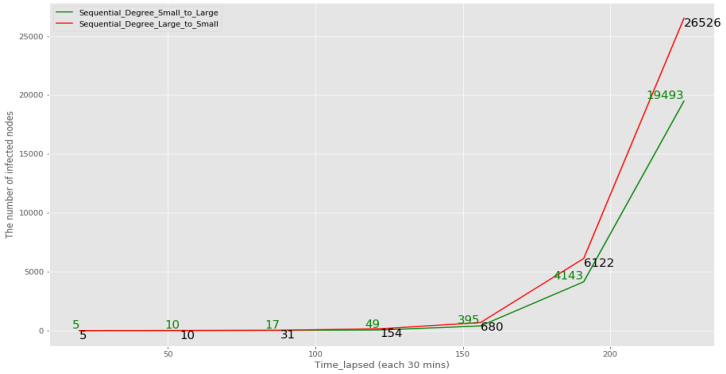

Figure 4.17: Total infected nodes based on different sequential strategies

Finally, the total infected nodes over time for these two types of sequential seeding strategy can be compared as well. It can be obviously seen that when the propagation time reaches 150 minutes, releasing seed in descending order, whether it is speed or range, obviously achieves a better performance than the ascending seeding strategy. Therefore, when the seed users have already been determined, within the future sequential seeding practice, the

seeds should be gradually released in order of their degree from large to small, contributing to achieving the influence maximization within one complex network.

4.2.2 Static and Dynamic Sequential Strategies

Seeds selection measures for the two sequential seeding strategies discussed above are all based on the static topological properties of the entire network, which are following called static sequential strategy. According to Jankowski J. [16], there is another approach regarding this called dynamic strategy. New seeds are determined according to the updated degree rankings during the spreading process by only taking the remaining network just containing the uninfected nodes into consideration. Although the previous research notes that this way can bring about wider spreading coverage in 90% of cases, its effect still should be examined in the particular Twitter topology. Meanwhile, based on the conclusion made in section 4.2.1, selecting seeds with the highest degree values can achieve the influence maximization. Therefore, all the seeds are determined based on their degree rankings in this scheme.

For this simulation plan, the evaluation interval from 0 to 210 minutes and a total of 20 seeds the same with simulating the other two kinds of sequential strategies are still set to make the related comparison. Initially, after sorting all the users depending on their out-degree values within the entire network, 5 seeds are set to begin the message spreading process. At the point of 30 minutes, the specific network containing only the uninfected users will be recalculated for its topological properties like degree and centrality values. Afterwards, the other 5 users with the highest degree are set as the seeds to continue the information dissemination. Furthermore, at the time of 60 and 90 minutes, the remaining 10 new seeds are determined using the same measure.

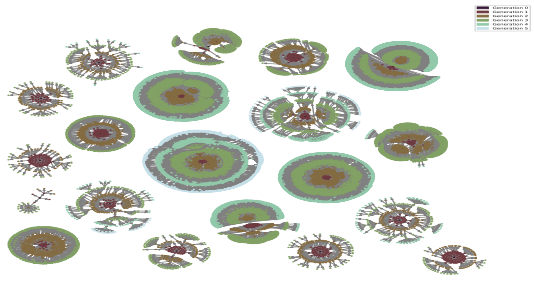

Figure 4.18: Dynamic sequential strategy

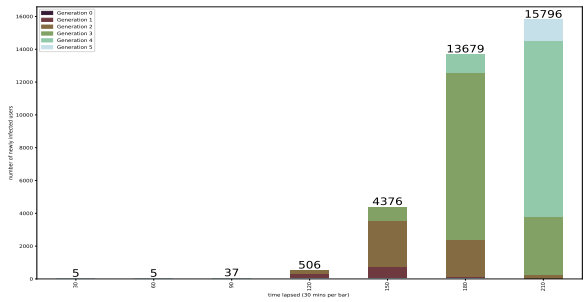

Figure 4.19: New infections over time

Compared with the spreading results in 4.11 and 4.12 of the static strategies, the dynamic one in 4.18 obviously activates more communities to spread the information in a wider range. Besides, the internal spreading pattern for most communities is like an annual ring. The potential reason for this result may be that when determining seeds during the later stages like 90 minutes, it can avoid picking the new seed users that should have been infected by the previous seeds. In this way, the possibility of new seeds appearing in different communities from the old seeds will be greatly increased. As a result, the total number of users reached by the dynamic strategy is 34,404, the most in three cases.

The distribution of the newly infected users over the time periods is also different in dynamic strategy. The information spread has been accelerated on a large scale from two and a half hours and the number of newly infected users at the 180th minute have reached more than 10,000, greatly different from the situation that began to spread from 180 minutes for the other two cases. Furthermore, compared its total amount of infections in each generation in figure 4.20 with the other two strategies in figure 4.15 and 4.16, it has already reached its peak in the third and fourth generations while the other two peaked in the fifth generation. All these results mentioned above could indicate that through utilizing the dynamic sequential seeding strategy both the propagation speed and the coverage can be enhanced a lot.

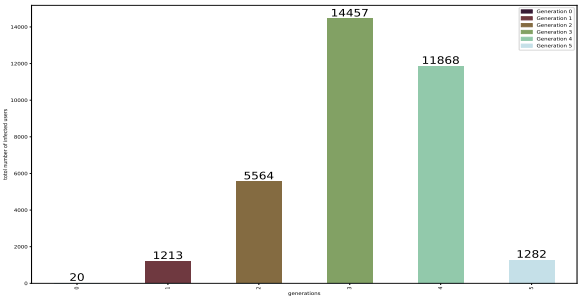

Figure 4.20: Infections with generation

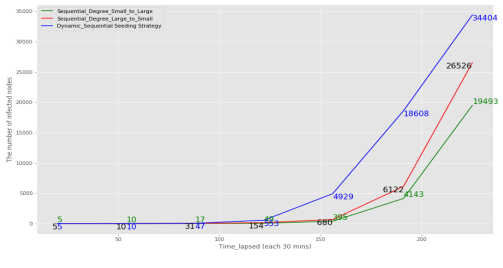

Figure 4.21: Static vs dynamic strategy

Finally, for all 3 broken lines in the right picture above regarding the total infected users with the time changes, the rising tendency of infections caused by these three strategies seems relatively consistent. However, when it reached 2 hours after starting diffusion, the dynamic strategy performs better with the faster speed and the significantly broader coverage. Therefore, in the future practice, when someone wants to spread relevant information faster and broader on Twitter, the dynamic sequential seeding strategy must be taken into the first consideration.

4.2.3 Single-stage and Sequential Strategies

After thoroughly analyzing the performance of various sequential seeding strategies in the process of information dissemination, it can be concluded that the dynamic sequential seeding depending on the constantly changing ranking of the degree values during the spreading process should be considered as the best strategy. Meanwhile, most previous researchers have concluded that the single-stage seeding performs not as effective as the sequential seeding strategy, which can better balance the speed and coverage. However, because of the new infection rate judgment model and the different simulator utilized in this research, simple analysis and comparison can be conducted to test the correctness of the above conclusion or to help analyze the more profound reasons.

In this section, the 20 seed users are determined the same as the ones in sequential seeding simulations. Then, what different from the sequential ones is that all these 20 seeds released the related message at the beginning to start the spreading. Within the same evaluation time interval from 0 to 210 minutes, their performance and influence in the information diffusion process can be easily compared.

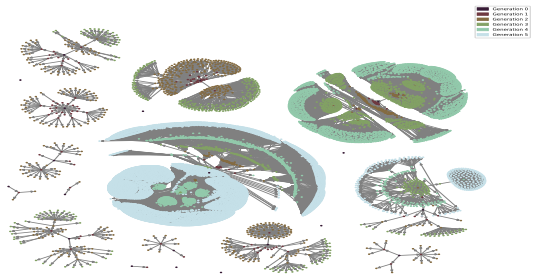

Figure 4.22: Single-stage seeding strategy

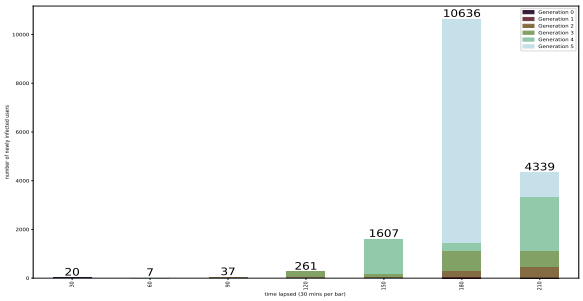

Figure 4.23: New infections over time

What should be emphasized firstly regarding the results is the final spreading coverage, only 16,902 users are reached in the single-stage case, far less than both two sequential cases. According to the figure 4.11, 4.12, and 4.22, the topological structure for all 3 resultant networks illustrates nearly the same amount of the generated communities resulted from the same 20 seeds, however, only around 3 large communities in size of the single-stage seeding one while more than 5 for the sequential results. This indicates the important role of the seeds appearance time playing in the information diffusion problem.

Transferring the attention into the newly infected users with the time extension, it also

displays the different patterns. For the single-stage seeding result, it reached the peak at 3 hours while began to drop dramatically since then. However, for the sequential results in figure 4.13 and 4.14, they are all in the rising form, especially after the point of 3 hours. Compared with the sequential seeding strategies, the way that releases all the seeds together at the beginning may perform well in the first few stages, but it will show the fading pattern later.

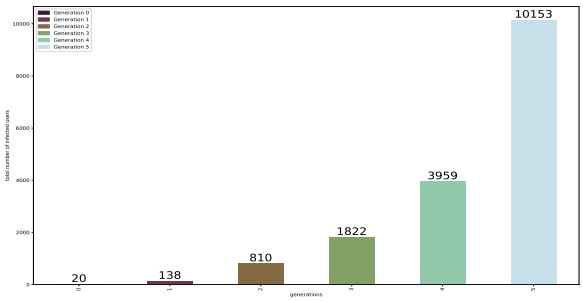

Figure 4.24: Infections with generation

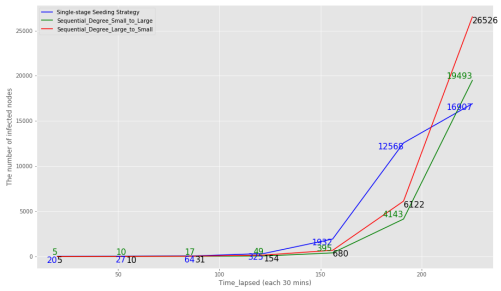

Figure 4.25: Single-stage vs Sequential

Meanwhile, as is shown in figure 4.24 for the single-stage case, the amount of infected users grows exponentially with the increase of the generation number, which is similar to the descending seeding strategy but with a greater exponent. In contrast, the ascending strategy performs not very stable in this respect. This can also reflect the dominant role of the nodes with significantly large out-degree when spreading messages in one network.

Furthermore, taking all the results of three types into one line chart 4.25, their performance including the spread speed and the coverage can be intuitively compared. It is obvious that although the single-stage seeding scheme shows a high spreading speed within the first 3 hours than the other two cases, the sequential strategies can better trade-off between the speed and the diffusion range. Finally, the descending seeding strategy reaches the widest coverage, therefore considered as the best strategy among all 3 of them.

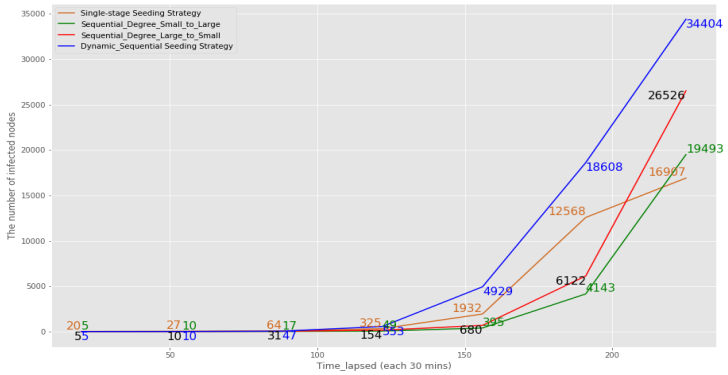

Figure 4.26: Four cases regarding the seeds appearance sequence

Till now, it can conclude the test result for hypothesis 2: *sequential seeding strategy can better trade-off between diffusion speed and coverage than single-stage one; moreover, the dynamic sequential strategy will generate the best performance.* According to figure 4.26, all three sequential seeding strategies can better balance between the spreading speed and their final impact range than the single-stage one. Moreover, it is obvious that the dynamic strategy is far ahead whether on speed or the coverage when dealing with diffusion influence maximization problem. Therefore, with these results, the hypothesis 2 can be well proven.

4.3 Influence from Seeds Activity Level

The top 20 influential features are listed in the figure 4.27 after the feature-importance analysis for the infection rate prediction model, including for each user, the number of days from the account registration, amount of tweets released with its normalized number and amount of tweets liked with its normalized number, which are represented by $UsM\_deltaDays$, $UsM\_statusesCount$, $UsM\_normalizedUserStatusesCount$, $UsM\_favouritesCount$ and $UsM\_normalizedUserFavouritesCount$ respectively. All those features mentioned can reflect the activity level of each user, which means it plays an essential role in information dissemination. Therefore, in the seed selection problem, in addition to the topological properties of the seed within the complex network, some characteristics of the seed user itself also need to be taken into consideration. Then, in this research, how the activity level since the users registered the accounts will affect the diffusion result is evaluated. In order to conduct this comparison, the method is to change the parameters about the number of tweets released.

In the beginning, 20 seeds are randomly selected from the network and then set the $UsM\_statusesCount$ and $UsM\_normalizedUserStatusesCount$ of the same seeds to 20179 and 7.211937 as high activity level, to 11945 and 3.661864 as medium activity level while to 201 and 0.075028 as low activity level. Meanwhile, all 3 simulations utilize the single-stage seeding strategy and the overall evaluation time interval is set from 0 to 300 minutes when started the information spread.

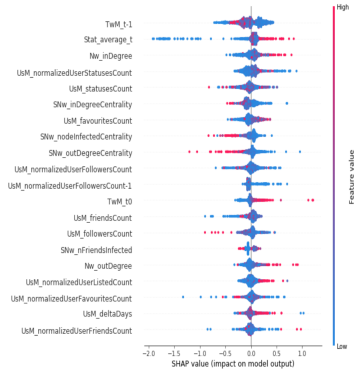

Figure 4.27: Importance

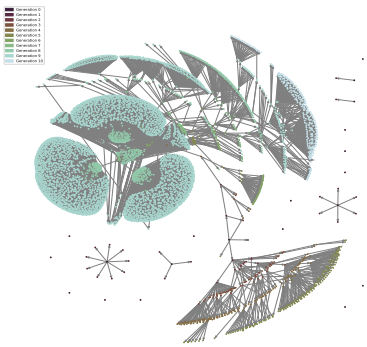

Figure 4.28: Low activity(a)

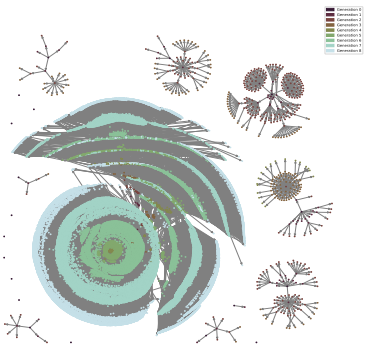

Figure 4.29: High activity(a)

Because of the similar topological pattern for two results with moderately and highly active seeds, only the high and the low are shown above to make a comparison. Firstly, from the total number of people infected by the news, the two cases with higher seed activity reached over 50,000, while the other one was even less than 6,000. Moreover, from the perspective of topology, the higher two are more concentrated and organized, which can already indicate the importance of the seeds activity level in message dissemination.

When transferring the attention to figure 4.30 and 4.33 for lower activity, the information began to spread from the 270th minute or the 8th generation gradually. However, just after half an hour or in the next generation, it suddenly infected more than 5,000 users. This phenomenon is also common in reality, for example on Sina Weibo, some popular stars always want to stop their negative news by usually controlling the accounts with a high activity level. However, due to insufficient control of the less active individual users, the related message usually bursts to an observable extent suddenly at a certain moment. Compared with this scheme, the other two began to spread the message continuously in the 150th minute and gradually reached the peak at the 210th minute or the seventh generation.

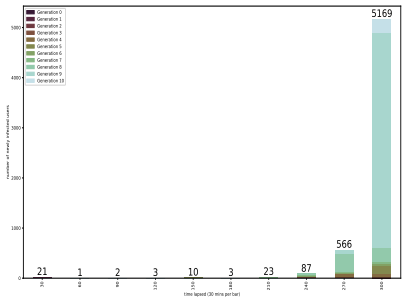

Figure 4.30: Low activity(b)

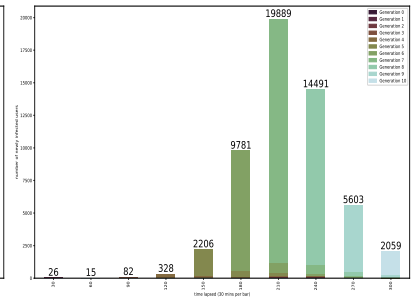

Figure 4.31: Mid activity(b)

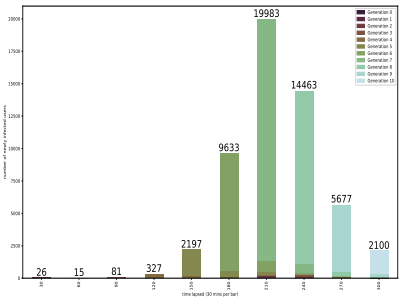

Figure 4.32: High activity(b)

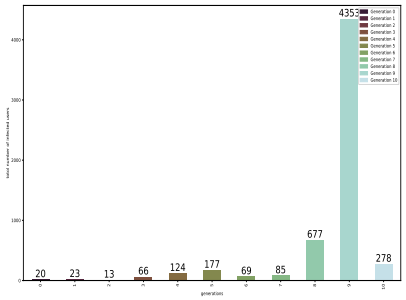

Figure 4.33: Low activity(c)

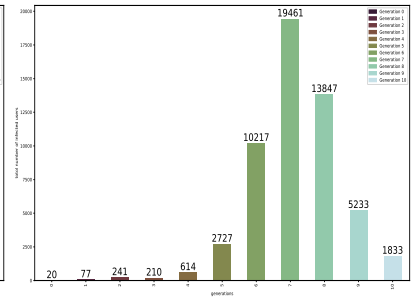

Figure 4.34: Mid activity(c)

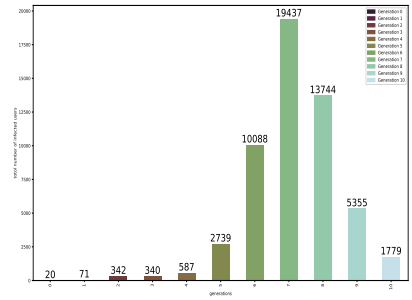

Figure 4.35: High activity(c)

Then, it can come back to the hypothesis 3: *when the seeds become inactive, their spread influence around the topics will drop considerably.* According to the line chart below, both the speed and the breadth are better for the case with highly active seeds. Therefore, the hypothesis can be proved intuitively, and it provides the meaningful instruction that when making seeds selection among the users with similar topological characteristics, more active netizens should be more appropriate for influence maximization.

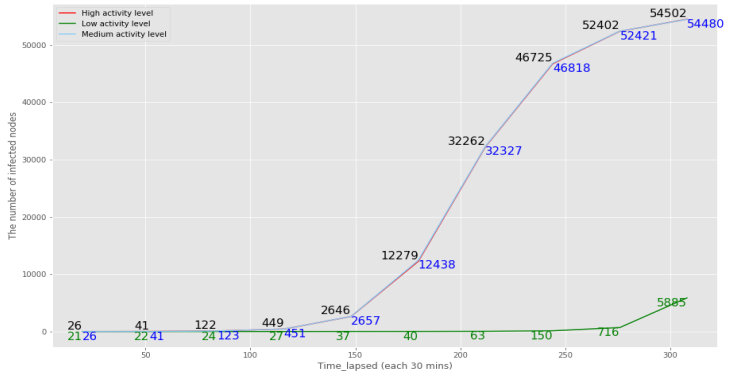

Figure 4.36: Total infected nodes with different activity level

Finally, what needs to be added here is that although the high activity level is twice the level of the moderate activity, it only differs by about 20 users in the end. Therefore, although the result shows that the higher the seeds activity, the wider coverage will be, in practice, as long as it is within a certain range, the ideal effect can be achieved. Also, people can make different choices regarding this aspect for different events.

4.4 Combined Strategy Generation

After studying how to use the characteristics of seeds and the order in which they emerged to maximize the effect of propagation, an idea that whether appropriately combining these aspects would achieve better performance or not comes into the mind. According to all results in the sections above, in the first few hours when the information started to spread, the total number of infected users generally meets a power function relationship with the degree values of the seeds, the level of the seeds activity, and the order of their appearance respectively. Also, their separate relationship can be presented below:

$$I_1(t) = x^{\lambda_1}, \qquad 0 < \lambda_1 < 1 \tag{4.1}$$

$$I_2(t) = y^{\lambda_2}, \qquad 0 < \lambda_2 < 1 \tag{4.2}$$

$$I_3(t) = z^{\lambda_3}, \qquad 0 < \lambda_3 < 1 \tag{4.3}$$

Then, the relationship between them can be simplified as following:

$$I(t) = ax^{\lambda_1} + by^{\lambda_2} + cz^{\lambda_3} \tag{4.4}$$

where I(t) means the total infections over time while a, b and c are the weighting coefficients between different strategies of x, y and z.

Afterwards, a new combined seeding strategy is formed in this research through taking advantage of the best situation in different aspects, including selecting the seeds based on the highest degree values with relatively high activity level as well and utilizing the dynamic sequential seeding strategy in the whole process of information diffusion.

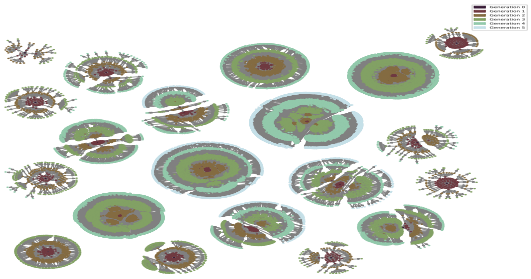

Figure 4.37: Combined seeding strategy

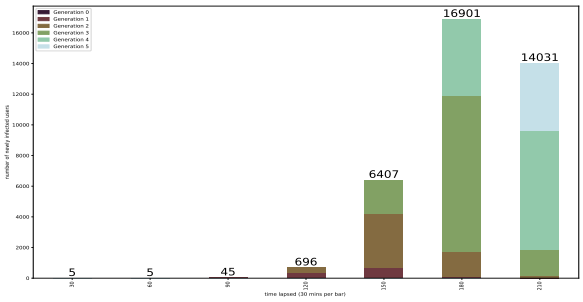

Figure 4.38: New infections over time

Completing the simulation for the combined strategy, what the first concern is the final coverage of the infected nodes. It reaches a total number of 38,090 users, which has been the most compared with the results for separate strategies of 24,867, 32,262 and 34,404 within the same evaluation interval. Concentrating on the topology in figure 4.37, it does activate the most communities, and the number of infections in each community is relatively evenly distributed, allowing maximum spread to small local groups within the network.

Afterwards, looking at the number of newly infected nodes over time, it has spread to 5,000 people on 150 minutes and even made the total number more than 20,000 within next half an hour, considerable speed and coverage. Similarly, it has reached the peak at the 4th generation and even covered over 30,000 users at the 5th generation. These are enough to prove that combining several great strategies could achieve better performance.

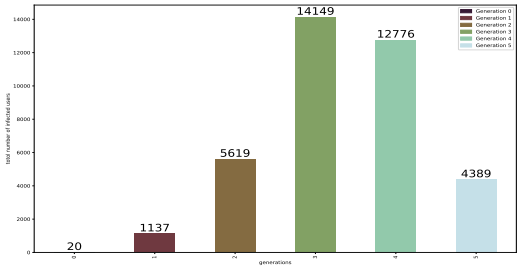

Figure 4.39: Infections with generation

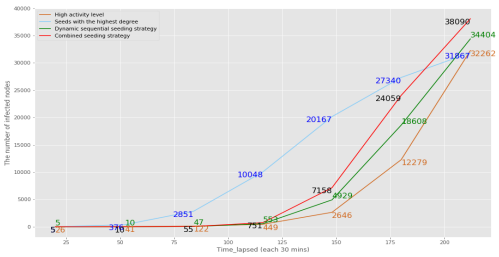

Figure 4.40: Combined vs Respective

Finally, illustrating the result of the combined strategy with the ones of separate strategies in figure 4.40, although the single-stage seeding strategy with the highest degree seeds give a faster speed and broader coverage with the first 3 hours, the other three cases can obviously better trade-off between the speed and coverage, finally reaching more users within the network. Meanwhile, compared the combined strategy with the single-stage highest

activity seeding strategy and the dynamic sequential seeding strategy, their performance was consistent and unsatisfactory before 120 minutes, however, after that, they made great progress both on speed and the covering range. Furthermore, the combined one has an outstanding performance from the 150th minutes, which seems almost straight up with a relatively large slope on a total number of users that infected by the related messages.

Due to the limitation of the calculation capacity, the parameters of these weighting coefficients were not adjusted and optimized. For example, setting the activity level within a certain range may bring about better results than just setting it to the highest, which should be similar when considering the degree values. Despite this, the current combined strategy has already performed better than the previous unilateral strategies. Therefore, such a combined seeding strategy provides an excellent theoretical basis for solving the problem of influence maximization in information dissemination.

Chapter 5

CONCLUSION AND DISCUSSION

5.1 Overview and Conclusion

The purpose of this research was to analyze how the seeds setting influence the information propagation on Twitter and then provide several theoretical bases and related data support for effective prediction and control of diffusion results in the future related practice. It mainly handled three research questions: what properties should be based on to select the seeds for influence maximization, how single-stage and sequential seeding strategy affect the information diffusion process and how does the activity level of seeds affect the spreading results. Moreover, it also put forward three hypotheses related to the above questions to make a wider information coverage: seeds selection should base on the highest betweenness centrality, sequential seeding strategies will perform better than the single-stage one while the dynamic one should be the best, and the final one was that inactive seeds would reduce the spread influence for the specific topic.

First of all, it established the network topology as the simulation space environment with more than 70,000 real existing Twitter users from various complete topic circles. Together with their real relationship on Twitter, the complex network with similar topological properties to the whole Twitter situation was generated. Through this, the analysis results could represent the information diffusion situation on the whole Twitter even other similar online social networking websites.

Secondly, this research constructed the essential simulator for different simulations. This special simulator primarily includes an infection rate prediction model, the iterative mech-

anism, seeds setting component and also the topology network mentioned above. The predictive model with relatively high accuracy was trained by XGBoost and the recent real event spreading situation on Twitter. For the iterative mechanism, the time window for each iteration is 30 minutes to meet the high accuracy requirements of the model, within which the features of all nodes will be updated to make a new prediction. Moreover, the seeds settings consist of the number of the seeds with their id and the appearance time for each. With this convenient simulation platform, all the simulations for various spreading cases with different seeds settings would be efficiently conducted to compare their unique influence on spreading speed and coverage.

	Seeding strategy	Time interval	150thmin	180thmin	210th min	Final
1	Single-stage with 5 highest degree seeds	0-210	20167	27340	**31867**	31867
2	Single-stage with 5 highest betweenness seeds	0-210	19722	26437	**30675**	30675
3	Single-stage with 5 highest eigencentrality seeds	0-210	2472	5896	**7957**	7957
4	Single-stage with 20 selected seeds	0-210	1932	12568	**16907**	16907
5	Sequential with 20 seeds from degree low to high	0-210	395	4143	**19493**	19493
6	Sequential with 20 seeds from degree high to low	0-210	680	6122	**26526**	26526
7	Dynamic with 20 dynamic ranking seeds	0-210	4929	18608	**34404**	34404
8	Single stage with 20 highly active seeds	0-300	2646	12279	**32262**	54502
9	Single stage with 20 moderately active seeds	0-300	2657	12438	**32327**	54480
10	Single stage with 20 lowly active seeds	0-300	37	40	**63**	5885
11	Combined with 20 dynamic ranking seeds	0-210	7158	24059	**38090**	38090

Table 5.1: Total number of infected users for different strategies

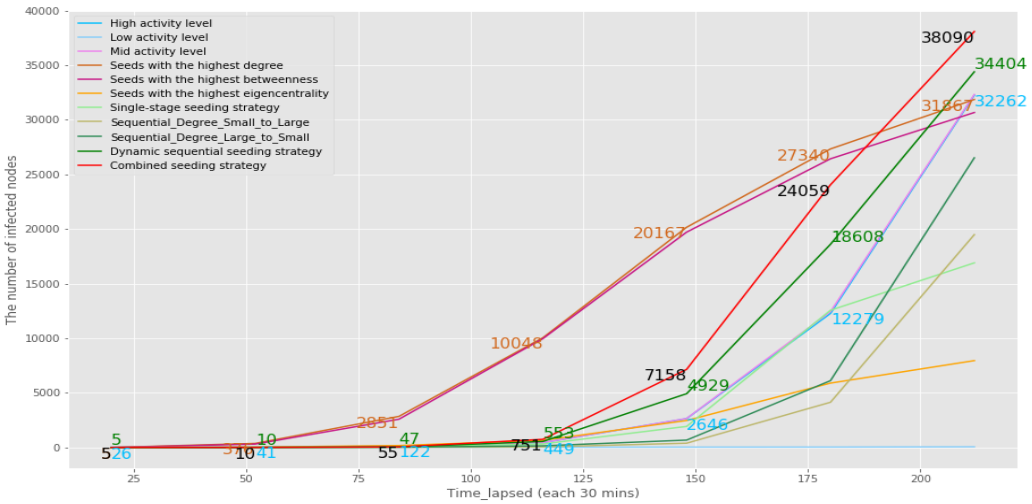

Figure 5.1: Final comparison for different strategies

Afterwards, all 10 different seeding strategies were well processed and compared to generate the better ones among them. According to the table and the line chart above, how the seeds properties and their appearance order influence the speed and the coverage of information diffusion can be concluded the following. When talked about the centrality values of the seeds, it is evident that seeds selection with the highest degree and betweenness have better performance than with the highest eigencentrality both on the speed and influence range reached while the highest degree seeds can contribute to the broadest coverage.

Moreover, when taken the seeds appearance sequence into consideration, in general, although the single-stage seeding strategy has faster speed within the first 3 hours, all 3 sequential ones can better balance between the coverage and the speed, finally reaching a wider coverage. Then, if the seeds are determined, the order according to their degree values from higher to lower performs better than the order from lower to higher. Finally, the best strategy with far better performance among these four should be the dynamic seeding strategy with seeds determination based on dynamic degree rankings within the whole spreading process.

Another topic regarding the seeds setting in this research is the influence of their activity level. It can be intuitively seen that when the seeds become inactive, their spread influence regarding the topics will drop considerably. Here what can be added is although

the higher the better, as long as their activity level reaches a certain range, they can always achieve satisfying results. Till now, the first hypothesis has been proved to be one-sided while the other two has been well proven.

Finally, a combined strategy taking advantage of all the best case within the different topics mentioned above was established, what should be emphasized is that this strategy can trade-off the spreading speed and coverage and finally can generate significantly great performance in diffusion influence maximization. Therefore, all the results above and this new strategy can provide several instructions on seeds settings in the future information spreading practice on OSNs, for example in the circumstance that the merchants want to spread the promotion information to more people in different regions or the stars want to curb the public opinion release about the scandal.

In conclusion, this research established a network and a simulator to analyze the influence of different seeding strategies on information diffusion primarily utilizing a simulation-based approach. All final results, analysis and comparison could provide several solid theoretical bases, and practice support for the prediction and control of the information propagation problem.

5.2 Limitations

Although this research generated many significant discoveries related to the seeds settings in information diffusion on Twitter, there also exist several limitations within the whole study process including the network topology, the predictive model, the evaluation time interval and the weight of each part in the combined strategy.

Firstly, although the network sampled in this research has similar topological properties with the structure of the whole Twitter, if it can utilize the entire Twitter network containing all the users with their relationships, the simulation result could be more accurate.

Secondly, the infection rate predictive model utilized in this research was trained by only one real event spreading situation on Twitter, which may contain the influence of the particularity of the event. If it can be trained from several different events and generate a more general model, it may provide more accurate predictions when applying to the spreading

control for future events.

Moreover, when taken the entire evaluation time interval into consideration, because the computing capacity is extremely limited for processing the enormous network and the most spread always occurs within five hours when the event happens, this research only evaluate the results like their speed and coverage only in an at most 3-hour interval. Therefore, all the situations after this time point would not be greatly evaluated or compared while the results in later periods would bring significant changes to the conclusion.

Furthermore, as is mentioned in the results section, the weights for different parts in the final generated seeding strategy are challenging to be determined accurately, resulting from the project time limit and the poor computing capacity. If they can be optimized through more simulations, the final performance of this unique seeding strategy for information diffusion would become far better than the current.

5.3 Future Improvements

For further study about this topic in the future, all the aspects mentioned in the limitations section should be taken into consideration to generate more accurate conclusions. Moreover, with respect to the factors that would influence the information spreading circumstances on OSNs, there are still many aspects worth exploring.

Firstly, one can generate the whole Twitter network as the spacial simulation environment to study the diffusion direction, speed, important nodes and also the coverage. Secondly, the infection rate predictive model could be generated through training more different events to make it more general and accurate for future practice. Thirdly, the entire evaluation interval could be extended longer to the point that the information about the specific topic will not be propagated any more on Twitter. Afterwards, the final combined strategy can be improved through optimizing the weights of different parts to achieve far better performance.

Moreover, the future seeding strategies can give more accurate classification and guidance based on the type of event, with which it will better meet the various detailed requirements from different individuals. Meanwhile, what this research considers more is the

influence maximization for information spreading problem, one can analyze more from another perspective on how to control information dissemination to make it propagate to fewer individuals.

Last but not least, another important factor that significantly affects the diffusion situation is the topological structure of the different network [25]. Therefore, one can also generate different complex networks form several online social networking websites and then compare their influence on spreading.

In conclusion, only through considering more about all the factors mentioned above that influence the information propagation results and proposing more effective and efficient measures to accelerate or control the spread, the communication on OSNs could be better utilized by various individuals, better enhancing the development of the economy and promoting the entirely social progress.

Bibliography

[1] Twitter network dataset, Apr. 2017.

[2] Akshay Java, Xiaodan Song, T. F., and Tseng, B. Why We Twitter: Understanding Microblogging Usage and Communities. In *Procedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007* (August 2007), Springer, pp. 56–65.

[3] Amaral, L. A. N., Scala, A., Barthélémy, M., and Stanley, H. E. Classes of small-world networks. *Proceedings of the National Academy of Sciences 97*, 21 (2000), 11149–11152.

[4] Barabási, A. L. Scale-free networks: A decade and beyond. *Science 325*, 5939 (2009), 412–413.

[5] Bonato, A., and Tian, A. *Complex Networks and Social Networks*. 2011.

[6] Bozzo, E., and Franceschet, M. Resistance distance, closeness, and betweenness. *Social Networks 35*, 3 (2013), 460 – 469.

[7] Buccafurri, F., Lax, G., Nicolazzo, S., Nocera, A., and Ursino, D. Measuring betweenness centrality in social internetworking scenarios. In *On the Move to Meaningful Internet Systems: OTM 2013 Workshops* (Berlin, Heidelberg, 2013), Y. T. Demey and H. Panetto, Eds., Springer Berlin Heidelberg, pp. 666–673.

[8] Chen, D., Lv, L., Shang, M., Zhang, Y., and Zhou, T. Identifying influential nodes in complex networks. *Physica A: Statistical Mechanics and its Applications 391*, 4 (2012), 1777 – 1787.

[9] Chen, T., and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), KDD '16, ACM, pp. 785–794.

[10] de C Gatti, M. A., Appel, A. P., dos Santos, C. N., Pinhanez, C. S., Cavalin, P. R., and Neto, S. B. A simulation-based approach to analyze the information diffusion in microblogging online social network. In *2013 Winter Simulations Conference (WSC)* (Dec 2013), pp. 1685–1696.

[11] Estrada, E. The structure of complex networks: Theory and applications, 2011.

[12] Ganesh, A. J., Massoulié, L., and Towsley, D. F. The effect of network topology on the spread of epidemics. *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies. 2* (2005), 1455–1466 vol. 2.

[13] Ghosh, S., and Ganguly, N. *Structure and Evolution of Online Social Networks*. Springer International Publishing, Cham, 2014, pp. 23–44.

[14] Hussain, O. A., Anwar, Z., Saleem, S., and Zaidi, F. Empirical analysis of seed selection criterion in influence mining for different classes of networks. In *2013 International Conference on Cloud and Green Computing* (Sept 2013), pp. 348–353.

[15] Isaac, M., and Ember, S. For election day influence, twitter ruled social media. *the New York Times 87* (Nov 2016), B3.

[16] Jankowski, J. Dynamic ran04kings for seed selection in complex networks: Balancing costs and coverage. *Entropy 19*, 4 (2017).

[17] Jankowski, J., Michalski, R., Bródka, P., and Karczmarczyk, A. Increasing coverage of information diffusion processes by reducing the number of initial seeds. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (New York, NY, USA, 2017), ASONAM '17, ACM, pp. 713–720.

[18] Jarosaw, J., Piotr, B., Przemysaw, K., Boleslaw, S., Radosaw, M., and Tomasz, K. Balancing speed and coverage by sequential seeding in complex networks. *Scientific Reports 7*, 891 (Apr 2017), 2045–2322.

[19] Kempe, D., Kleinberg, J., and Tardos, E. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2003), KDD '03, ACM, pp. 137–146.

[20] Kumar, S., Morstatter, F., and Liu, H. *Crawling Twitter Data*. Springer New York, New York, NY, 2014, pp. 5–22.

[21] Kwak, H., Lee, C., Park, H., and Moon, S. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web* (New York, NY, USA, 2010), WWW '10, ACM, pp. 591–600.

[22] Lee, S. H., Kim, P. J., and Jeong, H. Statistical properties of sampled networks. *Phys. Rev. E 73* (Jan 2006), 016102.

[23] Leskovec, J., and Krevl, A. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

[24] Liu, C., Zhan, X., Zhang, Z., Sun, G., and Hui, P. M. How events determine spreading patterns: information transmission via internal and external influences on social networks. *New Journal of Physics 17*, 11 (2015), 113045.

[25] Liu, Q., and Hong, T. Sequential seeding for spreading in complex networks: Influence of the network topology. *Physica A: Statistical Mechanics and its Applications 508* (2018), 10 – 17.

[26] Mattsson, L. G. *Industrial Marketing — The Network Perspective*. Gabler Verlag, Wiesbaden, 2004, pp. 175–201.

[27] McAuley, J., and Leskovec, J. Learning to discover social circles in ego networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* (USA, 2012), NIPS'12, Curran Associates Inc., pp. 539–547.

[28] McCorkle, D., and Payan, J. Using twitter in the marketing and advertising classroom to develop skills for social media marketing and personal branding. *Journal of Advertising Education 21*, 1 (2017), 33–43.

[29] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. *Measurement and Analysis of Online Social Networks*. IMC '07. ACM, New York, NY, USA, 2007, pp. 29–42.

[30] Nazemian, A., and Taghiyareh, F. Influence maximization in independent cascade model with positive and negative word of mouth. In *6th International Symposium on Telecommunications (IST)* (Nov 2012), IEEE, pp. 854–860.

[31] Okamoto, K., Chen, W., and Li, X. Ranking of closeness centrality for large-scale social networks. In *Frontiers in Algorithmics* (Berlin, Heidelberg, 2008), F. P. Preparata, X. Wu, and J. Yin, Eds., Springer Berlin Heidelberg, pp. 186–195.

[32] or Follow, F. Twitter: Most followers, Feb 2018.

[33] Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys. 87* (Aug 2015), 925–979.

[34] Ruhnau, B. Eigenvector-centrality a node-centrality? *Social Networks 22*, 4 (2000), 357 – 365.

[35] Salath, M., and Jones, J. H. Dynamics and control of diseases in networks with community structure. *PLOS Computational Biology 6*, 4 (04 2010), 1–11.

[36] Shakarian, P., Bhatnagar, A., Aleali, A., Shaabani, E., and Guo, R. *The Independent Cascade and Linear Threshold Models*. Springer, Cham, 2015, pp. 35–48.

[37] Siddiqui, S., and Singh, T. *Social Media its Impact with Positive and Negative Aspects*, vol. 5. International Journal of Computer Applications Technology and Research, 02 2016, pp. 71–75.

[38] Stolee, G., and Caton, S. Twitter, trump, and the base: A shift to a new form of presidential talk? *Signs and Society 6*, 1 (2018), 147–165.

[39] Varga, I. *Comparison of Network Topologies by Simulation of Advertising*. SciTePress, 2017, pp. 17–22.

[40] Wang, F., Wang, H., and Xu, K. Diffusive logistic model towards predicting information diffusion in online social networks. In *2012 32nd International Conference on Distributed Computing Systems Workshops* (June 2012), pp. 133–139.

[41] Wikipedia. *Countries and cities with local trending topics in Twitter*. Wikipedia, Jun 2015.

[42] Wikipedia. Facebook, Jan 2018.

[43] Wilson, E. Hubert de givenchy, master of romantic elegance, dies at 91. *the New York Times* (Mar 2018), A22.

[44] Zhang, Z., Liu, C., Zhan, X., Lu, X., Zhang, C., and Zhang, Y. Dynamics of information diffusion and its applications on complex networks. *Physics Reports 651* (2016), 1 – 34. Dynamics of information diffusion and its applications on complex networks.

[45] Zhou, L. Information spreading on twitter, 2018.

Appendix A

Code Instructions

All the related literature articles, datasets, codes, and results have already updated in Github within the following URL: *https://github.com/zym244855316/UCL_Final_project*. However, because of the size limit, several large files cannot be uploaded. If anyone needs them for further research, one can freely contact for sharing.

Firstly, the folders 'data' and 'Literature Review' contain part of the literature articles and the datasets for Twitter topology and simulation results.

Secondly, the data preprocessing regarding construction of the whole network and features extraction for all nodes are presented in folder 'preprocess'. Meanwhile, the code for the simulation process of all 11 seeding strategies is included in the folder 'simulation' with names clearly labelled.

Moreover, the code for visualization of the simulation results or other required evaluation graphs are all listed in the folder 'visualization' with all the results in the folder 'results'.

What needs to emphasize here is that part of the code was run on AWS while the remaining portion run in Jupyter Notebook, one can also contact me freely about the problems when running these code.