AN AGENT-BASED MODEL FOR INFORMATION DIFFUSION OVER ONLINE SOCIAL NETWORKS

A thesis submitted

To Kent State University in partial

Fulfillment of the requirements for the

Degree of Master of Arts

By

Zhuo Chen

December, 2016

Thesis written by

Zhuo Chen

B.S., Harbin Normal University, 2014

M.A. Kent State University, 2016

Approved by

JAY LEE, Professor, Ph.D., Department of Geography, Masters Advisor

SCOTT SHERIDAN, Professor and Interim Chairperson, Ph.D., Department of Geography

JAMES L. BLANK, Ph.D., Dean, College of Arts and Science

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all, I would like to wholeheartedly thank to my Master's advisor, Dr. Jay Lee, for endlessly supporting me during these past two years. I thank him not only for his patience, inspiration, and wisdom in the time of my research, but also for his warmth, kindness and care for his students in everyday life. Without his continuous support and insightful suggestions, I would not be able to finish my thesis and eventually get my Master's degree.

Second, I would like to give my sincere gratitude to my research committee, Dr. Xinyue Ye and Dr. Eric Shook. I enjoyed every moment we discussing the research and I have obtained plenty of helpful ideas and suggestion that sharpening my skills in research.

I would also like to thank dear friends and colleagues at the Department of Geography at Kent State University. I could not get on the path of research without their help. I miss the time we chatting, helping each other in the lab and the having lunch and dinner together. I am also grateful for their helpful suggestions on my research.

Finally, I would like to thank my families whom I miss greatly across the distance of a continent and an ocean. My beloved parents and wife offer their endless love and care from China. I feel sorry that I could only spent little time with them during these years, but their support and trust drive me working hard and finishing my thesis. In addition to my parents and wife, I am thankful for my aunt and uncle. They let me feel that I also have a home in a foreign land.

CHAPTER I. INTRODUCTION

People are born to be social. From the first second we were born, we were connected to our parents, then to relatives such as grandparents, uncles and cousins. As we grew up, we would hold a variety of social positions such as being a friend, a spouse, a classmate, a teacher, a colleague, a manager, or even the president of a company. These social positions connected us to each other. In turn, these connections form what we called social networks. Over thousands of years, people utilized their social networks for a wide variety of purposes, intentionally or not. The most important aspect of the functions of social networks is that they provide social resources that are embedded in the social networks (Lin, 1999). These social resources may be transferred via connections in the networks. Information as an intangible social resource also flows through the networks. With the advent of Internet, information could be generated with or without cost. In addition, social networks gradually evolved into cyberspace and emerged from many popular social network sites, such as Twitter, Flicker, Facebook, Tumbler, among others.

Thanks for the manifestation of social networks by these online services, observing how information diffuses has never been more feasible than ever. The study of diffusion of information has enormous implications for the society not only because of its effectiveness of marketing (Trusov et al., 2009) or political campaigns (Huckfeldt et al. 1995), but also its assistance in understanding overall human dynamics. Therefore, social networks could be an effective medium for us to analyze how information spreads via connections which composes half of the structure of social networks.

The pattern and extent of information diffusion via social networks depend greatly on the topology or the structure of the networks (Brown & Reingen, 1987). For example, information travels faster and wider in well-connected social networks than that in scarcely connected ones. Nevertheless, social networks are far more complicated than being described simply as "well or scarcely connected".

It is commonly agreed that social networks exhibit community or module/group structures (Newman & Park, 2003) because people naturally choose to connect with the likes in terms of geography, interest, ideology, or other aspects in life. Thus, social networks often consist of many communities in which within-community connections are much more than between-community connections (Newman, 2006). Information circulates rapidly within the communities via "strong ties" (i.e., interpersonal ties that usually have high emotional intensity and intimacy and high frequency of communication) but is barely passed out to other communities via "weak ties" (i.e., interpersonal ties that usually have low emotional intensity and intimacy and frequency of communication) (Granovetter, 1973).

Interestingly, the counter-intuitive fact is that weak ties are mostly the ones that contribute to the speed and extent of information diffusion (Granovetter, 1973; Bakshy, 2012). Moreover, transmission of information can be contagious and it can spread like a virus that can be transmitted through nodes in the networks. The contagion of information is more complicated than the contagion of an epidemic because many factors would affect whether a user is willing to adopt and transmit the information. Many seminal models, such as independent cascade (IC) model and linear threshold (LT) model had been proposed in the past (Guille, 2013). These models have been extended during the past decades, contributing to our understanding of the generalization of information contagion in social networks.

Because of the heterogeneous nature of individuals in social networks and the existence of a community structure, a simple model would not be capable of explaining or predicting how exactly a message would spread in social networks, especially those within and between communities. More importantly, recent works reported that information does not always spread via social network ties. This is especially true in the online social networks (Grabowicz et al. 2012; Pei et al. 2014). Users in Twitter, for example, can retweet (adopt the information and spread it) tweets even if they do not follow (or, are connected to) the users who initiated the tweets. This is because users could search for tweets on any given topics by using the search engine of Twitter. Alternatively, Twitter could provide users, based on their past tweeting histories, the tweets that are currently "trendy" or on topics related to their past concerns.

Given that information diffusion in online social networks is affected by the structure of social networks and the complex mechanism of a contagion, this thesis research focuses mainly on how efficiently information can flow through social networks based on the quantity and positions of seed nodes (or early adopters) in networks of different structures. In particular, this thesis research also examines such issues in networks with multiple communities. Furthermore, with the consideration of the difference between diffusion networks and social networks (i.e., diffusion process does not always follow the social network links), this thesis research also investigates how the early adopters function in these networks accordingly? To answer these questions, the thesis research uses the strength of agent-based modeling (ABM) approach, which is very powerful in modeling complex macro social phenomena emerged from simple, micro, and individual behaviors that could be aggregated for grouped behavioral patterns.

1.1 Research Purpose

Nowadays, social networks services such as Facebook, Twitter, Instagram, etc. seem to have become popular platforms for either celebrities, news media, organizations, or the general public to express their ideas and opinions. They have created a great opportunity for researchers to explore how information spread through online social networks when they allow anyone to post their words (thoughts) and even to examine the posts of all (any) users. Observations of people's communication in such an environment turn out to be easier than those in other traditional experimental settings. Just imagine how fascinating that thousands of people discuss a video game on online social services and within several minutes a plot of their tracks of communication could be drawn. Observing the way information diffuses and how it relates to the structures of social networks is important because, as they said, knowing someone seems to be more important than knowing something.

Investigating information diffusion in online social networks can never be easy. It requires the knowledge of human dynamics that explores human behavior overtime. The factors that contribute to the diffusion of information are complicated. Currently, it is believed that there are four critical components in the studies of information diffusion in online social networks: actors, content, the underlying network structure, and diffusion process. (Weng, 2014). With regards to the factors of a diffusion process, there are many challenges for understanding them, including finding influential spreaders (Guille et al. 2013) and maximizing the influence of information (Kempe et al. 2003). This thesis research studies the efficiency of information diffusion, which incorporates both finding influential spreaders and maximizing their influence of information transmission. Instead of finding a subset of nodes in the network that maximize the influence, this thesis attempts to find ways to choose the least number of nodes that speeds up

the information diffusion the most. In addition, this thesis also compares different types of networks in terms of the performance of propagating information, aiming to find the kind of network that triggers large cascades of information adoption.

In a nutshell, the thesis aims at contributing to studies of online social networks on information diffusion from the perspective of efficient diffusion with agent-based modeling and simulations. Outcomes from this study should provide hints to the geography likely behind information diffusion in social networks.

1.2 Research Questions

Overall, the questions asked in the present study are divided into three parts:

- On ***Network structure***: How a certain message spreads via social networks depends on the structure of the networks (Watts, 2002). Specifically, this research asks that, given different network structures:

  (1) To what extent the structure of a social network, for example, the different classic network structures, facilitate the process of information diffusion?

  In online social networking services, the diffusion networks and social networks are not the same because information does not exactly flow through links of social networks. Jumps over disconnected nodes (i.e., no links between them) are often allowed (and occurred in real social networks). This situation happens even more likely in online social networks with the advance of web 2.0 that online social networking sites are also tightly connected with traditional media. Will social networks still be an important channel of information diffusion? That being asked, an important subsequent questions would be to explore:

(2) ***To what extent would social networks account for the process of information diffusion since information does not always spread through social links, i.e., other avenues being the traditional channels such as TV/radio/newspaper broadcasting?***

In Twitter, during the event or incident, people spread out information by retweeting the tweets from others. Retweeting the tweets as the process of information diffusion, however, does not require a social link between the two users. Thus, in the collection of all the retweets, there may exist a good portion of them whose information sources are not through social links. What is the portion of the retweets that are due to social links between two users? If the percentage of these social-link based retweets are small, does it mean that social network lost its importance in modern social networking sites? These questions will be further explored and discussed in the thesis.

- On ***Diffusion efficiency***. In order to maximize the speed of information diffusion with a limited time budget*:*

3. ***How many early adopters (seed nodes) would be needed to disseminate the transmission of information in a certain social network so to ensure wide enough coverage and where are their best locations in the network if to achieve such coverage (the identification methods of early adopters)?***

Because of the limited resources and the law of diminishing marginal returns, it is impossible to make everyone to be early adopters in the early stages of information diffusion, that is, to adopt the information intentionally. Finding the optimal number of early adopters is critical in that it leverages resources in a more efficient way.

The rest of the thesis is structured as follows. Chapter 2 reviews concepts and works related to information diffusion in social networks. Chapter 3 presents the data and methods utilized by the thesis. Analyses and results are described in Chapter 4. Finally, Chapter 5 gives the conclusion and a discussion of the research.

CHAPTER II. CONCEPTS AND RELATED WORK

This chapter reviews four different but related topics: graph theory, the characteristics of social networks, information diffusion through social networks, and the maximization of influence on information diffusion in social networks. Though there are other challenges and topics with regard to information diffusion, these four topics make up the most flesh and bone of the present thesis.

2.1 Graph

2.1.1 Centralities and Heuristic Algorithms

Centrality refers to an indicator of how central a vertex is within a graph (in this case, a network). It is usually used in graph theory and network analysis. In addition, *heuristics* are approaches that can be adopted for solving well-defined problems, usually more quickly but less exhaustively than classic methods. Heuristic methods often employ a practical approach, or a short cut, in search of a solution that can get a solution but the solution is not guaranteed to be optimal or perfect. However, solutions found by applying heuristics are often sufficient for the immediate goals. In order to measure how suitable a node is to be selected as an early adopter or a seed node, an ample amount of methods has been proposed in the literature. This research investigates the performance of six centralities and heuristics with respect to the spread of influence.

2.1.2 Betweenness Centrality

Betweenness is one of the earliest centrality measures in the field of social network analysis. Linton Freeman is generally credited as the scholar who contributed to this method in 1976. Each node's betweenness in a network is derived by calculating the fraction of that node being on the shortest paths between paired nodes that pass through the node (or, a target node). In other words, given a target node, if it can be a bridge node that connects any pair of nodes in the network such that the more shortest paths between paired nodes going through the target node, the more central the target node would be. The betweenness centrality of a node v is given by the following expression:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths between node s and node t, and $\sigma_{st}(v)$ denotes the number of those shortest paths between s and t that pass through node v.

2.1.3 Closeness Centrality

Closeness centrality is another centrality measure in graph theory. Similar to betweenness centrality, closeness centrality utilizes shortest paths as well. However, closeness centrality emphasizes the length of the shortest path that can be the natural distance or geodesic distance metric between pairs of nodes (Sabidussi, 1966). For a node v, the *farness* of node v is defined as the sum of its distances along the shortest paths from all other nodes. The *closeness* then is defined as the reciprocal of the farness:

$$C_C(v) = \frac{1}{\sum_{s \in V \setminus v} d_G(v, s)}$$

where $d_G(v, s)$ is the geodesic distance between v and s. Thus, the more central a node is, the lower its total geodesic distance from all other nodes would be. Notice that some nodes may not be reachable from node v – two nodes could belong to separate "components" of a network with no connections between these separate components. It follows that closeness can be only computed within a well-connected network that each node can be reached by others through their connections.

2.1.4 Eigenvector Centrality

Eigenvector centrality, different from *degree centrality*, addresses the importance of a node's neighbors. It acknowledges that an influential node is not just a node with more neighbors but with more neighbors that are influential. The number of connections of a particular node still counts for something, but having fewer high-influence neighbors may outperform those who have more low-influence neighbors in the system (Bonacich,1987; Newman, 2006).

To obtain eigenvector centrality mathematically, an adjacency matrix is used. For node v in the network, the eigenvector centrality of node v can be defined as:

$$v_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} v_j$$

where λ is a constant. A_{ij} is the element in adjacency matrix in that $A_{ij} = 1$ when there is a link between node i and j, otherwise $A_{ij} = 0$.

2.1.5 K-shell

K-shell or *K-core decomposition* (Seidman, 1983; Batagelj & Zaversnik. 2011) is another centrality measure that decomposes a network from outside to inside (from less connected to most connected portions of the network). While the decomposition is in progress, each node is given an integer index or *coreness* value k_s (i.e., remaining degree after the decomposition) to indicate their centrality. More specifically, given a social network, we start by trimming or cutting off all nodes with degree $D1$.

After removing all the nodes with k_{D1}, some nodes may be left with only one link. Continue pruning the graph of the network iteratively in this manner until there is no node left with k_{D1}. The removed nodes, along with the corresponding links that were also removed, form a K-shell with index $k_s = D1$. With a similar pattern, we iteratively remove the next K-shell, $k_s = D2$, and continue removing higher K-shells until all nodes are removed. Ultimately, each node is associated with one k_s index, and the network can be viewed as the union of all K-shells. The resulting classification of a node can be very different from that when the degree k is used.

In general, K-shell decomposition is useful for understanding both (i) the presence of a hierarchical structure and (ii) clusters of tightly connected nodes in a graph (Miorandi & Pellegrini, 2010). Although K-shell index is derived from the degree of nodes in a graph, it tells more information than just the raw degree in terms of the role played by a node in the network. In addition, K-shell decomposition serves as a good tool for dissecting and visualizing many network structures in the real world. (Carmi et al., 2006; Alvarez-Hamelin et al., 2008)

2.1.6 Greedy Algorithm

Greedy algorithm (GA) is widely used in machine learning, business intelligence (BI), artificial intelligence (AI), and computer programming as a problem-solution method. An algorithm is a set of robust processes, sometimes defined mathematically, that can reach a solution to a problem if the processes are followed with pre-defined parametric values. For use in efficient information diffusion, GA looks for the best candidate nodes that maximizes their influence (i.e., the number of further adopted nodes influenced by them) during the process of information propagation. It has been proved that finding the most influential nodes in a network is NP-hard and GA can be used to find the optimal solution for influence maximization with provable approximation guarantees. (Kempel et al., 2003). Kempel's empirical experiment also proved that GA outperforms other degree- or centrality-based heuristics in maximizing the influence of selected nodes on information propagation. However, GA used limited centrality-based methods and it has been many years during which many other efficient centrality measures have been introduced. Therefore, a more comprehensive study that compares GA with other centralities and heuristics becomes necessary.

The general idea of implementing GA in the analysis of social networks is as follows. For a social network $G = (V, E)$ with V as the set of vertices and E being the set of edges, let S be the subset of vertices selected as seed nodes that trigger the influence cascade. Let $Diff(S)$ denote the process of propagation simulation that depends on the type of models chosen for information diffusion. An information model decides whether the information can be transmitted from one to another. Different models have different mechanisms in diffusion and thus will have different diffusion results. The mechanism of different models for information diffusion will be explained more specifically in the *Data and Methods* chapter of this thesis. The result of

$Diff(S)$ is a subset of vertices influenced by S, or the set of nodes that have been influenced by S. Based on GA, we keep adding nodes into S, but only one at a time, such that, with multiple simulations of $Diff(S \cup v)$, the number of the set of vertices from the result of $Diff(S \cup v)$ can be maximized.

2.1.7 Degree Discount

Degree discount heuristics is a much faster algorithm than GA for locating those influential information spreaders (Chen at al., 2009). Chen at al. shows that degree discount heuristics run only in milliseconds while GA's often run in hours in a network with tens of thousands of nodes. In the meantime, the selected influential spreaders calculated by using degree discount heuristics achieved even better influence spread than by using most centralities and degree-based heuristics.

The general idea of degree discount is as follows. If a vertex u has been selected as an early adopter and when considering selecting v, a neighbor of u, as a new early adopter based on its degree, the edge $\overline{vu}$ should not be counted towards its degree. The reason for discounting v's degree by one is that u is already selected as a seed node and adding the neighbors of u into the seed set is believed to be redundant. For every neighbor of v that is already in the seed set (e.g., node u), we do the same discount on v's degree.

Algorithm: Degree Discount (Chen et al. 2009)	
1:	initialize $S = \emptyset$
2:	**for** each vertex v **do**
3:	Compute its degree d_v
4:	$dd_v = d_v$
5:	initialize t_v to 0
6:	**end for**
7:	**for** $i = 1$ to k **do**
8:	select $u = \arg max_v \{dd_v \mid v \in V \setminus S\}$

9: $S = S \cup \{u\}$
10: **for** each neighbor v of u and $v \in V \setminus S$ **do**
11: $t_v = t_v + 1$
12: $dd_v = d_v - 2t_v - (d_v - t_v)t_v p$
13: **end for**
14: **end for**
15: Output S

In the algorithm, S denotes the seed set as the result and t_v denotes the number of neighbors of a vertex v that are already selected as early adopters. Propagation probability is represented by p with dd_v being the discount degree of node v.

2.2 Social Network

Social networks are complex networks that have attracted a great deal of attention from researchers in many fields such as computer science, sociology, epidemiology, physics, etc. As complex as they are, social networks are not formed randomly. They bear substantive features and often display their own particular patterns. Two prominent characteristics of social networks are the small world phenomenon (Watts & Strogatz, 1998) and the social networks' property of being scale free (i.e., power law) (Barabási et al., 2000).

Small-world phenomenon is the manifestation of a popular theory, known as *six degrees of separation*, which was originally described by Frigyes Karinthy in 1929. It had been experimentally tested by Stanley Milgram (1967), showing that two arbitrary people can reach each other by at most six intermediaries of their connections. Scale-free networks are networks whose degree distribution follows the power law, implying that only a few nodes dominate the total degree (i.e., number of connections) of networks while the majority of the nodes contribute little.

Many algorithms have been introduced to generate these structures in simulated networks. The first and most famous is Barabási -Albert model. It generates a scale-free network with two simple mechanisms: continuously adding new nodes into the system ("growth") and connecting with other nodes with a preference of connecting the new nodes to the high-degree (well-connected) ones ("preferential attachment") (Barabási and Albert, 1999). Extensions of the Barabási -Albert model emerged in subsequent years. One of the models, for example, is Two-Level Network model (Dangalchev, 2004) that adds a second-order preferential attachment to the network. The attractiveness of a node is not only determined by its degree but also by the total degrees of its neighbors.

Recently, a consensus has been reached among researchers in this area that the structure of a network can affect the processes of information diffusion in many ways (Moreno et al 2004; Watts, 2002; Dodds & Watts, 2004). For example, Centola (2010) investigated the effects of network structure on behavioral diffusion by studying the spread of health behaviors. This study demonstrated that clustered-lattice networks were more efficient than corresponding random networks for behavioral diffusion.

It should be noted that social networks are more complicated than clustered-lattice networks or random networks. By studying the structures of online social networks, Mislove et al. (2007) found that these networks contained a large strongly connected core of high-degree nodes, surrounded by many small clusters of low-degree nodes. In addition to this core structure, another important structure in social networks is that social networks are often community-based (Newman, 2006; Girvan & Newman, 2002; Ahn et al. 2010; Fortunato, 2010; Weng et al., 2013). Social networks naturally have a community structure. It can be detected within complex networks (Rosvall & Bergstrom, 2007). The structure of communities has been demonstrated to

have effects on information diffusion in that it could speed up or impede the information flows, or the spread of diseases (Granovetter, 1973; Onnela et al., 2007; Grabowicz et al., 2012; Weng et al., 2013, 2014). By considering the community structure in information diffusion models, it is likely that we would be able to dissect and study the process of information diffusion more comprehensively.

2.3 Information Diffusion

Information diffusion is the process that information propagates over time through certain intermediaries among individuals of a social network. Information diffusion in social networks is now a hot issue that attracts a tremendous amount of attention by researchers working on social behaviors. In order to understand the functionality of information diffusion, there are some notable challenges, such as detecting interesting/trending topics, modeling diffusion process and identifying influential spreaders (Guille et al., 2013). The problems that this thesis is addressing could belong to identifying influential spreaders of information. However, finding influential spreaders really depends on the scale of the social networks such that large and complex networks would have a large group of influential spreaders while small or simple networks have few. Thus, if we want to locate all of these influential nodes in a social network and employ them to start an information cascade, for the purpose of maximizing influence of information diffusion, tons of resources would need to be invested. Therefore, an efficient way of generating large information cascade should be introduced so that with limited resources, the information propagation in social networks can be maximized.

There are many implications for studying efficient information diffusion. Viral marketing can be the first application of socially based information propagation. It suggests that ideas are

spread by the *word of mouth* (WoM) effect in the marketplace. WoM effect effectively encourages people to adopt the information or a product (Herr, Karades, and Kim, 1991). As we enter the modern digital world, WoM effect goes online, especially with the emergence of social media.

People are heterogeneous. Some people may like the advertisement of a product while some may not. In addition, some people may have great influence on their neighbors while others barely do. These people who have great influence on others are called *opinion leaders*. In the process of information diffusion, the existence of opinion leaders may significantly affect the final outcome of adoption and speed of diffusion (Peter S. van Eck, 2011). Most of the studies of maximizing information diffusion did not consider the role of opinion leaders and they assumed that all nodes in a network were the same (i.e., each node had the same probability of adopting and propagating information in the diffusion process). These researchers focus only on connectivity but not considering the potentially significant influence of personality traits or knowledge among influential adopters.

In order to meaningfully mimic information diffusion in online social networks, the role of opinion leaders should not be ignored. Though opinion leaders are usually located in the central parts of the networks, they differ from early adopters when considering the efficient way of diffusion. Early adopters are those who first adopt information in the networks and start to spread it to their neighbors. Thus, their quantity and locations are the most important aspect when we are interested in efficient diffusion.

A different perspective of studying an efficient information diffusion is how it has been done in epidemiology. For information diffusion, the goal is to maximize the spread of information in the networks. However, epidemiologists are trying to minimize the spread of

diseases. That being said, the common goal is to locate these most influential people in the networks so that the speed of diffusion can be controlled.

2.4 Efficient Diffusion on Social Networks

Social networks may serve as an efficient tool for information propagation. People would express their behaviors, ideas, innovations, or memes in the social network sites, in a hope to convey them to the public. Social networks exhibit complex structures. That complexity makes it difficult to find who could be early adopters who could trigger and maximize the spread of information flows in social networks? An early adopter may or may not be an opinion leader though mostly the people who are opinion leaders usually function as early adopters because of their special positions in the network. In addition, opinion leaders are those nodes that are most capable of spreading the ideas, behaviors, and information within and between communities (Rogers, 2003).

In some early works, opinion leaders were assumed to be initial adopters that triggered the information cascade. Valente & Davis (1999) showed that using opinion leaders to disseminate information within networks could be a process that significantly outperformed those just using random nodes in the social networks. Their experiment sample, however, was derived from a small physician community in Illinois, which might be different from large online social networks. To that end, a large number of studies had been conducted to look for efficient methods for choosing the most influential opinion leaders in the larger social networks (Valente & Pumpuang, 2007; Kimura et al., 2007; Leskovec et al., 2006). Moreover, the most common methods for identifying opinion leaders in social networks used to be conducting interviews of network entities. The interviews, needless to say, were time-consuming and expensive.

Nowadays, in online social networks, it is simple to identify opinion leaders using social network analytics.

Nevertheless, in the challenge of finding influential spreaders in a social network, opinion leaders can be different from early adopters. In this thesis, opinion leaders are those influential people who have most connections in their communities. Early adopters, however, refer to people who adopted the information the earliest. Opinion leaders and early adopters sometimes may overlap, but chances are still good that a node with few connections may be very important in the network and thus being an early adopter.

Therefore, algorithms for efficiently finding influential early adopters are needed. This problem is known as *Influence Maximization* problems which was first formulated in Kempe et al.'s seminal paper (2003). Kempe and his colleagues described this optimization problem as NP-hard (Non-deterministic Polynomial-time hard), and proposed a heuristic algorithm (greedy algorithm, or GA) for solving the problem. Their experiments were conducted based on three different information diffusion models: the independent cascade model, the weighted cascade model, and the linear threshold model.

Subsequently, researchers (Leskovec et al., 2007; Chen et al., 2009) started to delve into improving the GA's for this maximization problem. These improved algorithms reduced the search time significantly. However, in spite of the efficiency of identifying early adopters with improved algorithms, the diffusion models could not be confirmed as being accurately finding an optimal set of early adopters in online social networks. As such, this issue should be investigated comprehensively. Along this line of thought and for the fact that information also spreads out of social networks, whether identified early adopters had maximized influence on information spread remains to be investigated further.

The problem of influence maximization is meaningful and worthy of investigation because the budget for information dissemination is often limited. In addition, because of the law of *diminishing marginal return*, more early adopters do not always warrant wider spread of information. In fact, this problem also pertains to the effectiveness of information diffusion in which the ultimate goal is to find the optimal way of disseminating information in terms of speed and/or in terms of extent. This problem is important also because nowadays information is shifting onto online social media and being exchanged more intensely via online social networks. With such a revolutionary trend, will the mechanism of information diffusion and the role of opinion leaders change from their influential roles in physical space to being influential in cyberspace accordingly?

How information diffuses is still mysterious to most of us. In spite of the existence of a large variety of information diffusion models, none can fully explain the process of information diffusion. Furthermore, the existing models assume that information spreads solely on the social networks, which in many cases deviates from the reality. Real diffusion process needs to be further explored if we want to confidently utilize social networks to facilitate efficient diffusion of information such as policies or advertisements. To that end, this research attempts to address this problem directly and uses real diffusion networks (in Twitter) to investigate the difference between diffusion networks and underlying social networks. More importantly, this research also explores the evidence of efficient diffusion on artificial and real social networks.

CHAPTER III. DATA AND METHODS

3.1 Twitter

The dataset of real networks was obtained from Twitter. Twitter, now as one of the largest social networking sites, is a huge platform on which users express their feelings, share interesting materials (e.g., texts, graphics, or videos), and discuss news. As of December 31, 2015, the monthly active users have reached 320 million. Different from other social networking sites such as Facebook, this microblog service only allows users to post less than 140 words in their single posts or tweets in Twitter. Another difference is the way that users form connections in their social networks. In Twitter, users can follow and also be followed, which means that a bi-directional relation is not necessary. A user can receive the posts from the users they followed, yet users that are being followed may not receive their followers' posts.

Social networks in Twitter provide abundant resources from the perspective of scholars, advertisers, and political activists. Many of them see online social networks as a great opportunity to study the different aspects of information diffusion, viral marketing, and the formation of great communities. However, it has been pointed out that these connections in online social networks do not imply the existence of an interaction between two users that are connected (HuberMan et al., 2008). In other words, most of the connections in the social networks of Twitter were meaningless. One can easily buy 2,500 Twitter followers for the low price of about $25. These followers will do nothing interactive but following. Thus a different way of detecting online social networks becomes necessary in order to investigate the interaction among users.

Twitter, like most social networking services, provides a set of functional symbols that improves their customers' experience of usage. The @ ("at") symbol means that a post could be intentionally written for someone. The user after the @ symbol would receive the notification to view the post. For instance, *Figure 1* is a tweet from President Barack Obama. In his tweet, @TheEllenShow is the *handle* of the twitter account of TheEllenShow (i.e., screenname of Ellen DeGeneres), which means President Obama mentioned TheEllenShow in his tweet and TheEllenShow should be notified to see this tweet by Twitter.

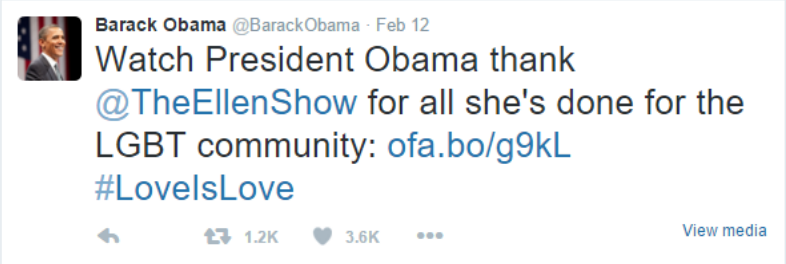

Figure 1: An example of tweet from Barack Obama

Mention symbol has been seen as the representation of active interaction among users. Another important reason that the @ symbol plays an essential role in Twitter is that every reply to a tweet starts with an @ symbol. With the existence of the @ symbol, we are then able to track users' friends not only by their following or follower's relationship, but also by the activity that users mention others using the @ symbol in their tweets. Twitter provides API that allows us to collect tweets by almost any user. Although there is a limit that we can only retrieve the most recent 3,200 historical tweets of each user. Such limit of 3,200 tweets is almost a one-year amount of tweets even for active users who tweet 10 times a day averagely.

With the ability to collect the most recent 3,200 tweets of each twitter user, we could form a social network based on interaction of users by extracting @ symbol in their tweets. The social network formed by this way may be more reliable when studying the information

propagation through these links. In fact, it is not feasible to collect all the follower/following relationships in a reasonable time because Twitter API restricts the number of queries to only one run for every 15 minutes.

3.1.1 Data Collection

In the thesis, the dataset of tweets was collected on topic level, i.e., the dataset of tweets is only a collection of tweets concerning particular topics and would not cover the whole twitter world. The reason for this is twofold. The first is because with the limitation of Twitter API, as described previously, it is impossible to obtain the complete information of all the tweets and their posters within the short time that Twitter permits. The second consideration is that topics are often related to events happened in somewhere in the real world, therefore it is meaningful to study the dynamics of information concerning a specific event over a certain period of time.

The dataset used in the thesis was collected for a topic of Bernardo wildfire that happened in May, 2014. The wildfire lasted five days and was not 100% contained until May 18th. By using the Twitter Search API, all of the tweets that were related to Bernardo wildfire were collected. A total of 1,997 tweets were retrieved and most of the tweets were generated during the first three days after the initial ignition. The 1,997 tweets were posted by 1,290 unique Twitter users. Because Bernardo wildfire was a local fire near San Diego, most of the users were located in San Diego, according to their profiles. It indicated that it was a local topic in the Twitter and it might also imply the existence of local communities in the virtual world.

In order to obtain the relationship among the users of this topic, the most recent 3,200 historical tweets of each user were collected and their friends were also extracted by using the @ symbol in the tweets. Because some users had set a privacy protection in their account, it was not

possible to collect their tweets using Twitter API. Thus, in total, 795 users were accessed and their historical tweets were collected. An online social network of 6,950 edges (links) and 726 nodes (not 795 because some were not connected to any of the other users in the topic) are formed among these users, which showed a power law degree distribution.

In addition to online social networks, online diffusion or information networks are also an essential part of Twitter networks. In Bernardo tweets dataset, around 61% of the tweets were retweets, which meant forwarding the messages of specified users in their own posts. It could be extracted by collecting retweets (*RT @username*). In other words, *retweets* stand for tweets that are not originally posted by the same user. The original user's name will appear in the retweets followed by RT @. For example, if user A retweets user B's tweet: "I love this cat!" the retweet posted by user A would be "*RT @B* I love this cat". Following this pattern, a retweet network, which represents a diffusion network, could be structured.

In summary, using the tweets about Bernardo wildfire, a social network and diffusion network are extracted in the study. We could see these two networks as different layers of the network from the wildfire community (i.e., the same users in the community but have both social networks and diffusion networks). Therefore, an integrated network that contains both social network and diffusion network was established in the study.

3.1.2 Limitation

Different from using the common Followers/Following relationship as a representation of the social networks, the method of data collection utilized "@" symbol as the representation of social link. While it has the advantage of filtering out unnecessary and ineffective social links (e.g., purchasable followers), it might have the limitation of not including effective social links.

Therefore, this data was collected under the assumption that users interact with their friends in Twitter in their recent 3,200 tweets. While classic Followers/Following relationship captures redundant social links, the methods in this thesis takes the snapshot of the essence in users' social relationship. It has the risk of omitting real social links of certain individuals, but it compensates by saving tons of time from collecting the entire social networks in Twitter.

3.2 Agent-based Modeling and Simulations

Since the goal of the research includes maximizing the influence of information in online social networks with the consideration of opinion leadership, the identification of seed nodes (i.e., early adopters) of information and opinion leaders become the most important aspects in this study. Seed nodes and opinion leaders are not the same because diffusion of information may start with seed nodes but the spread of information generally does not necessarily begin with opinion leaders. For this reason, choosing opinion leaders as seed nodes would not ensure the maximization of their influence in the networks.

It may be plausible to assign the most connected opinion leaders in the social networks as early information adopters so that the influence of the information could flow through the networks the most effectively. To that end, it should be noted that social networks are complex and intricate and that connectivity alone sometimes could not account for all of the complexity in social networks. Highly centered opinion leaders may cluster together (i.e., connected with strong ties but not necessarily geographically clustered), leaving most of the nodes in the peripherals barely reachable except through only weak ties. Consequently, it is difficult to use the conventional mathematical equations to identify seed nodes in social networks. Agent-based model, with its ability to model complex interactions with feedback loops, can therefore be

suitable for analyses and experiments of how text information flows and diffuses through social networks.

Agent-based modeling and simulation (ABMS) is an effective approach to modeling systems that are comprised of autonomous but interacting agents (Macal & North, 2005). An agent could be either an individual or a set of collective entities such as an organization or a group. Each agent follows its own goal, behavior pattern, and internal states according to a set of rules and contextual information from the history, relationships to other agents, and possibly other environmental settings (Weng, 2014).

Unlike conventional theories and classic models expressed in natural language, agent-based model is a computer program, meaning that it is forced to be precise. Agents in an ABM completely follows the rules that were pre-defined just like things that are logically coded in computer programs. The core of ABM is the rules of how agents interacting and having impacts on each other mutually. This allows the system changes dynamically, not because the rules set for the system, but because the rules set for the agents that compose the system. The greatest part of ABM is that it allows interrelation between model parameters. Unlike regression models that independent variables cannot have mutual feedbacks, ABM allows such interactions in full scale.

In addition, the use of agent-based modeling approach is based on the consideration that the interactions between nodes that are connected directly or indirectly as links can be interactive, dynamic, and changing with the progress of time. This type of complex relationships cannot be modeled by the traditional approaches such as regression models. Agent-based modeling has the benefits of 1) capturing emergent phenomena; 2) providing a natural description of a system and 3) being flexible and dynamic in allowing feedbacks between influencing factors (Bonabeau, 2002). The advantage of using agent-based modeling stands out

especially when a system of complex elements becomes hard to capture and generalize whereas each element and the interaction between each other of the system is clear and expressible in an algorithm.

3.2.1 NETLOGO

This thesis research uses NetLogo as the platform to establish agent-based models. NetLogo (https://ccl.northwestern.edu/NetLogo/) is a "low-threshold, high-ceiling" programming platform for agent-based simulation/modeling. NetLogo provides an excellent GUI (graphic user interface) with switches, sliders, choosers, inputs, and other interface elements. These elements dramatically simplify the construction of models/simulators. In addition, NetLogo has its own language which is an extension of Logo programming language (Papert, 1980). The difference between the original Logo language and the current NetLogo language is that, instead of controlling one single *turtle* at a time in the traditional Logo, programmers can now control thousands of them in NetLogo (Tisue & Wilensky, 2004) at the same time. *Turtles,* along with *patches*, *links*, and *observers* are the representation of agents in NetLogo (Railsback & Grimm, 2011). The NetLogo world is two dimensional and is divided up into a grid of *Patches*. *Turtles* are those agents that can move around in the world of NetLogo over *Patches*. Each *Patch* is a square piece with coordinates. For example, *Patches* can represent the land on which people, represented by *Turtles*, live. *Turtles* can move around over *Patches* just like people can move around over land. *Links* are agents that connect two *Turtles*. Each *Link* has two ends; each end is a *Turtle*. A *Link* dies if either its end-*Turtles* dies. Finally, the *Observer* is the agent that gives instructions to the other agents. It is invisible in the world.

It is important to note that, since a social network contains connections or edges, these edges can be represented by *Links* breed in NetLogo. One can easily create a link between two turtles with a single line of NetLogo code. To support the breed such as *Links*, NetLogo also features a variety of predefined functions/commands (called *Primitives* in NetLogo). For example, there is a predefined function called *Link-neighbors*. It can access any neighbor that is being connected to a specific turtle or individual. Mostly importantly, NetLogo also supports different extensions for data manipulation, including networks extension and GIS extension. These extensions enable NetLogo to deal with real world data in this research. In addition, NetLogo is also suitable for simulating the generation of random social networks as well as the spread of social media messages through individuals.

NetLogo does have its drawbacks. The major problem comes from the execution speed for models in NetLogo and it leaves much to be desired. For a network that exceeds thousands of nodes and edges, a NetLogo model of such network cannot be easily executed due to limitations in computing resources that ordinary PCs have. However, when not simulating a huge and complex system, NetLogo is a good alternative for modeling small social networks.

3.3 Information Diffusion Models

Information diffusion model plays a crucial role in the experiments here. This section introduces the mechanism of general information diffusion models.

Scientists from different fields, including sociology, epidemiology, and ethnography, have dedicated themselves for a long time for investigating how information spreads on social networks. Though this mission was as complex as decoding the patterns and processes of human behaviors, some progress has been achieved nevertheless. Benefited from the advances of

anthropology, geography, physics sociology, and other related research fields, several modeling approaches have been suggested for addressing the issue of how information spread from person to person and what factors could affect such information propagation.

In the challenges of modeling diffusion processes, two most popular seminal models for information diffusion have been proposed: The *Independent Cascade model* (IC) (Goldenberg et al., 2001) and the *Linear Threshold model* (LT) (Granovetter, 1978). As a matter of fact, besides those two well-known models, other models do exist and can be used to investigate information diffusion. But many of them are actually variations and extensions of the IC and LT models, including AsIC (asynchronous independent cascade) or AsLT (asynchronous linear threshold) (Satio et al., 2011). AsIC and AsLT deal with diffusion processes in contiguous time instead of discrete time stamps that are characterized by IC and LT.

Other related information models such as SIS (S: Susceptible, I: infected, S: Susceptible) or SIR (S: Susceptible, I: infected, R: Recovered) and their variants are epidemic models and analogues of virus propagation processes (Newman, 2003). Models like SIS or SIR imply non-graphical networks, which means they do not consider the topology of the network. Thus SIS and SIR are not appropriate models when considering social networks or for identifying influential nodes.

3.3.1 Independent Cascade Model

Independent cascade model is a stochastic information diffusion model (Goldenberg et al., 2001). In this model, information spreads over the networks through cascades. Nodes in the social networks have two states in each discrete time step. 1) Active: Nodes that are in such state are

those that are already aware of the information and have adopted it. 2) Inactive: It means that inactive nodes are unaware of the information or are not influenced by the active nodes.

The process that diffuses information through a network often starts with a small set of nodes known as *seed nodes* or *early adopters* (Rogers, 2010). These nodes become active by accepting the information being diffused through the network. At each time step, an active node has a single chance for activating one of its currently inactive connected neighbors. The probability of inactive node becoming active depends on the tie between the two nodes. In general, that probability can be structured as a parameter of the model. Note that each node only has one attempt for trying to activate each of its neighbors, whether it was successful or not. The activation process ends when no more attempts are possible.

3.3.2 Linear Threshold Model

Linear threshold model is different from the independent cascade model in their diffusion mechanism (Granovetter, 1978; Kempe et al., 2003). Specifically, a linear threshold model is receiver-centric while an independent cascade model is sender-centric. (Guille et al., 2013). Again, in a linear threshold model, a set of seed nodes would first be set to attempt the activation of their connected neighbors. Each inactive node has a uniform random threshold [0, 1] as the probability of being influenced. Each neighbor of that node can be given a *weight* to represent how influential it is. At each time step, the probability of successful activation depends on the total weight of its *active* neighbors, i.e., if the total weight of its active neighbors exceed the threshold of the inactive node, it will become active. More formally, an inactive node v becomes active if

$$\sum_{u \in Nbor(v)} w_{u,v} \geq \theta_v$$

Where $Nbor(v)$ denotes the set of active neighbors of the target node v. In addition, $w_{u,v}$ denotes the weight between node u and v. Furthermore, the total weights should satisfy the constraint that $\sum_{u \in Nbor(v)} w_{u,v} \leq 1$. The threshold v of node is denoted by θ_v.The activation process terminates if no more attempts are possible.

CHAPTER IV. ANALYSES AND RESULTS

4.1 Model Description

This section introduces the agent-based model designed using NetLogo and the basic functions that enable the model to manage most of the analyses and experiments. The graphic user interface (GUI) of the agent-based model in NetLogo mainly contains *View* (the visual representation of the NetLogo), *Buttons*, and *Sliders* to control the model and *Monitors* and *Plots* to show data the model is generating (Figure 2).

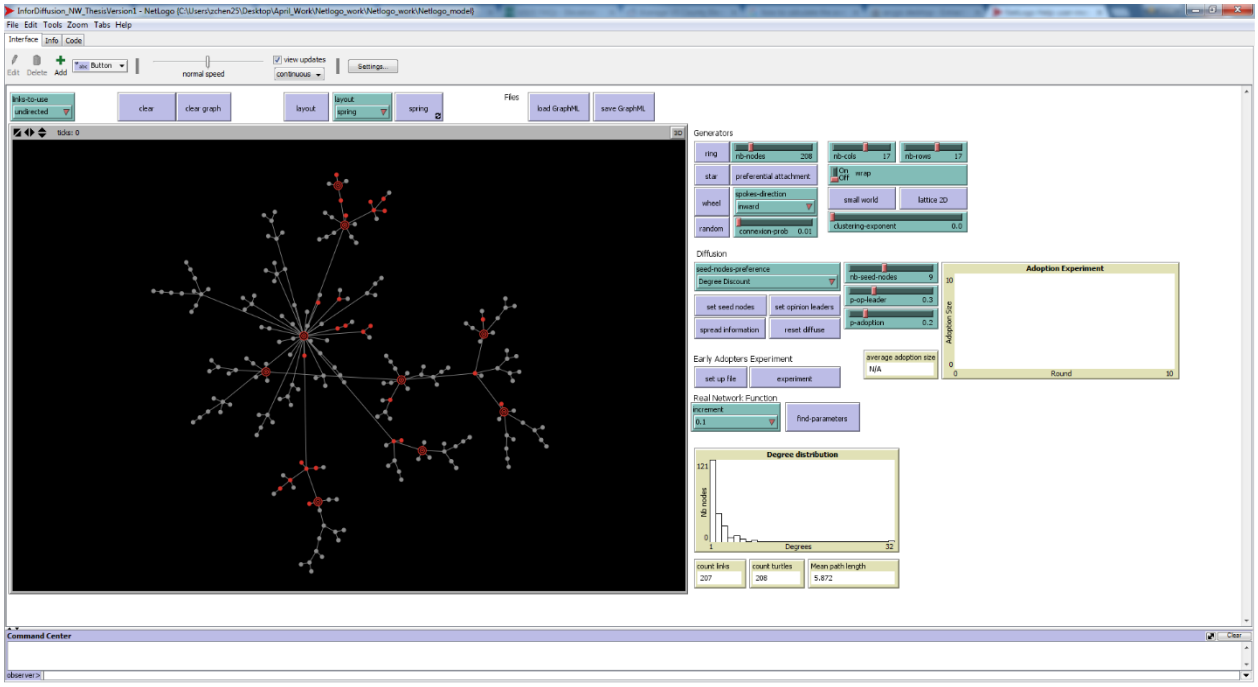

Figure 2: The interface of Netlogo and the integration of models of the information diffusion and network generation using Netlogo

The agent-based model in the thesis consists of two main submodels: (1) a Network Evolution submodel and (2) an Information Diffusion submodel:

(1) A *Network Evolution* submodel (Figure 2) simulates the evolution of a network's structure. The structure of a simulated network is controlled by a set of parameters with regard to the type of the networks. This submodel is designed for manipulating the structure of a network to emulate a real social network as close as possible. This is due to the fact that we cannot change the structure of real social networks so we can only change the structures of simulated networks to emulate the real ones.

It should be noted that an ABM has the ability to import real social networks and generate a visual description of that network. This function can be utilized to validate the accuracy and to calibrate simulated models. In fact, the *Network Extension* of NetLogo enables users to import networks that are in different formats, such as GraphML, VNA, GEXF, etc. In this thesis, GraphML was chosen as the standard format of networks that NetLogo used to load and save network information.

Figure 3 shows an example for the graphic user interface (GUI) of the Network extension in NetLogo that can be used to generate networks. Most of the functions are borrowed from NW-Extension demo: *Network Extension General Demo*[1]. These buttons and sliders allow users to generate the different types of networks they want.

[1] https://github.com/NetLogo/NW-Extension/blob/5.x/demo/Network%20Extension%20General%20Demo.nlogo

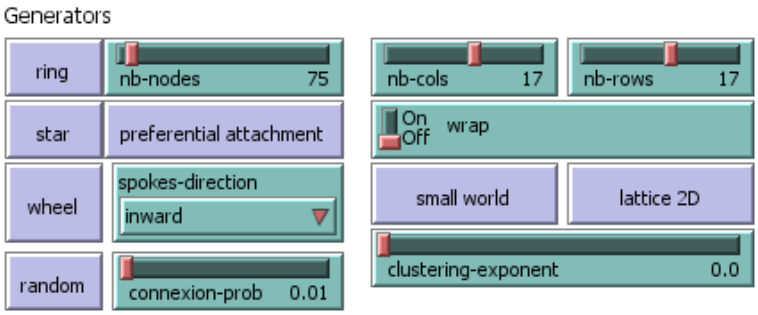

Figure 3: GUI of network generator in NetLogo

There are two ways to generate networks depending on what type of network to create. The left group of buttons (Figure 3) takes the parameters in the slider *nb-nodes* which controls the number of nodes in the network. The right group of buttons (*small world* and *lattice 2D*) takes the parameters from the sliders *nb-cols* and *nb-rows* because the algorithm proceeds by generating a lattice of the given number of rows and columns.

A total of four types of networks may be generated from the Network Evolution submodel to support the experiments. These included varying network structures to generate those such as preferential attachment, random (variant of Erdős–Rényi model), small-world (Kleinberg Model), and lattice networks. These functions have corresponding buttons in the Network extension as shown in Figure 3.

The first network that Network Evolution submodel can generate is *Preferential Attachment model* (Barabási and Albert, 1999; Wilensky, 2005). Basically, a preferential attachment model simulates the formation of some networks, such as connections between websites or the collaborations between actors. These networks have the characteristics of "preferential attachments": a few "hubs" that each has an extremely huge number of connections, while everyone else has only a few. The evolution of such a network starts with two ordinary nodes connected by an edge, indicating that these two

are socially related. At each time step, a new node is added and randomly picks an existing node to connect to, but with some bias. More specifically, an existing node's chance for being selected for connecting with the newly added node is directly proportional to the number of connections it already has, or its "degree." This is known as the "preferential attachment." Because of this mechanism of preferential attachment, the parameter to generate the network is number of nodes only and the number of edges is always one less than the number of nodes. (e.g., N = 200, E = 199)

The second type is *random model* which is a variant of *Erdős–Rényi* model (Erdős & Rényi, 1959). In a random model, it starts with a specific number of nodes and each node has a *connection-probability* (between 0 and 1) of being connected to each of the other nodes. Thus, the parameters for controlling random models include the number of nodes and the connection probability. The third network is of *lattice* or *grid* networks that has the format of placing nodes in a lattice or arranged as a grid. To control the number of nodes in the network, one has to specify the rows-count and column-count in the lattice.

The last model is *small-world* model which is implemented in *Kleinberg* Model (Kleinberg, 2001; Easley & Kleinberg, 2011). The small-world model is created according to a lattice of a given number of rows and columns and subsequently adding additional links between the nodes in the lattice. Small-world network is controlled by a parameter of *clustering coefficients* in addition to *column-count* and *rows-count*. The higher the *clustering-exponent*, the more the algorithm will favor already close-by nodes when adding and rewiring new links.

(2) *Information Diffusion* submodel (Figure 4) is the primary model in the NetLogo model. It simulates and evaluates the information diffusion process so that an efficient strategy could be found to maximize the propagation influence. The information diffusion submodel also supports experiments for the choices of optimal early adopters in terms of their positions and the quantity in the social networks.

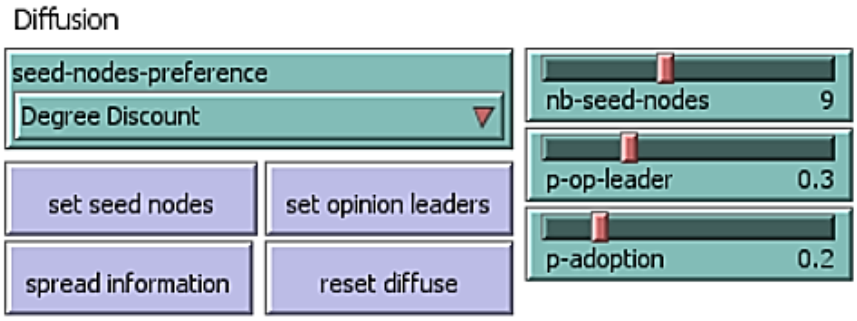

Figure 4: GUI of information diffusion model in NetLogo

The GUI of information diffusion submodel (Figure 4) contains the *Chooser* (*seed-nodes-preference*) which allows user to choose specific method of locating seed nodes/early adopters. With the *Sliders* on the right, users are able to specify the number of seed nodes that will be distributed in the network (*nb-seed-nodes*), the propagation probability from opinion leaders (*p-op-leader*) and the propagation probability from common people (*p-adoption*). After users set these *Chooser* and *Sliders*, they could press the buttons to *set seed nodes* and *set opinion leaders* that assign the early adopters and opinion leaders in the network. This aforementioned procedure is the prerequisite for hitting the button *spread information* that lets the information start to diffuse from the early adopters. Finally, *reset diffuse* cleans the status of adoption and sets the network back to the beginning where early adopters and opinion leaders have been assigned.

Inspired by the two classic models (IC and LT model), information diffusion

submodel incorporates two concepts from them and uses the mixed diffusion model as the major model to simulate information propagation. The idea is that while IC model considers only the source node that sends the message with a probability, the mixed diffusion model takes into account the status of the target node. The model starts with an initial set of seed nodes. Theses seed nodes each has a single chance for activating their neighbors just like IC model. Instead of an arbitrary propagation probability to influence (i.e., activate) their neighbors and similar to LT model, it considers the fraction of active neighbors of the target node. When the aggregated weights of the target nodes and its neighbors exceed the threshold, the target node becomes active.

Having the information diffusion model in place, we could have the ability to simulate the early adopters. Early adopters, also called seed nodes, are those chosen for adopting the information at the beginning of diffusion processes. The purpose of choosing early adopters is to trigger an information cascade that is as large as possible. The methods for choosing early adopters are betweenness centrality, greedy algorithm, closeness centrality, K-shell, eigenvector, and degree discount heuristics. These methods are introduced in the early chapter and used as parameters whose values control the simulation of the agent-based models in this thesis research. To run the diffusion model, it is also required to specify the parameters such as p_{op} and p_n which are propagation probability of opinion leaders and common people respectively.

Different from early adopters, *opinion leaders* are those with high centrality (high degree) in their local networks and they are more influential in spreading the information. Because of its position and popularity in the network, opinion leaders would have more connections and better chance for influencing its followers. Thus, the probability of

successfully influencing a node by another mainly depends on the role and function of a node in a network.

Even though opinion leaders may not be early adopters in the social networks for maximizing their influence, their role as opinion leaders in the social network is still important. Traditional opinion leaders serve a critical role of mediating information diffusion. They function in both political and economic ways in different scenarios. In marketing, for example, opinion leaders help to disseminate a product or innovation quickly in their communities than any great salesman (e.g., via "word of mouth" effect). Same thing happens in political campaigns. Political leaders usually play the role of opinion leaders that advocate their political stands. As compared with common people, opinion leaders often have more social connections and are always seen as trustworthy information sources *(*Katz*,* 1957*)*. They facilitate the diffusion of information because their existence improves the probability of adopting information by the public.

There is a large body of literature addressing the problem of identifying opinion leaders (Valente & Pumpuang, 2007; Ning et al., 2012; Bodendorf & Kaiser, 2009; Aral & Walker, 2012). Some utilize self-identification (Valente, 1999) while the other employ network centrality to locate opinion leaders (Ma et al., 2012). Among the problems of network centrality, different methods are used to identify local and global opinion leaders. *Global opinion leaders* ensure the information extent to include many communities and *local opinion leaders* facilitate the information adoption within the communities. In the present research, the number of opinion leaders is assumed to be 10 percent of the general public based on a survey. (Doumit et al., 2011) The methods of locating them in the network is simply degree centralities for their simple role in the

situation. Indeed, chances are that some common people may be the focus of the topic because they are at presence in the event or they are involved in the incident. In this case, these people are either early adopters nor opinion leaders, but their messages could be disseminated by many people. It is rather difficult to detect their locations because these people could be anywhere in the networks. Considering it is not very commonly occurred in real topics, simulating such situation is not implemented in this work.

In summary, opinion leaders are a group of people that are popular (well connected) in their social network (i.e., community). They usually play an essential role in a vast information cascade. It is assumed that opinion leaders are also those who possess the most connections in their communities, which makes them more identifiable. People in the same community trust or follow their opinion leaders. Therefore, if the opinion leadership is considered in an information diffusion submodel, it would be more possible to unfold the truth of diffusion mechanism.

In order to find the optimal early adopters in the network, experiments were set up in the Netlogo model. (Figure 5). Together with the functions depicted in Figure 4, users who ran the experiment needed to select the method for choosing early adopters as well as defining the propagation probabilities. Since the goal of the experiment was to find the optimal number of early adopters for each network, users did not need to specify the number of seed nodes in the network as the experiment would go through each number of early adopters from 1 to 20 automatically. The procedure in NetLogo is rather simple: *set up file* to create the file that saves the results of each simulation in the experiment and *experiment* to start the experiment that goes through each number of early adopters from

1 to 20 under the propagation probability and methods of choosing early adopters that introduced in Figure 4.

Figure 5: Early adopters experiment to find optimal early adopters in the network under different propagation probabilities

Although none of the two submodels is an agent-based model in their conventional forms, they can be translated to an agent-based framework without much effort based on the discussion above. In fact, these transformations have been carried out before (Rand & Rust, 2011; Herrmann et al., 2013). These works assumed that information started from a number of early adopters, these early adopters might or might not be opinion leaders. Early adopters were chosen by using algorithms or heuristic for maximizing their influence (Chen et al., 2009; Kempe, et al., 2003). For the purpose of comparison, approaches of using network analysis (centrality) to identify seed nodes were also implemented.

4.2 Network Topology and Spread Efficiency

This section examines the topology of networks and tries to answer the questions: what type of networks (exclude fully-connected) better facilitates information diffusion? What unique characteristics does it possess if it does?

To date, there are many measurements that have been developed to characterize network topology. Table 1 displays some of the commonly recognized characteristics in a network. These include modularity (strength of division of a network into modules), average clustering coefficient (a measure of how complete the neighborhood of a node is), diameter (the shortest

distance between the two most distant nodes in the network), and average path length (average number of steps along the shortest paths for all possible pairs of network nodes). Four of the listed networks are artificial networks generated by *Network Evolution* submodel with the same number of nodes (in this case, $N = 800$). The real network was derived from tweets about Bernardo wildfire as introduced in Data and Methods chapter.

In order to assess how these networks perform, in terms of information diffusion, the networks were imported into information diffusion submodels and let them run in multiple diffusion simulations. The success (i.e., efficiency) of information diffusion was measured by the total number of adopted individuals by the end of a diffusion process.

Table 1. lists the topological characteristics of four artificial networks and one real network studied in this thesis research. The characteristics of different networks include the number of nodes (N), the number of edges (E), average degree, density, modularity, average clustering coefficient, average path length and diameter.

Table 1: Lists of the topological characteristics of four artificial networks and one real network studied in this thesis research

Network	N	E	Avg.degree	Density	Modularity	Avg.C.Coefficient	Avg.path length	Diameter
Preferential	800	799	1.998	0.002	0.924	0	8.8857	22
Lattice	800	1543	3.858	0.005	0.802	0	19	55
Random	800	3145	7.862	0.001	0.31	0.01	3.4611	6
Small world	800	2337	5.842	0.007	0.486	0.007	4.3068	7
Real network	1,300	8,025	5.899	0.005	0.253	0.124	3.483	10

In the experiments of four artificial networks, random network performed the best in terms of influence maximization. As an example, in the 800-node random networks, with probability of diffusion p_{op}= 0.4 and p_n= 0.3, about 90% of the population were influenced even though there was only one seed node (or early adopter) at the beginning. While in other

networks, adoption rate was 72% for small world, was 4% for preferential attachment, and was 1% for lattice respectively.

It is interesting to notice that the adoption rate for preferential attachment was greater than that of lattice networks. When there were few seed nodes, preferential attachment showed greater adoption rates. However, such advantage of preferential attachment started fading off when more seed nodes were added. This was probably due to the different nature of these two types of networks. Preferential attachment networks follow the power-law degree distribution in that most nodes in the network have few connections with a small number of nodes having most of the connections. Lattice networks, on the other hand, tend to have an even distribution of degrees of nodes, with most of the nodes having degree of four, three, and two. When more early adopters were selected from nodes with top centrality as simulations proceeded, there were fewer influential nodes left in the pool for influential nodes in the preferential attachment network. As opposed to this, in the lattice network, this did not seem to cause any problem because most nodes would have similar influence.

From Table 1, it is obvious that the number of edges and average degree in a network were not the factors that affected the adoption rate when the number of seed nodes t were limited. For example, the Lattice network with $N = 800, E = 1543$ and seed set size (t) = 10 was shown to have 76 diffusion influence. Comparing to this, a preferential network with $N = 800$, $E = 799$ and t = 10 was shown to have nearly 129 diffusion influence already. It should be noted, however, when we looked at the relationship between average path lengths, average clustering coefficients or diameters, and diffusion influence, there seemed to exist negative correlations.

First, to confirm the correlation between average path length ι and adoption rate f (or diffusion influence), an experiment was conducted. A small world network was used for testing the correlation. In the experiment, same numbers of nodes ($N = 400$) were generated 20 times using the small-world network generator but with an increasing clustering-exponent. These led to generated networks to have different average path lengths. Diffusion simulations were executed 1,000 times for each network so that the diffusion influence can be collected under each ι.

Path length in a graph represents the geodesic distance of the shortest path between two nodes in a network. The average path length is calculated by averaging the lengths of all possible pairs of network nodes. By definition, let G to be the graph with the set of vertices (nodes) V and $d\ (v1, v2)$ being the shortest distance between v_1 and v_2, where v_1, $v_2 \in$ V. Therefore, the average path length is:

$$\iota_G = \frac{1}{n \cdot (n-1)} \cdot \sum_{i \neq j} d(v_i, v_j)$$

Average path length is one of three robust measures that describe the social network, along with degree distribution and clustering coefficients.

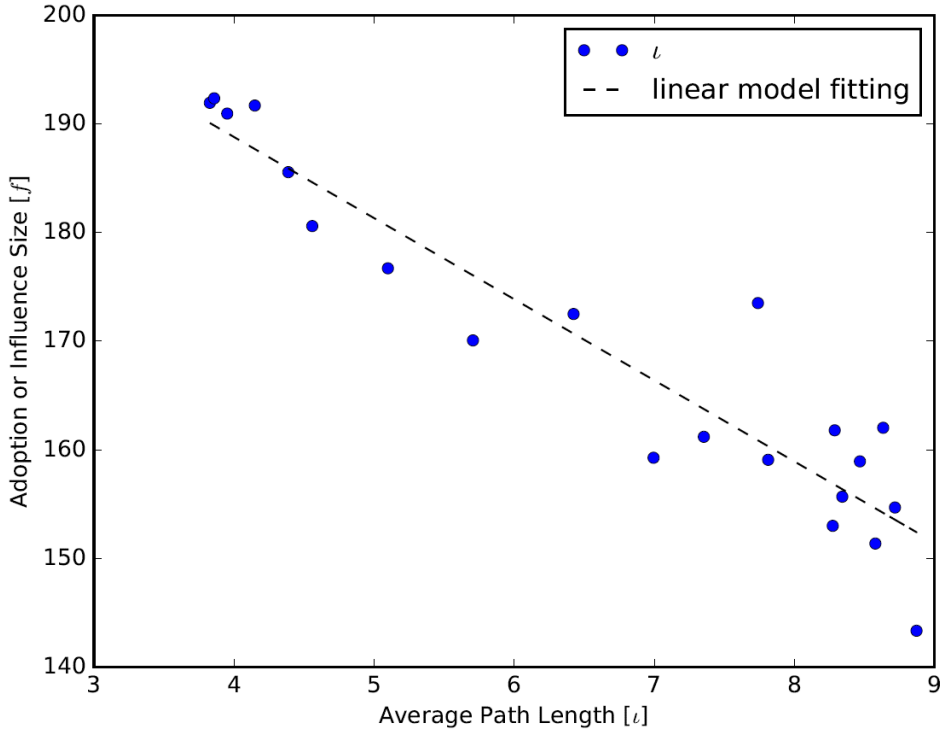

Figure 6: Scatter plot of average path length ι versus influence size f; the black dashed line is the linear regression line. Experiment is based on an artificial small-world network with $N = 400, p_{op} = 0.3, p_n = 0.2$. *It shows ι is a good predictor for f and the correlation is significant. (* $p < 0.01$*)*

Figure 6 clearly shows a linear negative correlation between average path length and adoption or influence size in the simulated information diffusion. The network that has a shorter average path length tends to trigger a bigger information adoption cascade. The result corresponds to what is expected because a small average path length means that nodes are easily reachable between each other and the information could be easily spread to all of nodes. This explained in some way why many social networks have relatively small average path length. In order to be well informed, we tend to form our networks with small average path length. This phenomenon is the well-known 'small-world'.

Using the data from the same experiment, similar analysis was performed for average clustering coefficients. Pearson Product Moment correlation coefficient (Figure 7) was used to

measure the linear relationship between average clustering coefficients and levels of diffusion influence. With the same number of nodes and close number of edges, average clustering coefficients have a negative correlation with adoption size. It implies that the more clustered a social network is, the more it may impede the spread of message within the social network. It is an interesting phenomenon and is the opposite to the conventional intuitive assumption that the more clustered network, the easier information can be transmitted.

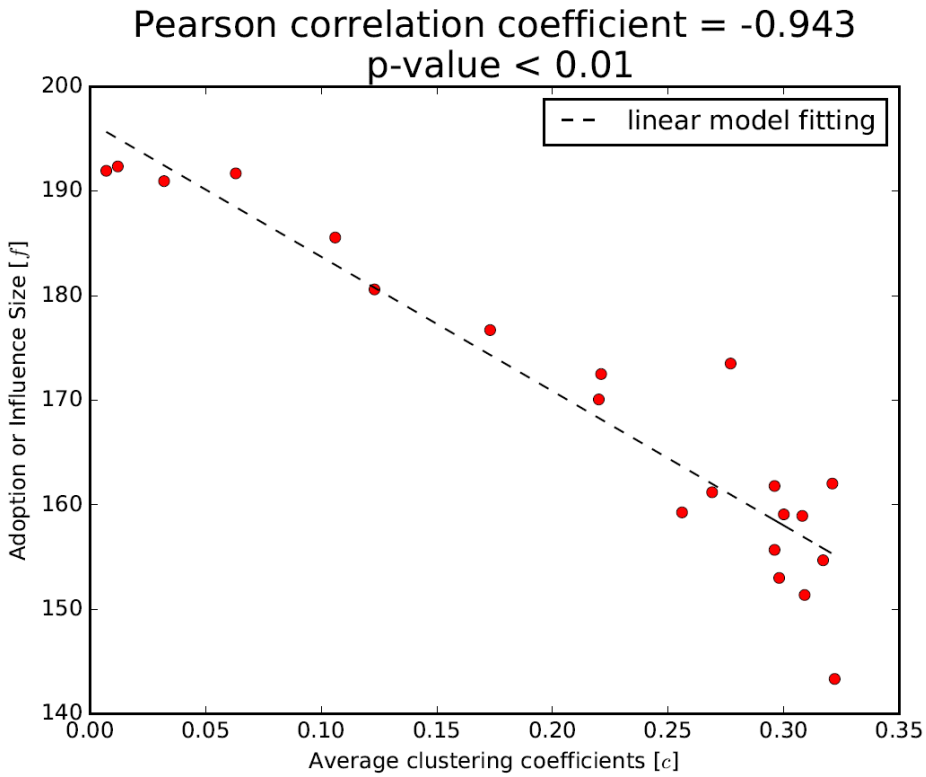

Figure 7: Scatter plot of average clustering coefficients c and adoption size f

As can be seen in Figure 7, the network that has smaller average clustering coefficient tends to trigger a bigger information adoption cascade. The result deviates from what was expect because it is common to think that the more cluster a network is, the easier the information can propagate.

More analyses were conducted in order to solve this puzzle. First of all, a review of the definition of clustering coefficient is necessary. The clustering coefficient measures the extent to which one's friends are also friends of each other. It is computed by dividing the number of triangular connections containing node v by the number of possible connections between its n neighbors . The possible clustering coefficient value ranges from 0 to 1. When a node v's n neighbors have no connections to each other (e.g. a star network), the clustering coefficient is 0. When n neighbors are fully-connected, i.e., everyone of n connects to each other in n, the clustering coefficient then is 1.

Given G as the graph with the set of vertices (nodes) V and edges (connections) E, a clustering coefficient c of a node $v \in V$ is:

$$c_v = \frac{N_{v \in \Delta}}{N_{v \in \Lambda}}$$

Where $N_{v \in \Delta}$ denotes the number of triangular connections that contains v, and $N_{v \in \Lambda}$ denotes the number of triplets (of nodes) that contain v. In a graph, a triangle means a subgraph with 3 edges and 3 vertices. Triplet stands for a subgraph with 2 edges and 3 vertices, one of which is v and such that v is an *incident* to both edges, which means both edges share the same end node v.

Having defined the clustering coefficient of node v, the average clustering coefficient of the graph is simply the average of clustering coefficient values calculated for all n nodes :

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} c_i$$

In order to derive a higher clustering coefficient value while keeping the number of vertices and edges the same, compensation has to be made by rewiring the edges in the graph such that the neighbors of any given node v are tightly connected. While this process increases

the efficiency of information exchange within the community (a group of well-connected nodes), interaction between two communities, however, may be impeded because of the rewiring. Overall, it becomes more difficult to spread the information to more communities when the in-community clustering coefficient values increase while the between-community clustering coefficient values decrease. This may explain the reason why a network with higher average clustering coefficient has less adoption size at the end of the information diffusion.

4.3 Centralities and Heuristics Experiment

In this section, different centralities and heuristics are described and discussed to help understand how choosing early adopters in each of the four types of networks may make information diffusion more effective. These methods include finding a balance between betweenness centrality, greedy algorithm, closeness centrality, K-shell, eigenvector and degree discount heuristics. All of these centralities were tested on four network structures: preferential attachment, random, small-world, and lattice networks, with different number of nodes (Table 2). Additionally, each network structure was used in the simulations with three different sets of propagation probabilities: a) $p_{op} = 0.4, p_n = 0.3$, b) $p_{op} = 0.3, p_n = 0.2$ and c) $p_{op} = 0.2, p_n = 0.1$.

Table 2: Characteristics of tested networks

Network	E	Avg.degree	Density	Modularity	Avg.C.coefficient	Avg.path length	Diameter
				N = 800			
Preferential	799	1.998	0.002	0.924	0	8.8857	22
Lattice	1543	3.858	0.005	0.802	0	19	55
Random	3145	7.862	0.001	0.31	0.01	3.4611	6
Small world	2337	5.842	0.007	0.486	0.007	4.3068	7
				N = 400			
Preferential	399	1.995	0.005	0.891	0	6.547	14
Lattice	760	3.8	0.01	0.76	0	13.333	38
Random	1608	8.04	0.02	0.306	0.019	3.102	6
Small world	1159	5.795	0.015	0.489	0.018	3.876	6
				N = 200			
Preferential	199	1.99	0.01	0.856	0	6.041	13
Lattice	370	3.7	0.019	0.699	0	10	28
Random	508	5.08	0.026	0.405	0.019	3.435	7
Small world	563	5.63	0.028	0.44	0.233	3.427	6

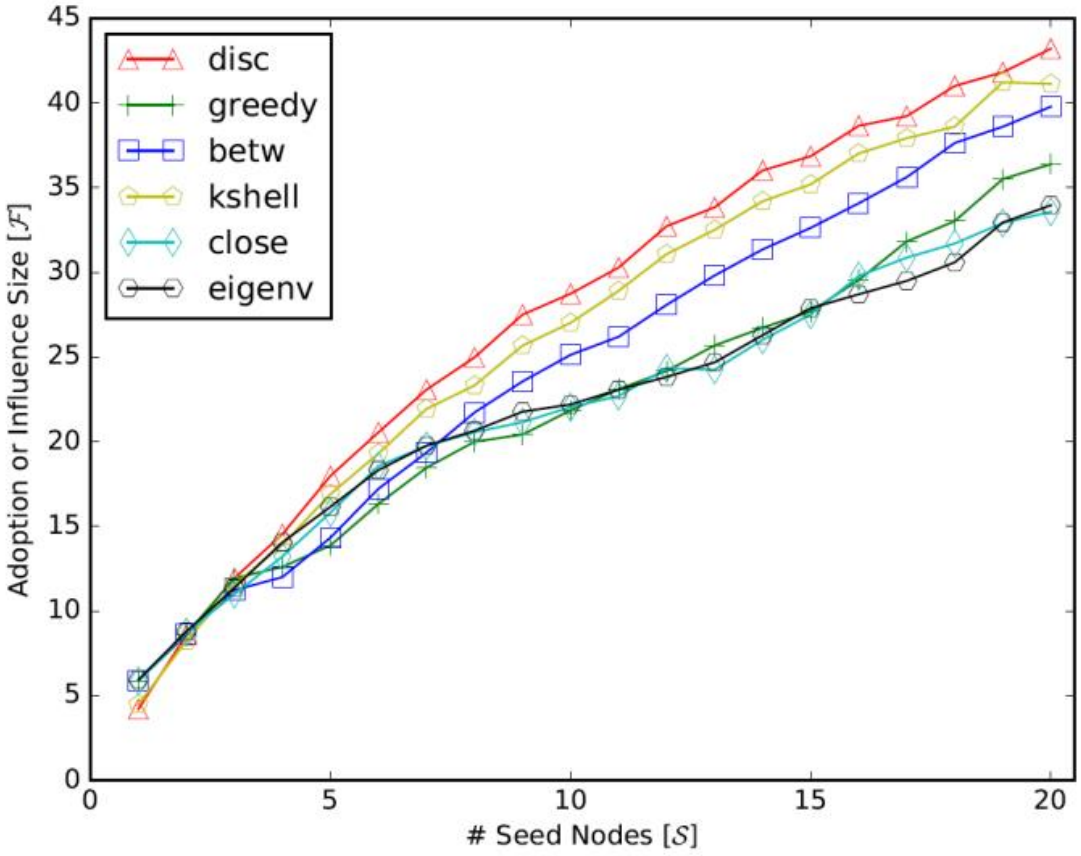

Figure 8: Information diffusion on preferential attachment artificial network with six centralities and heuristics (Disc: degree discount; Greedy: greedy algorithm; Betw: betweeness centrality; kshell: K-shell; close: closeness centrality; eigenv: eigenvector centrality). The total number of nodes is $N = 200$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.2, p_n = 0.1$

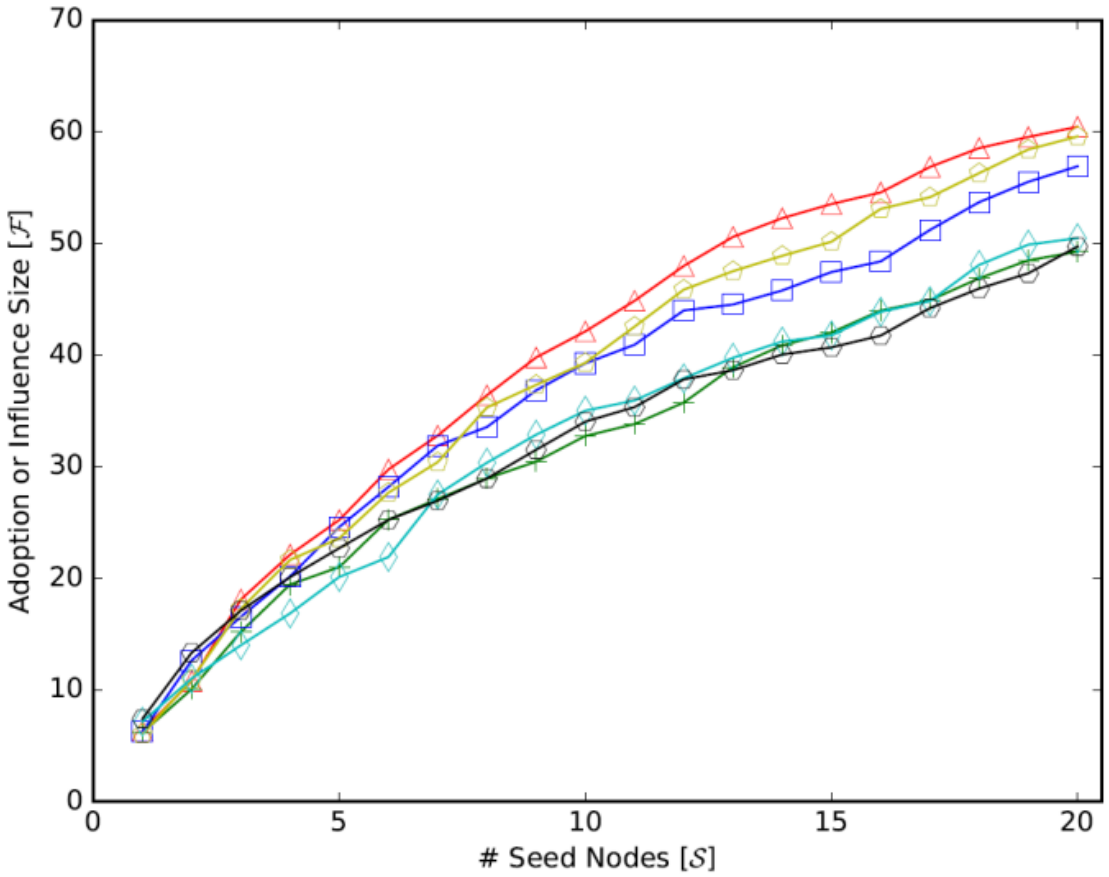

Figure 9: Information diffusion on random artificial network with six centralities and heuristics. The total number of nodes is $N = 200$; Propagation probabilities for opinion leaders and normal people are $p_{op} = 0.2, p_n = 0.1$

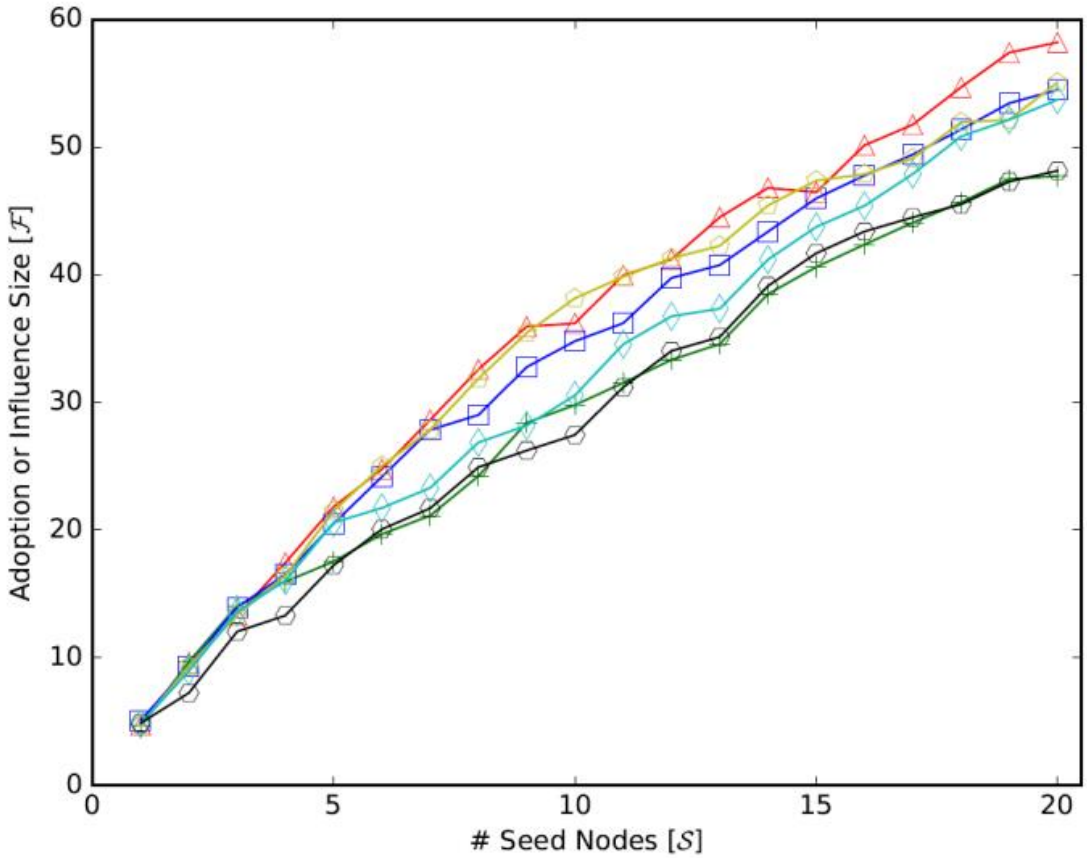

Figure 10: Information diffusion on small-world artificial network with six centralities and heuristics. The total number of nodes is $N = 200$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.2, p_n = 0.1$

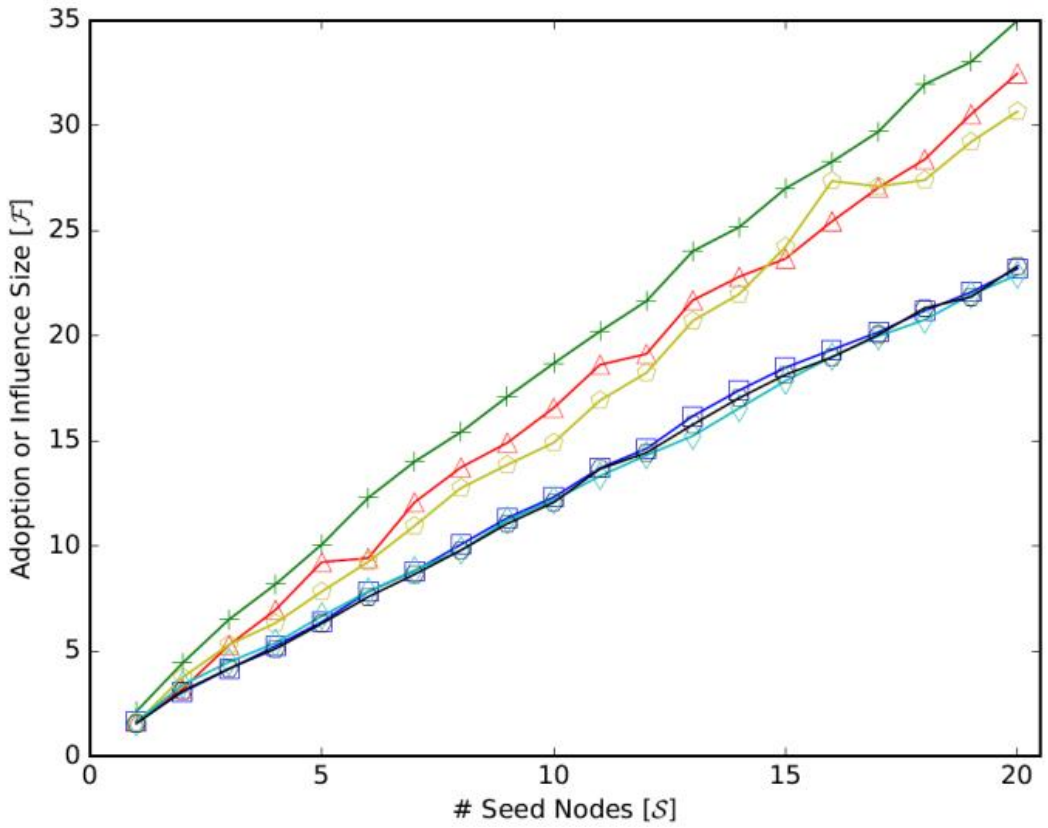

Figure 11: Information diffusion on lattice artificial network with six centralities and heuristics. The total number of nodes is N=200; Propagation probabilities for opinion leaders and normal people are p_op=0.2,p_n=0.1

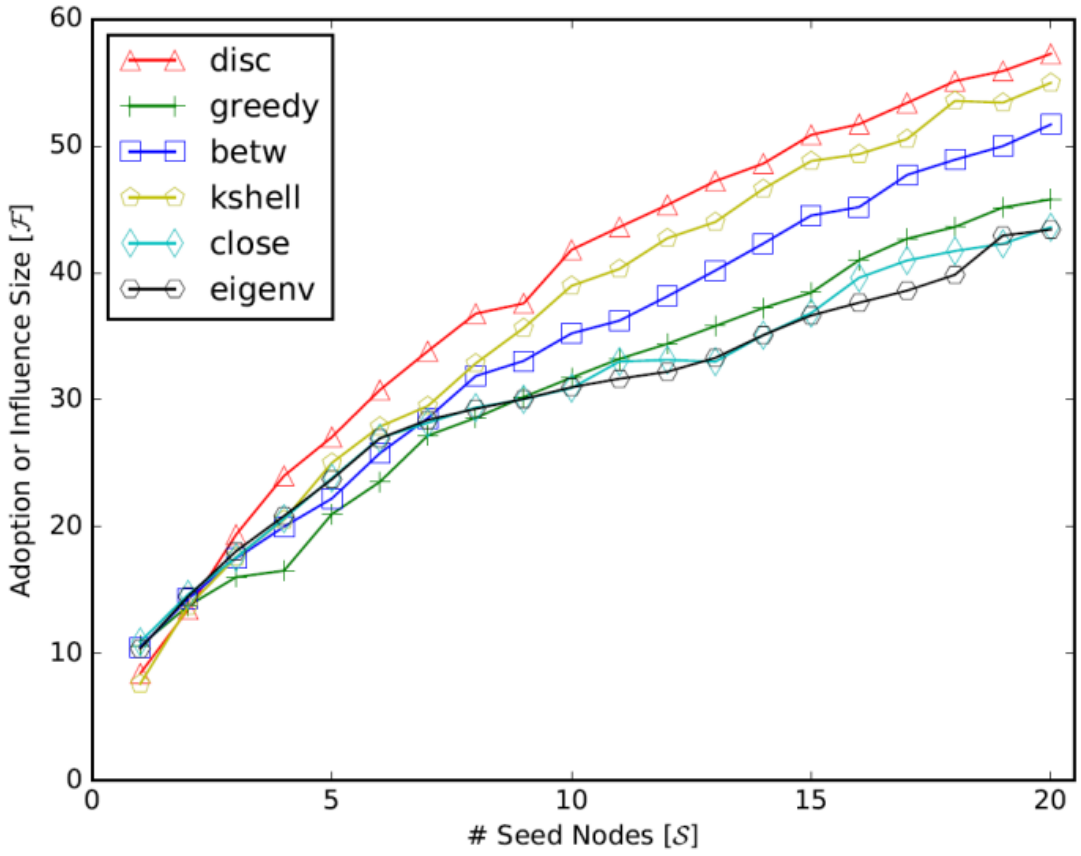

Figure 12: Information diffusion on preferential atttachment artificial network with six centralities and heuristics (Disc: degree discount; Greedy: greedy algorithm; Betw: betweeness centrality; kshell: K-shell; close: closeness centrality; eigenv: eigenvector centrality). The total number of nodes is $N = 200$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.3, p_n = 0.2$

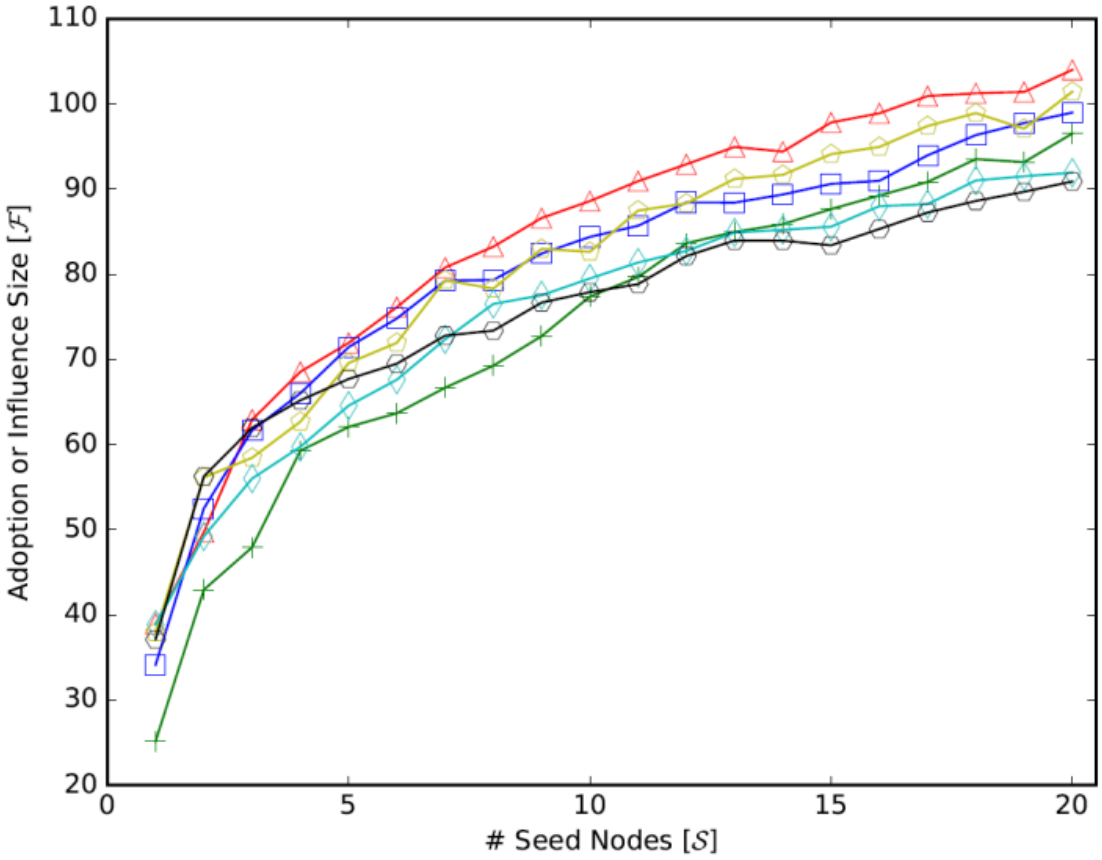

Figure 13: Information diffusion on preferential atttachment artificial network with six centralities and heuristics. The total number of nodes is $N = 200$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.3, p_n = 0.2$

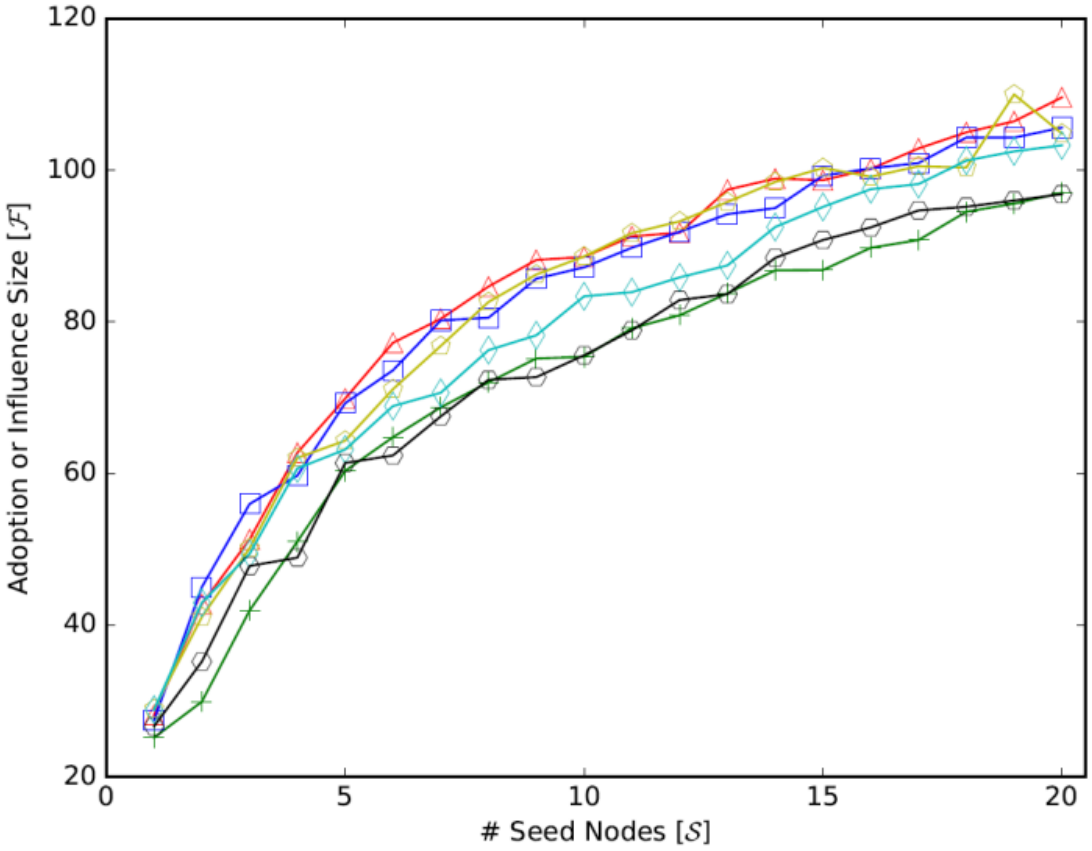

Figure 14: Information diffusion on small-world network with six centralities and heuristics. The total number of nodes is $N = 200$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.3, p_n = 0.2$

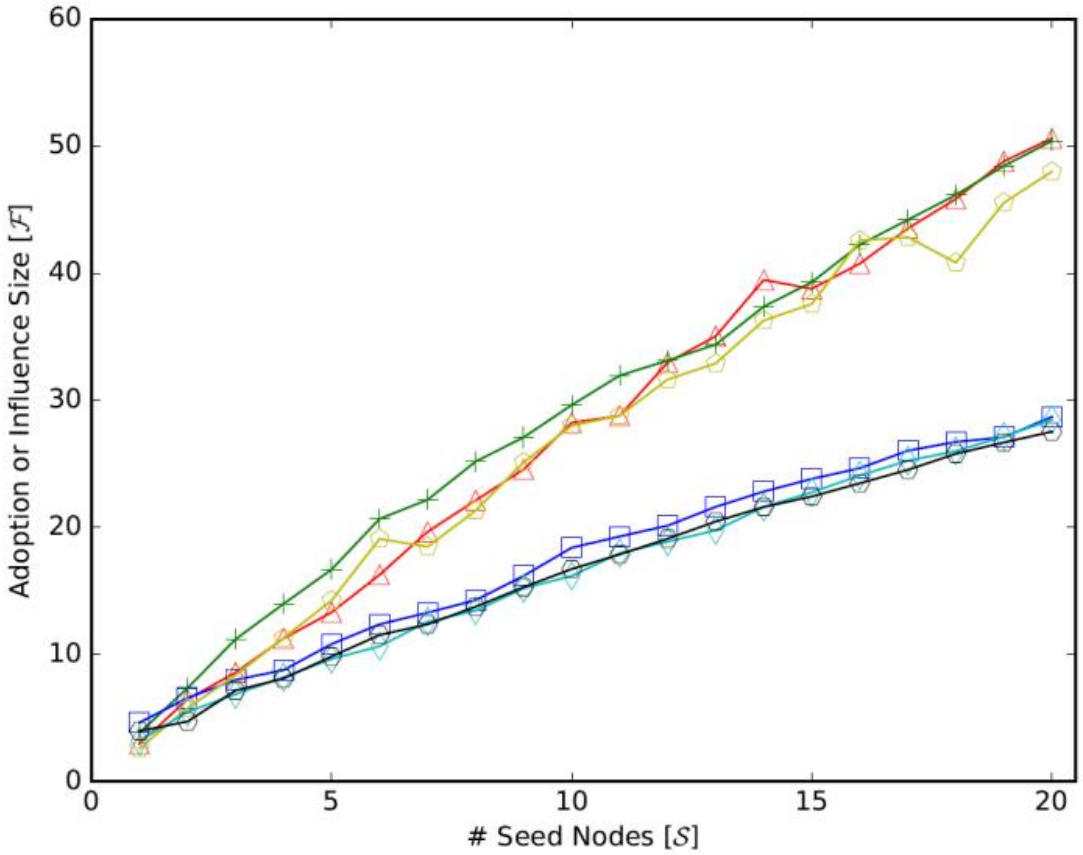

Figure 15: Information diffusion on lattice artificial network with six centralities and heuristics. The total number of nodes is $N = 200$; Propagation probabilities for opinion leaders and normal people are $p_{op} = 0.3, p_n = 0.2$

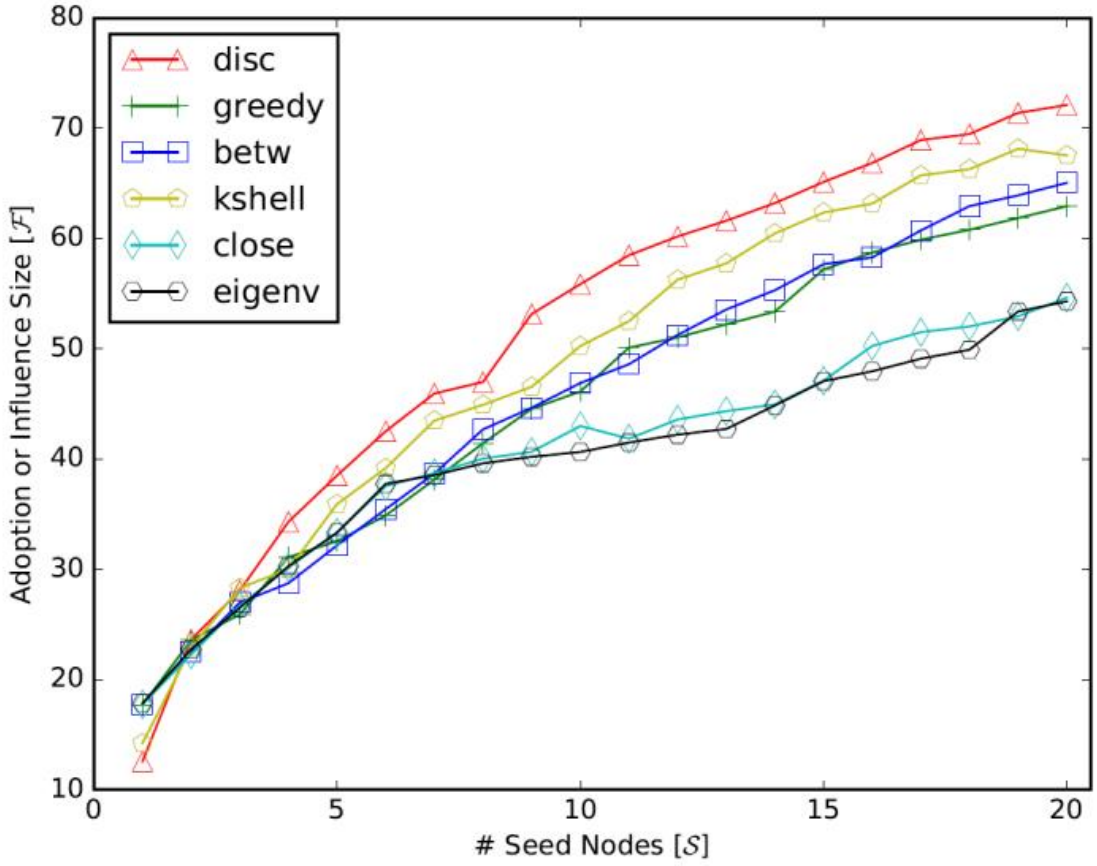

Figure 16: Information diffusion on preferential attachment artificial network with six centralities and heuristics (Disc: degree discount; Greedy: greedy algorithm; Betw: betweeness centrality; kshell: K-shell; close: closeness centrality; eigenv: eigenvector centrality). The total number of nodes is $N = 200$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.4, p_n = 0.3$

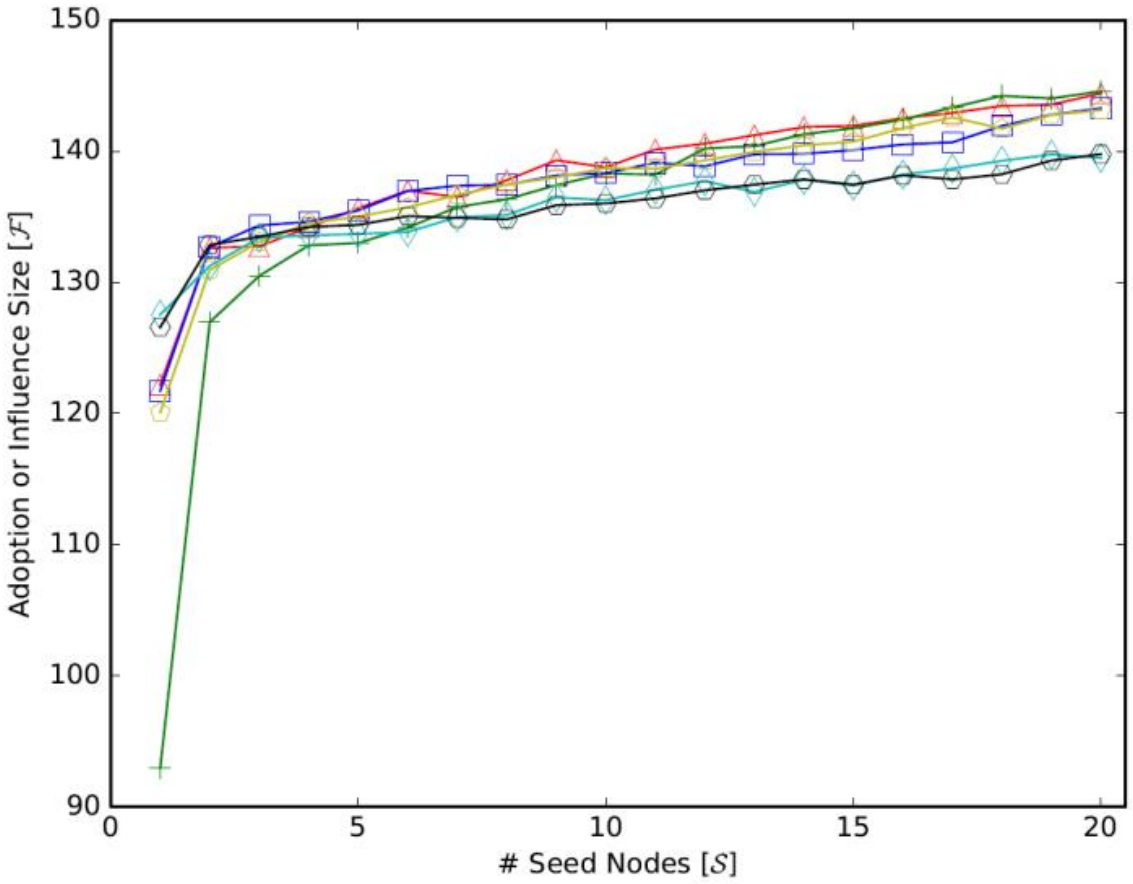

Figure 17: Information diffusion on random artificial network with six centralities and heuristics. The total number of nodes is $N = 200$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.4, p_n = 0.3$

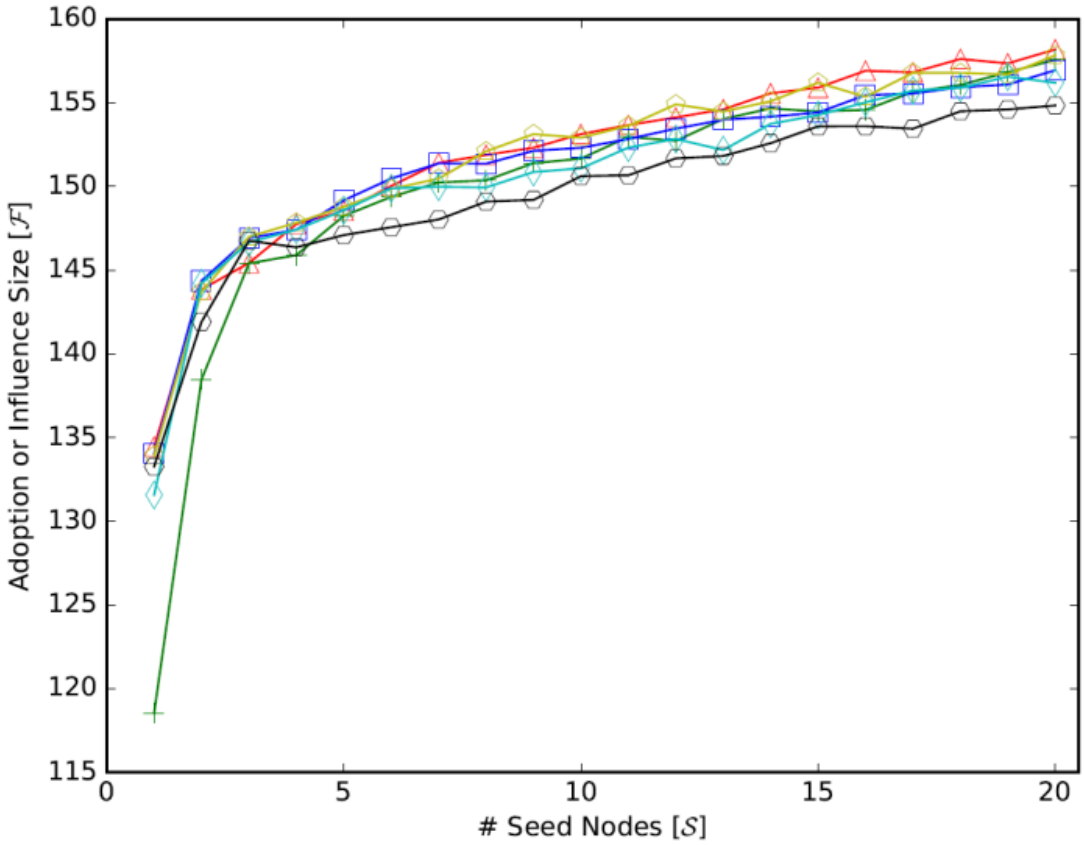

Figure 18: Information diffusion on small-world artificial network with six centralities and heuristics. The total number of nodes is $N = 200$; Propagation probabilities for opinion leaders and normal people are $p_{op} = 0.4, p_n = 0.3$

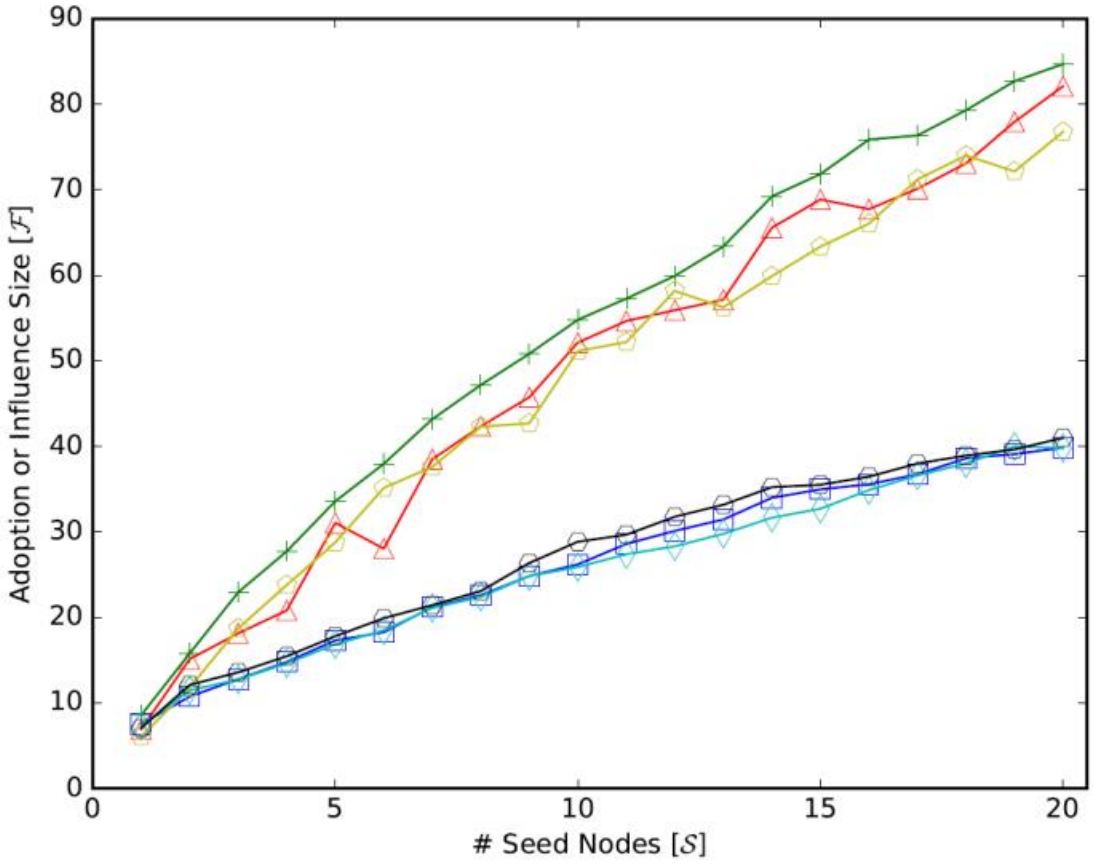

Figure 19: Information diffusion on lattice artificial network with six centralities and heuristics. The total number of nodes is $N = 200$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.4, p_n = 0.3$

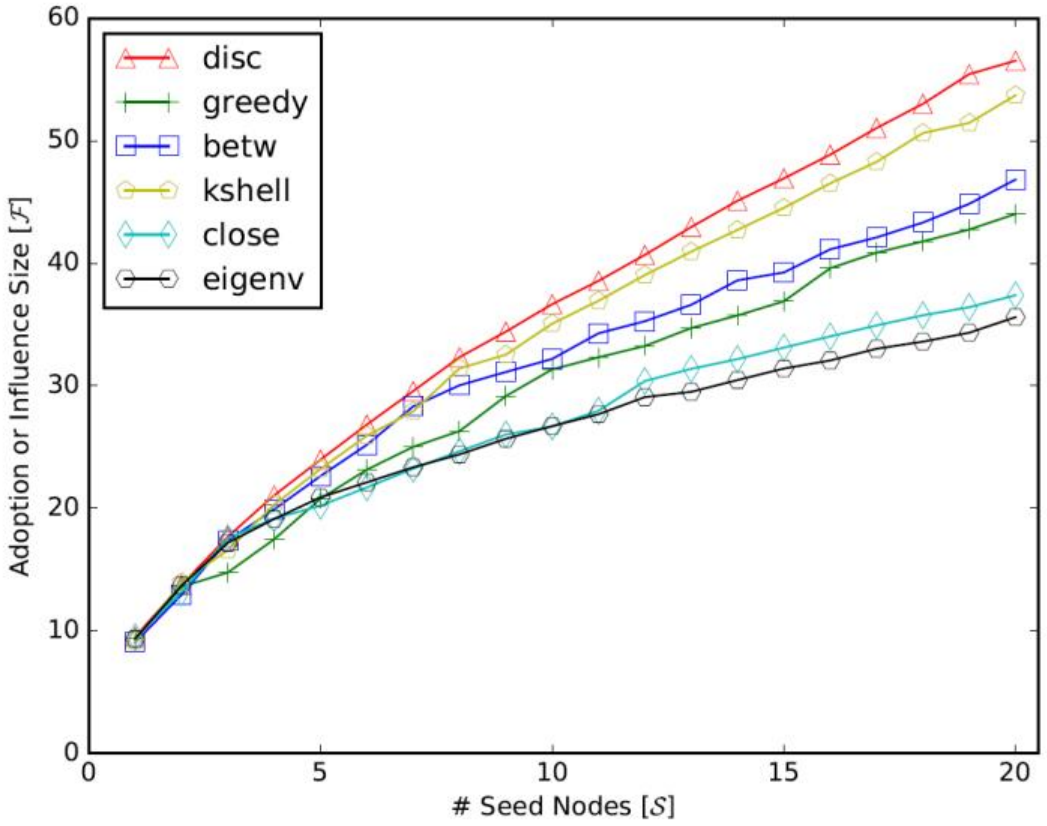

Figure 20: Information diffusion on preferential attachment artificial network with six centralities and heuristics (Disc: degree discount; Greedy: greedy algorithm; Betw: betweeness centrality; kshell: K-shell; close: closeness centrality; eigenv: eigenvector centrality). The total number of nodes is $N = 400$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.2, p_n = 0.1$

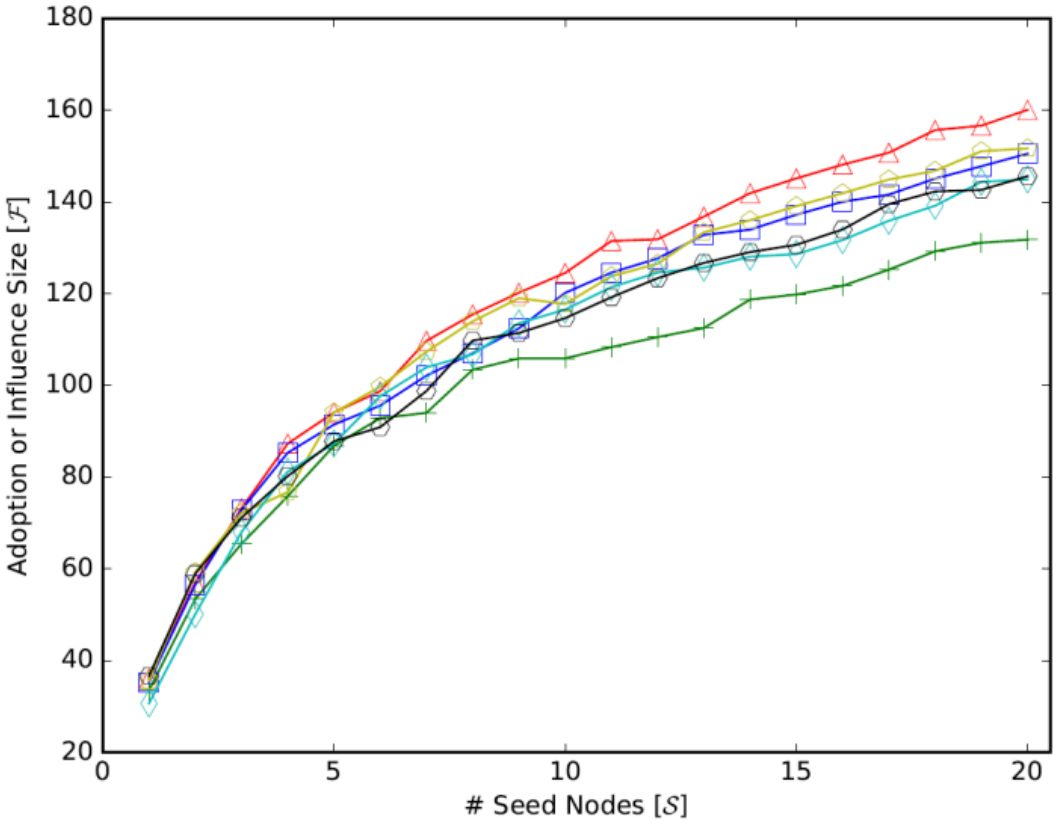

Figure 21: Information diffusion on random artificial network with six centralities and heuristics. The total number of nodes is $N = 400$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.2, p_n = 0.1$

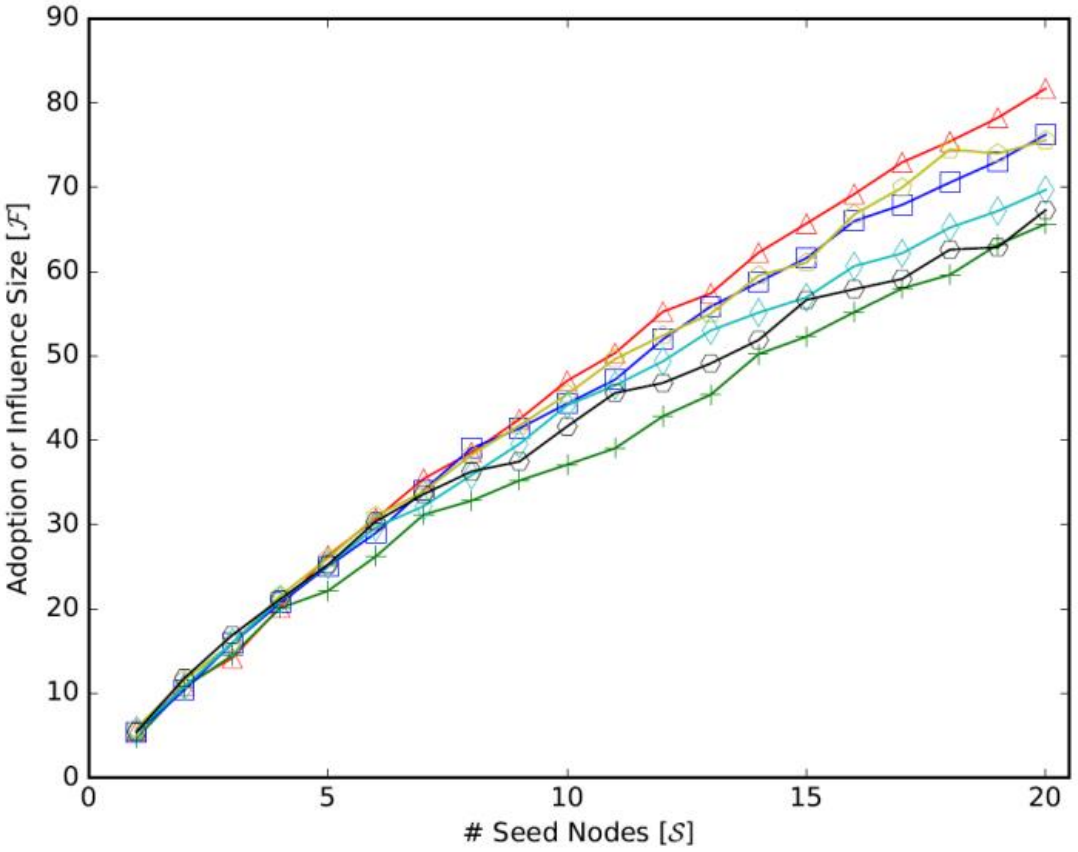

Figure 22: Information diffusion on small-world artificial network with six centralities and heuristics. The total number of nodes is $N = 400$; Propagation probabilities for opinion leaders and normal people are $p_{op} = 0.2, p_n = 0.1$

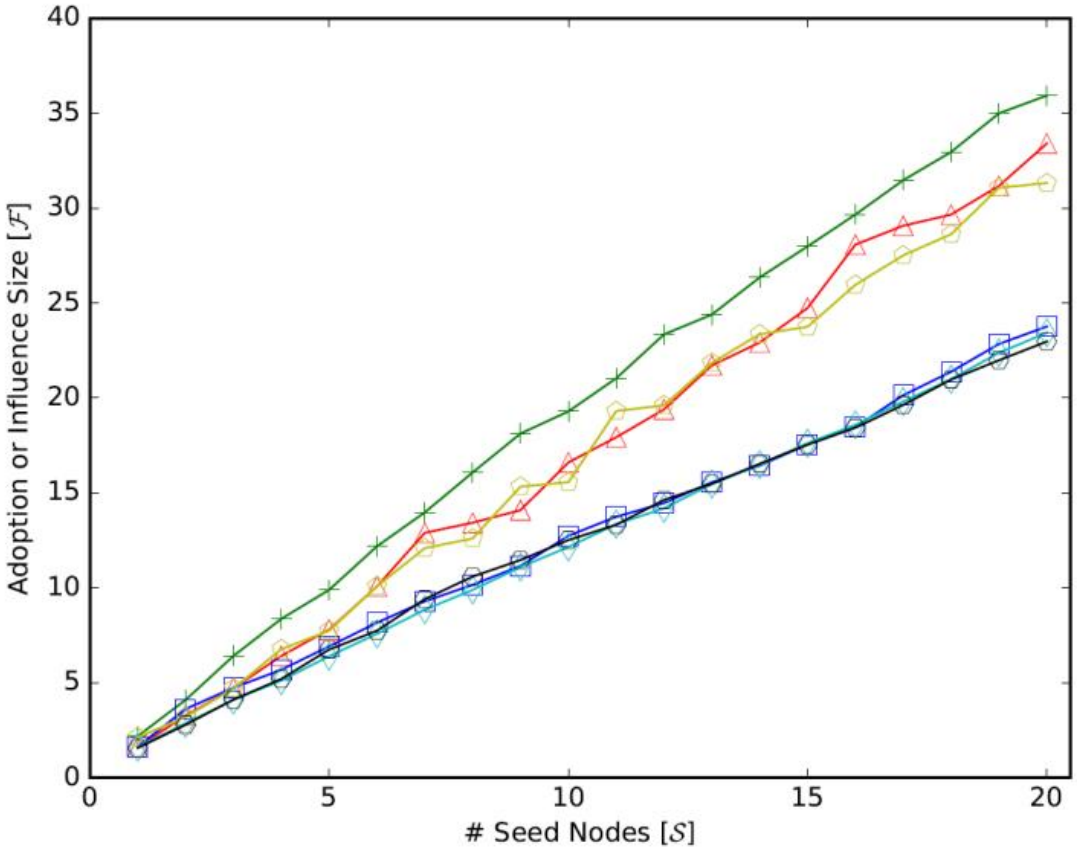

Figure 23: Information diffusion on small-world artificial network with six centralities and heuristics. The total number of nodes is $N = 400$; Propagation probabilities for opinion leaders and normal people are $p_{op} = 0.2, p_n = 0.1$

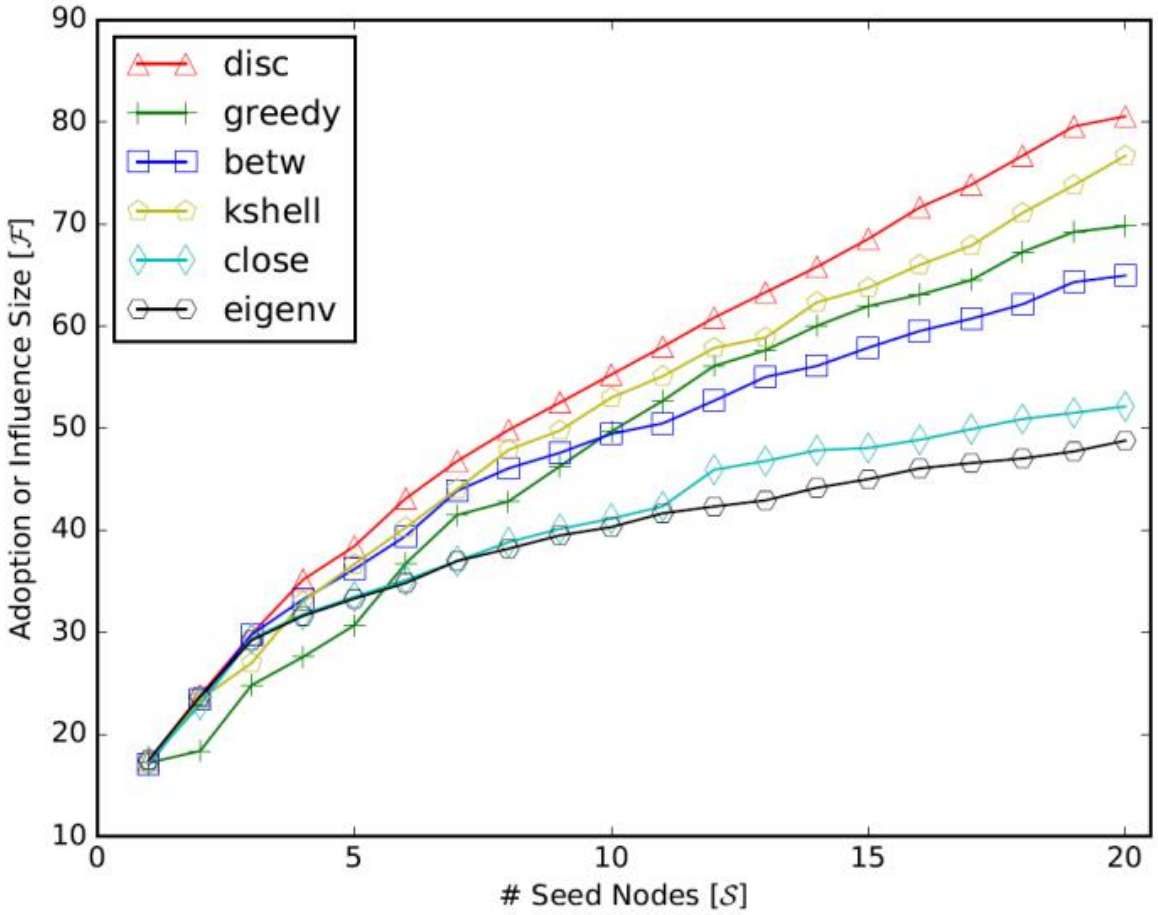

Figure 24: Information diffusion on preferential attachment artificial network with six centralities and heuristics (Disc: degree discount; Greedy: greedy algorithm; Betw: betweeness centrality; kshell: K-shell; close: closeness centrality; eigenv: eigenvector centrality). The total number of nodes is $N = 400$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.3, p_n = 0.2$

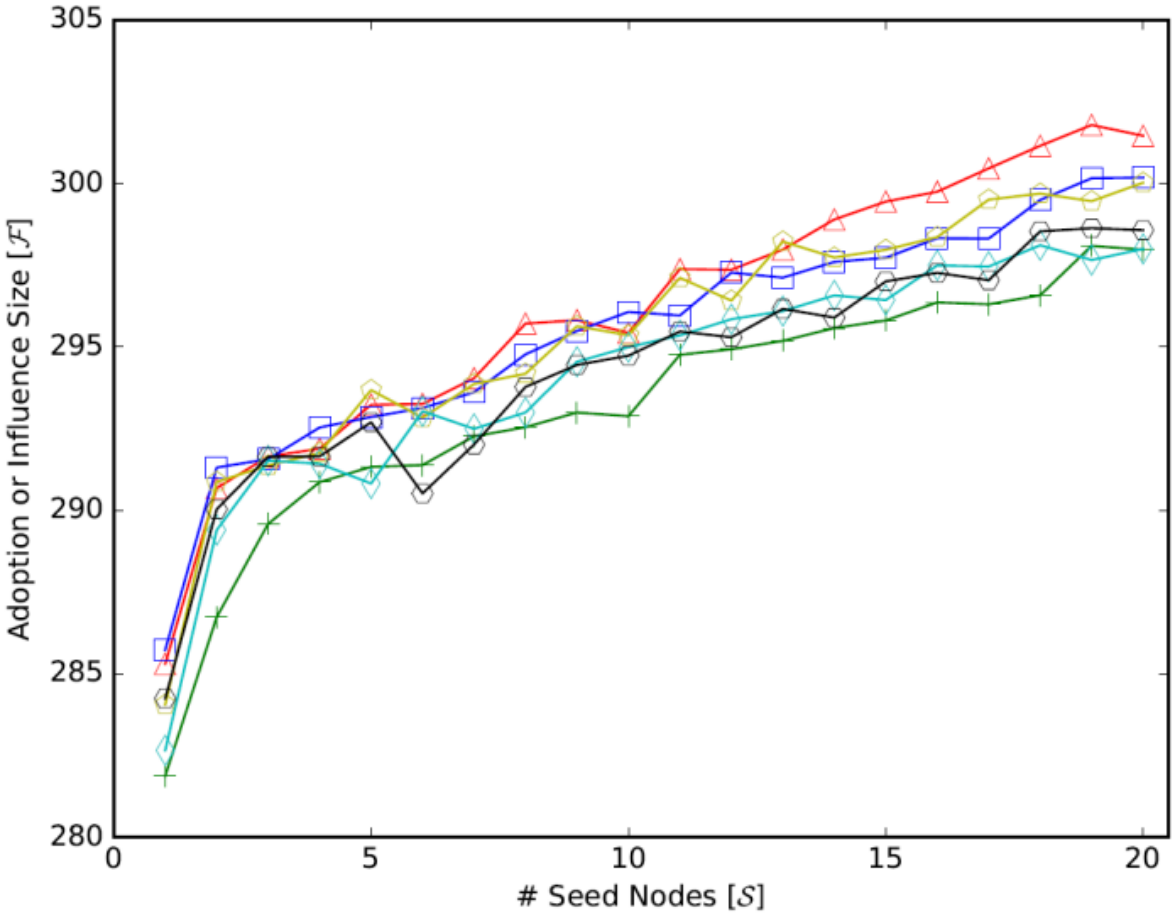

Figure 25: Information diffusion on random artificial network with six centralities and heuristics. The total number of nodes is $N = 400$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.3, p_n = 0.2$

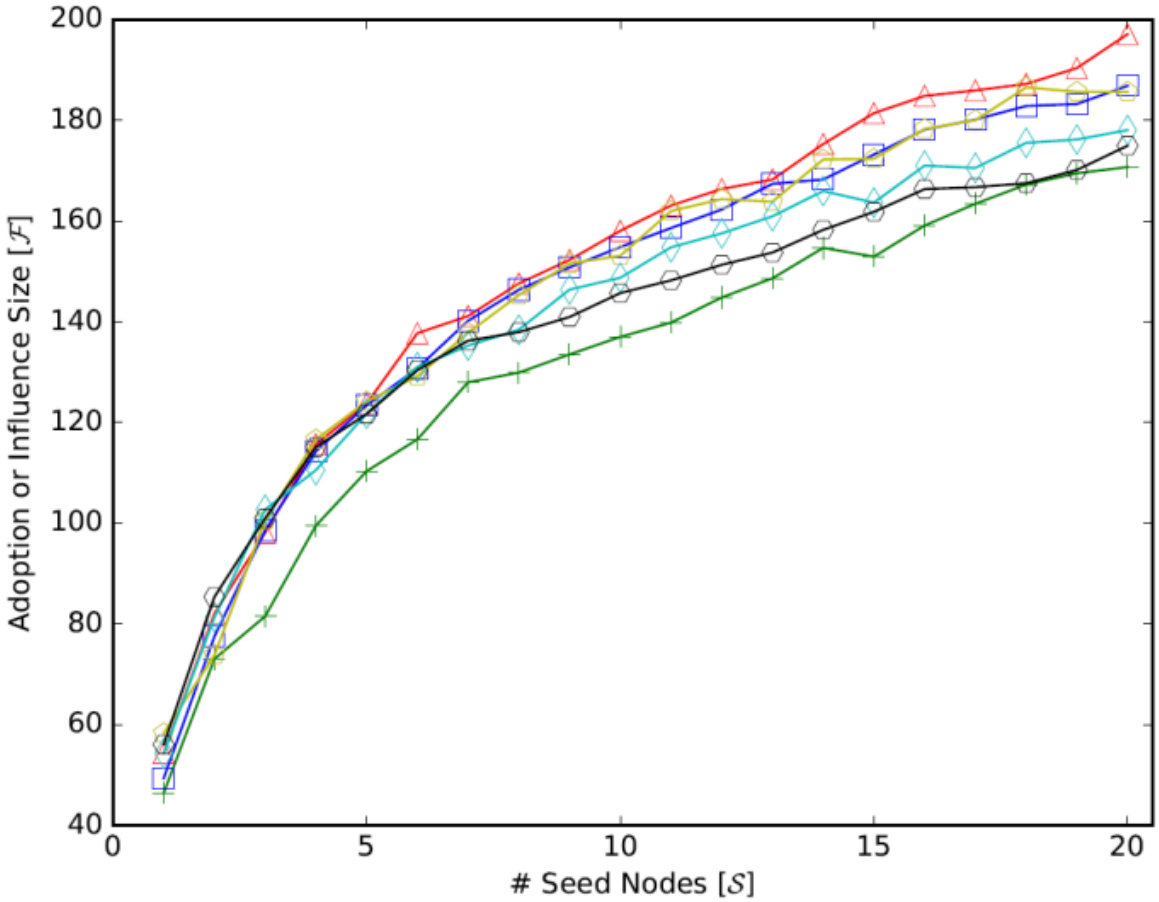

Figure 26: Information diffusion on small-world artificial network with six centralities and heuristics. The total number of nodes is $N = 400$; Propagation probabilities for opinion leaders and normal people are $p_{op} = 0.3, p_n = 0.2$

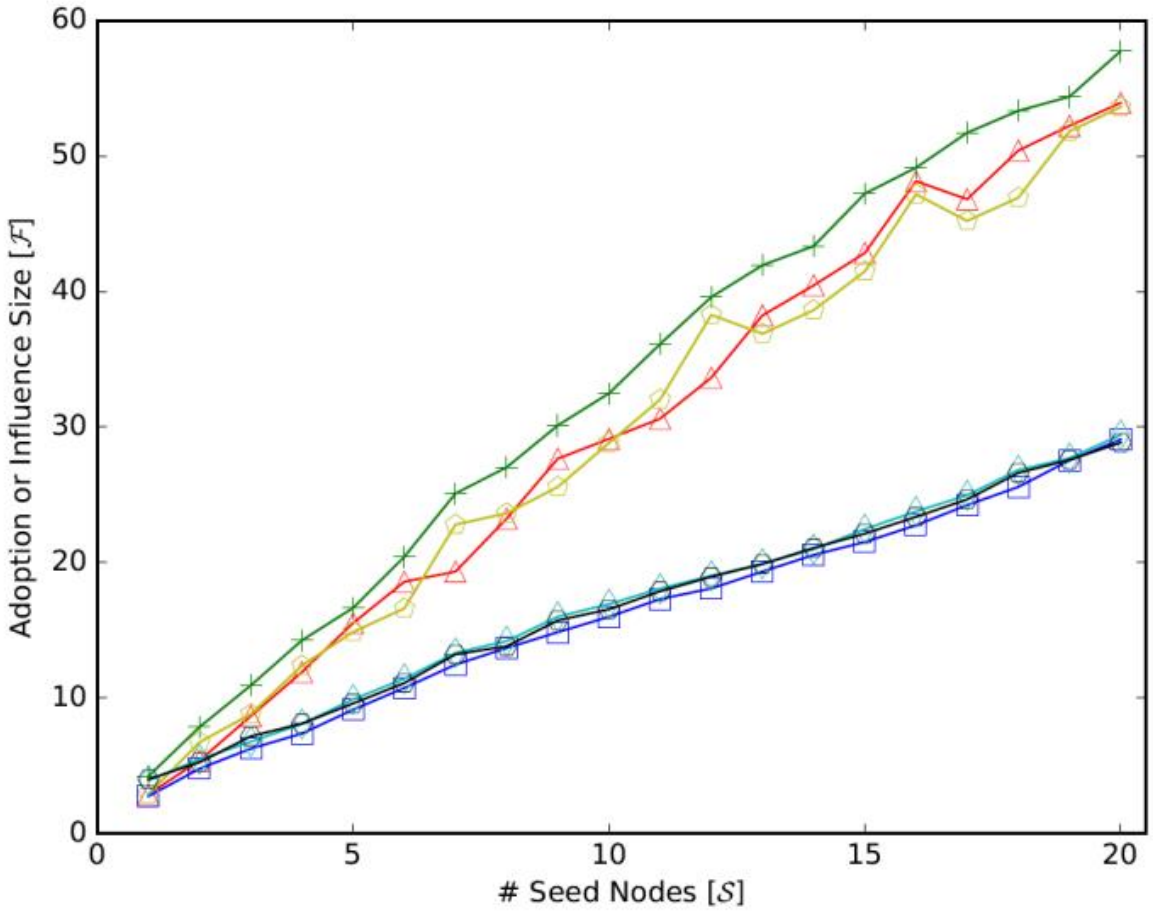

Figure 27: Information diffusion on lattice artificial network with six centralities and heuristics. The total number of nodes is $N = 400$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.3, p_n = 0.2$

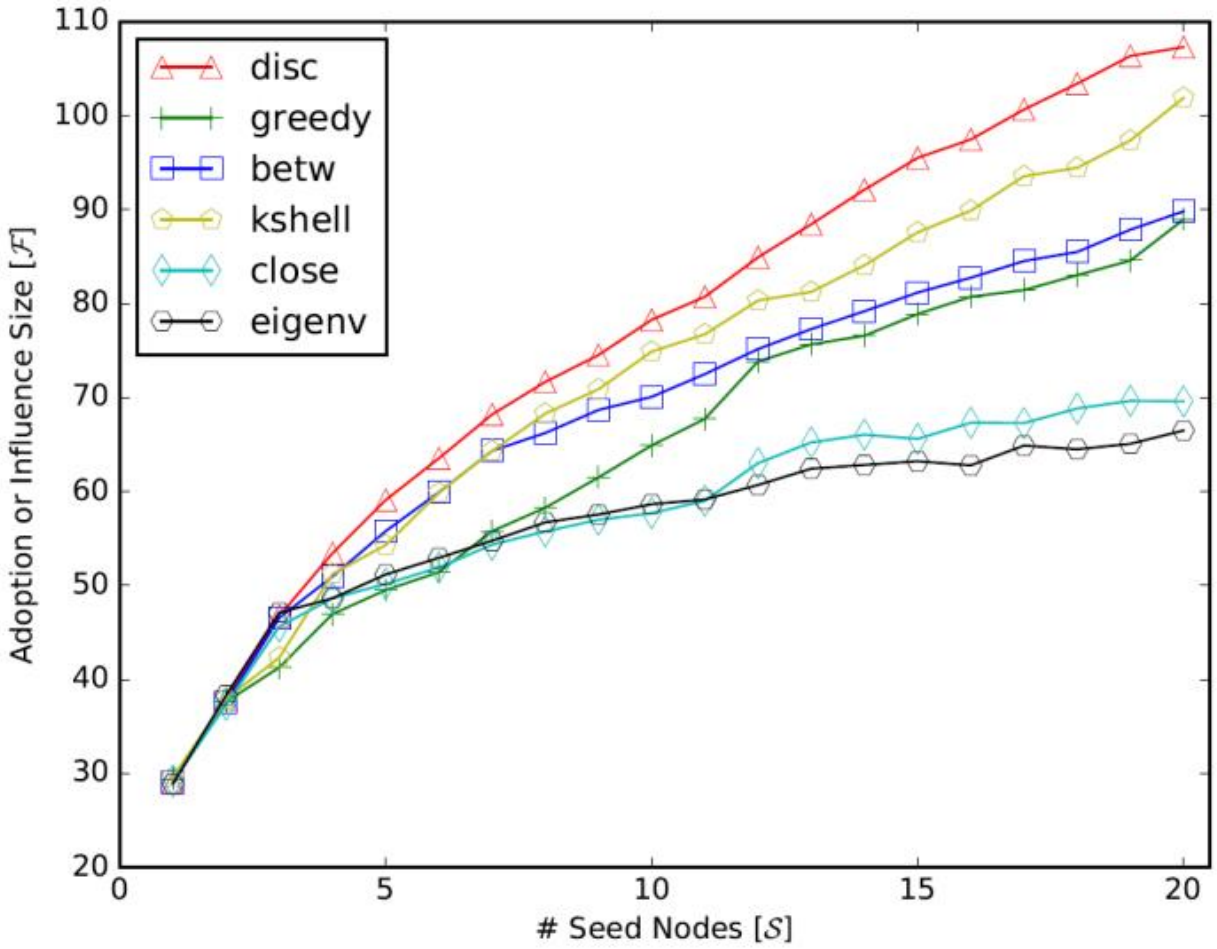

Figure 28: Information diffusion on preferential attachment artificial networkswith six centralities and heuristics (Disc: degree discount; Greedy: greedy algorithm; Betw: betweeness centrality; kshell: K-shell; close: closeness centrality; eigenv: eigenvector centrality). The total number of nodes is $N = 400$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.4, p_n = 0.3$

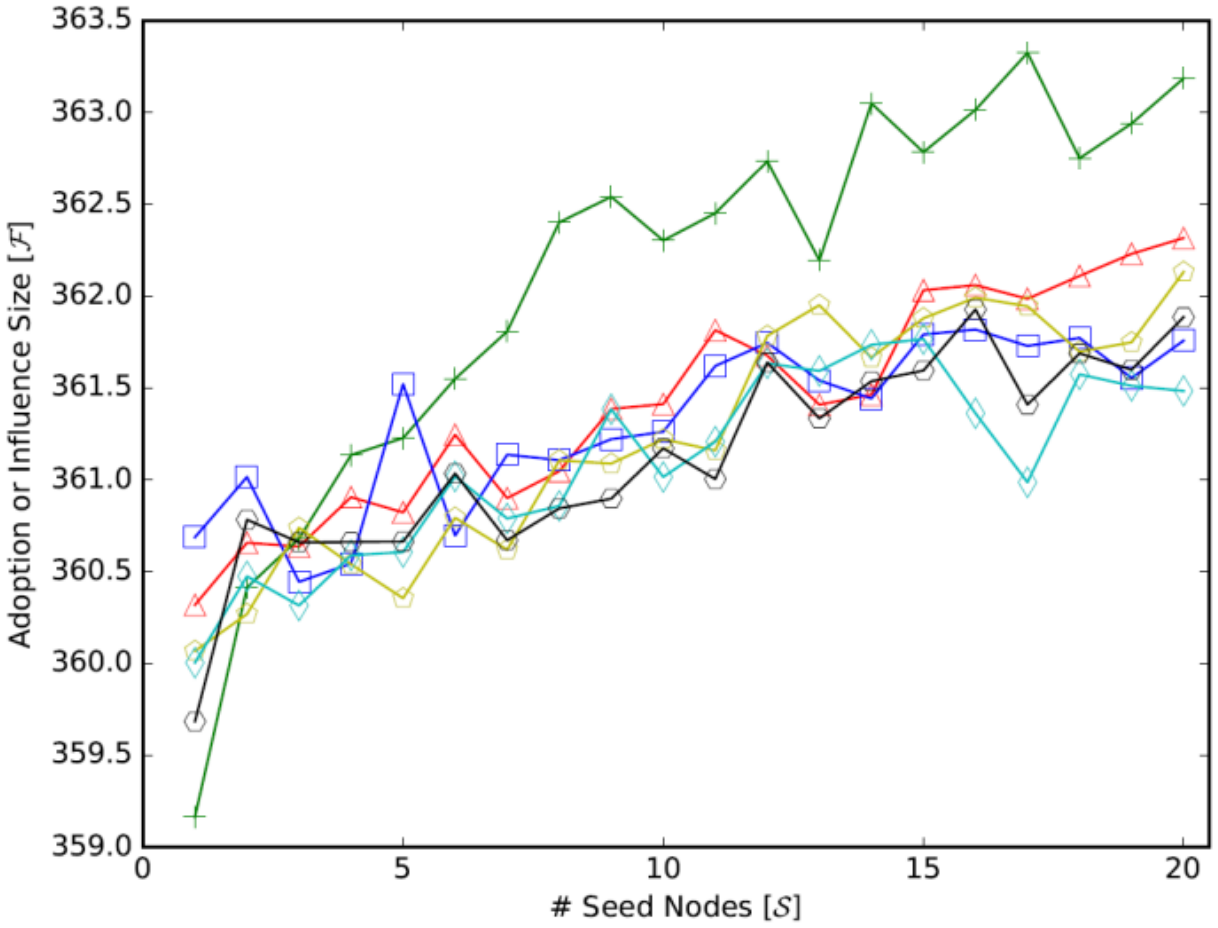

Figure 29: Information diffusion on random artificial network with six centralities and heuristics. The total number of nodes is $N = 400$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.4, p_n = 0.3$

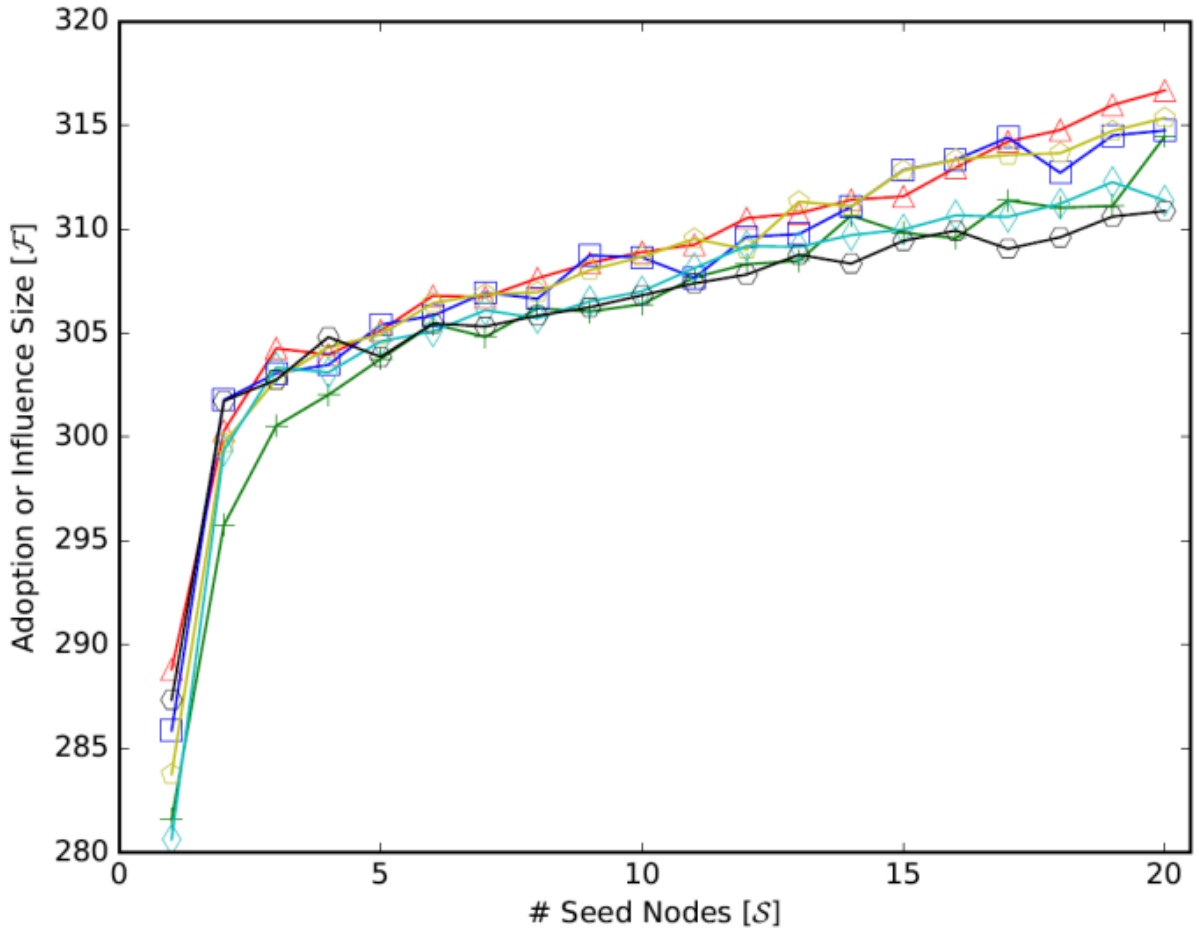

Figure 30: Information diffusion on small-world artificial network with six centralities and heuristics. The total number of nodes is $N = 400$; Propagation probabilities for opinion leaders and normal people are $p_{op} = 0.4, p_n = 0.3$

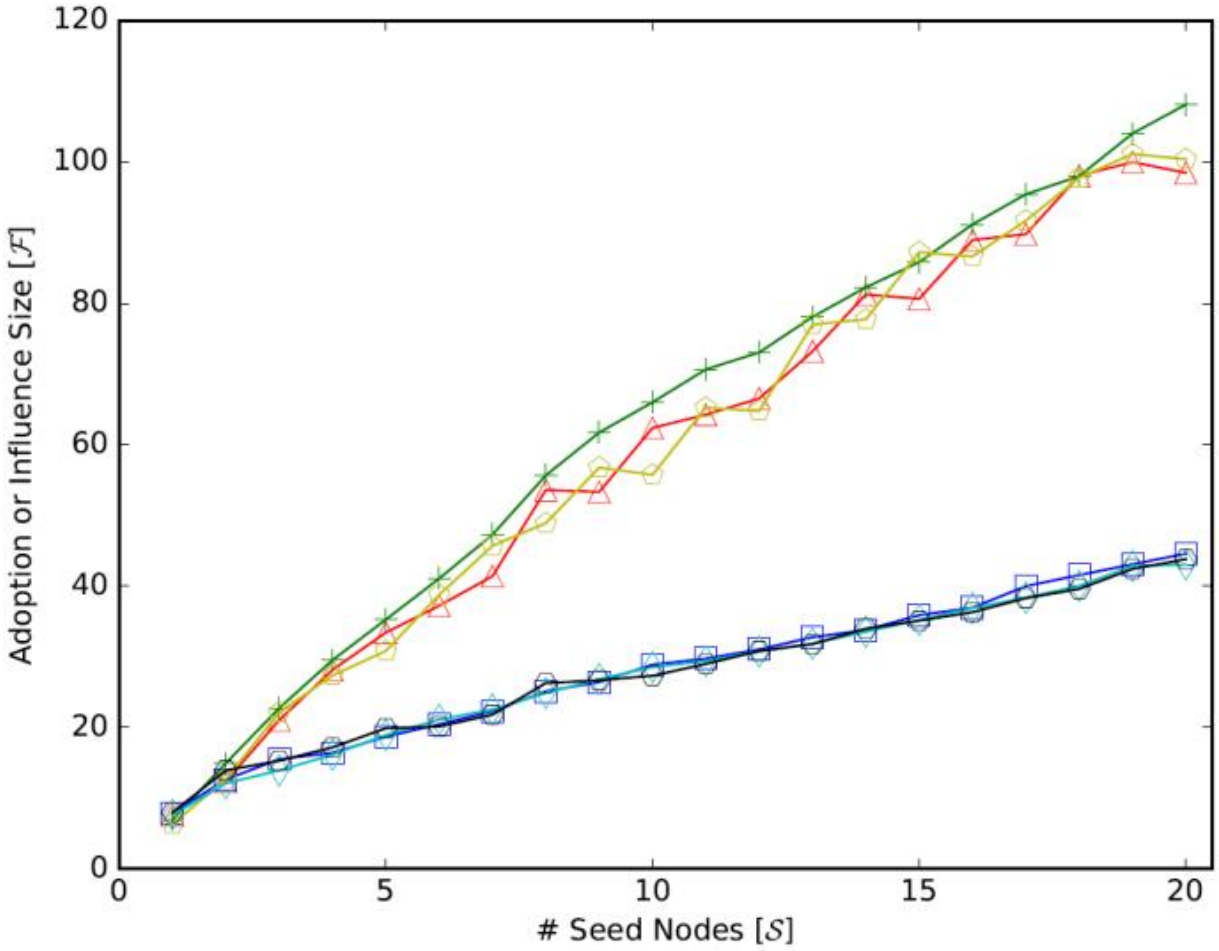

Figure 31: Information diffusion on lattice artificial network with six centralities and heuristics. The total number of nodes is $N = 400$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.4, p_n = 0.3$

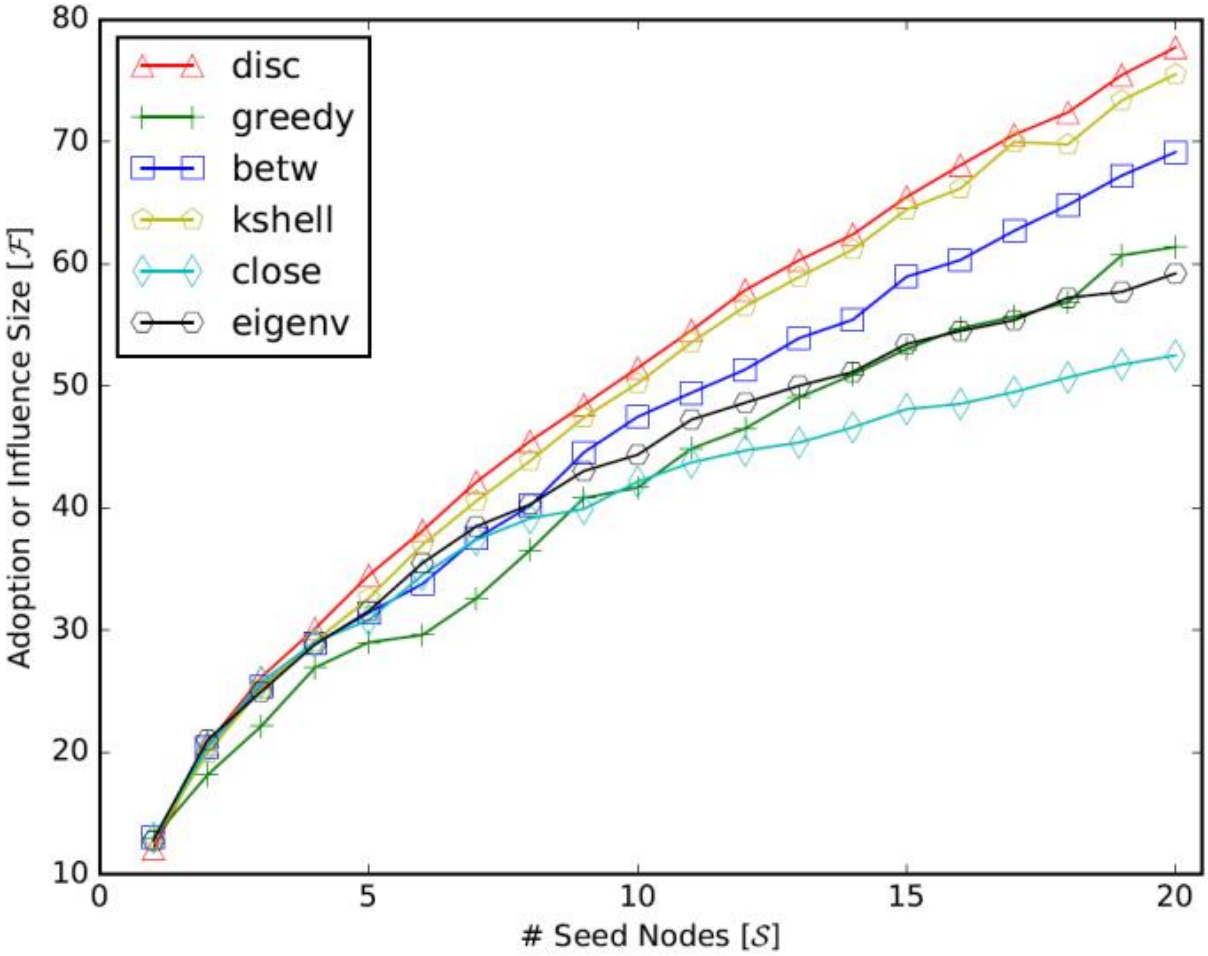

Figure 32: Information diffusion on preferential attachment artificial network with six centralities and heuristics (Disc: degree discount; Greedy: greedy algorithm; Betw: betweeness centrality; kshell: K-shell; close: closeness centrality; eigenv: eigenvector centrality). The total number of nodes is $N = 800$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.2, p_n = 0.1$

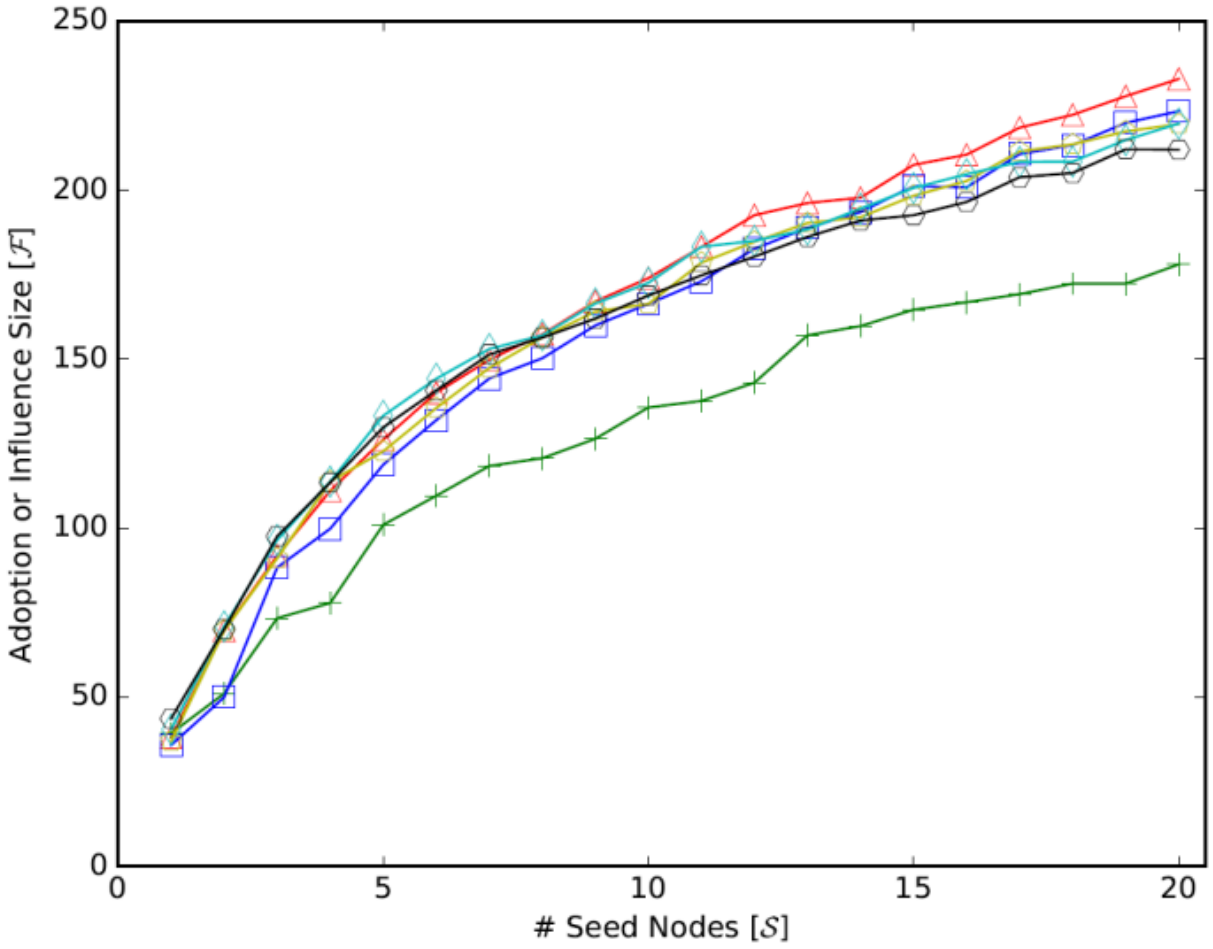

Figure 33: Information diffusion on random artificial network with six centralities and heuristics. The total number of nodes is $N = 800$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.2, p_n = 0.1$

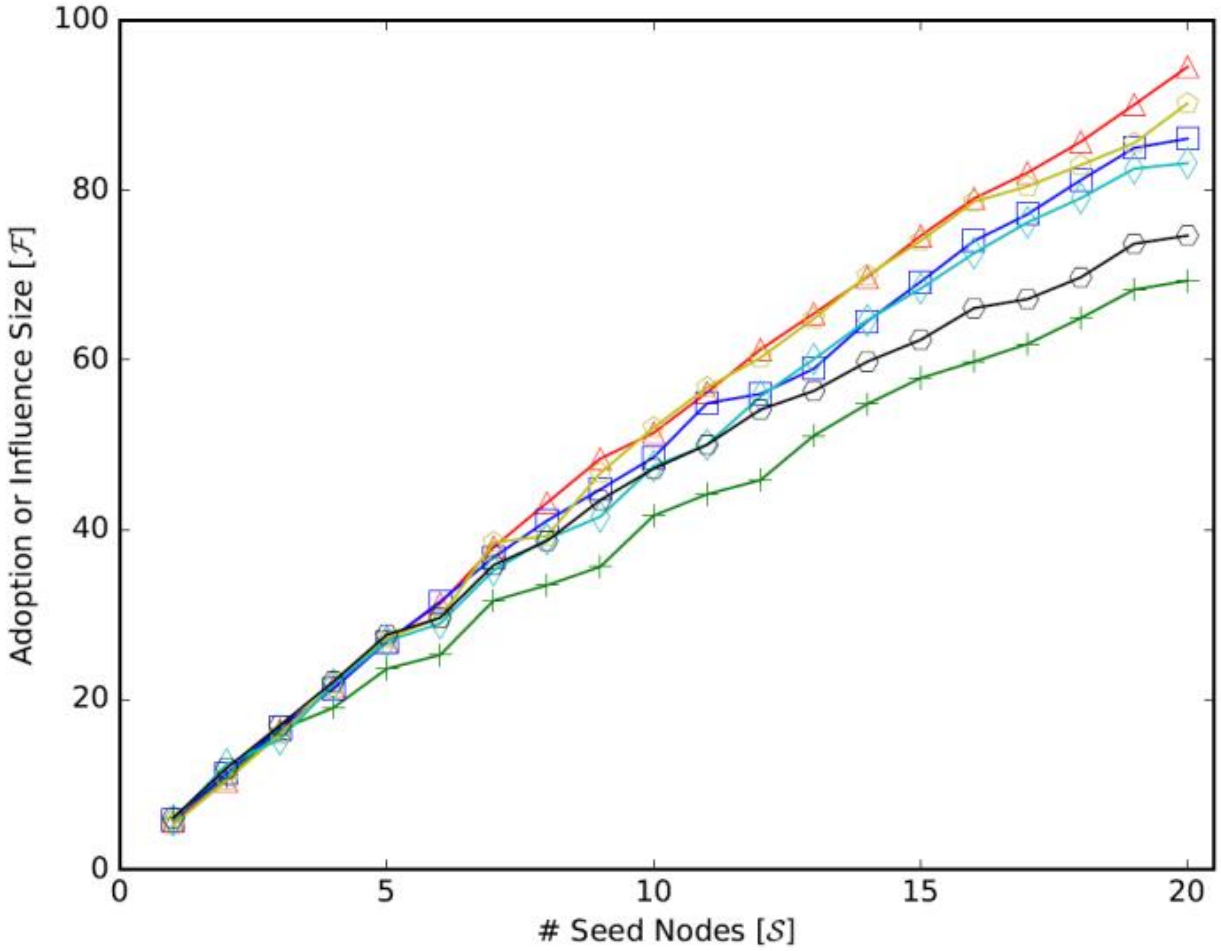

Figure 34: Information diffusion on small-wolrd artificial network with six centralities and heuristics. The total number of nodes is $N = 800$; Propagation probabilities for opinion leaders and normal people are $p_{op} = 0.2, p_n = 0.1$

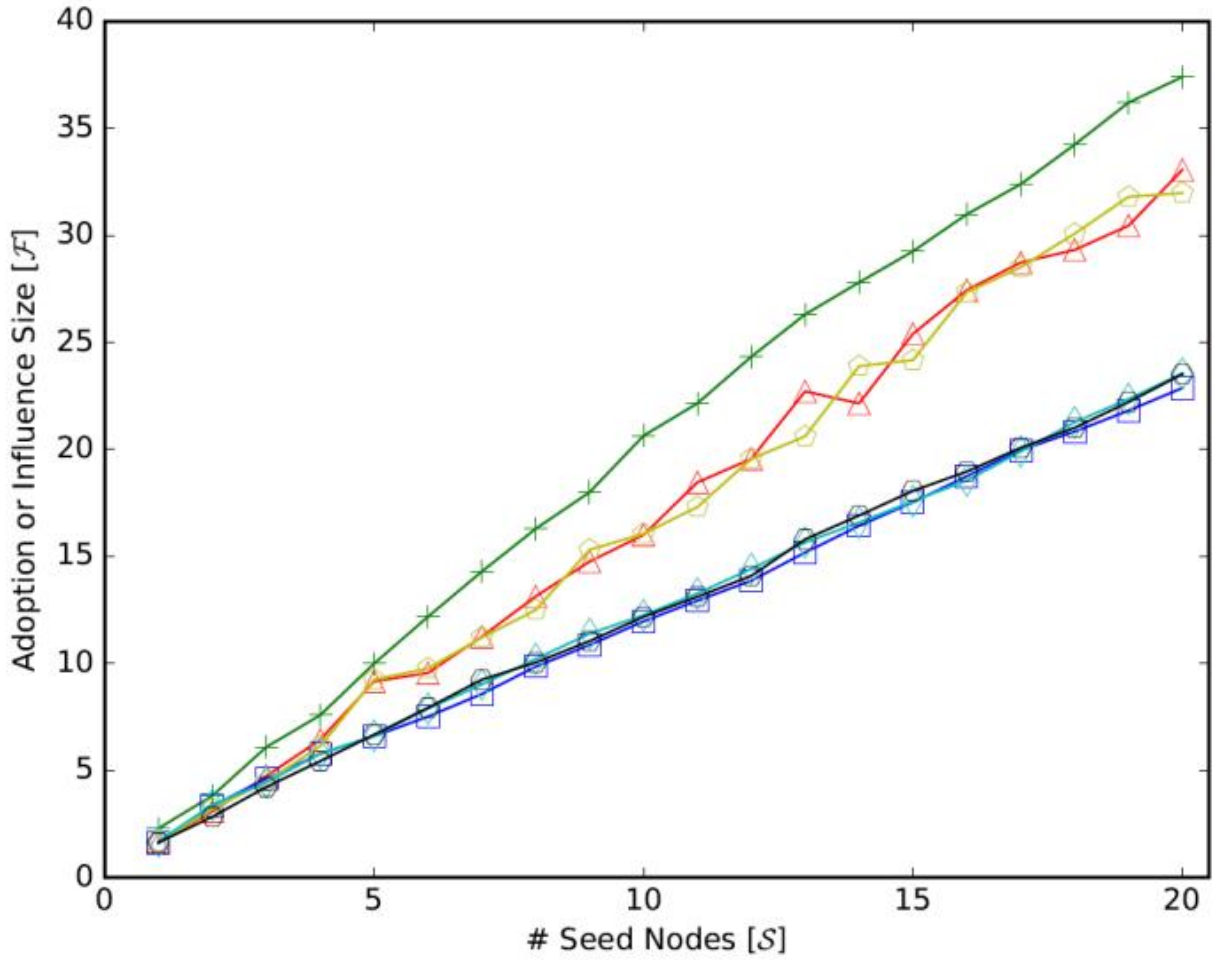

Figure 35: Information diffusion on lattice artificial network with six centralities and heuristics. The total number of nodes is $N = 800$; Propagation probabilities for opinion leaders and normal people are $p_{op} = 0.2, p_n = 0.1$

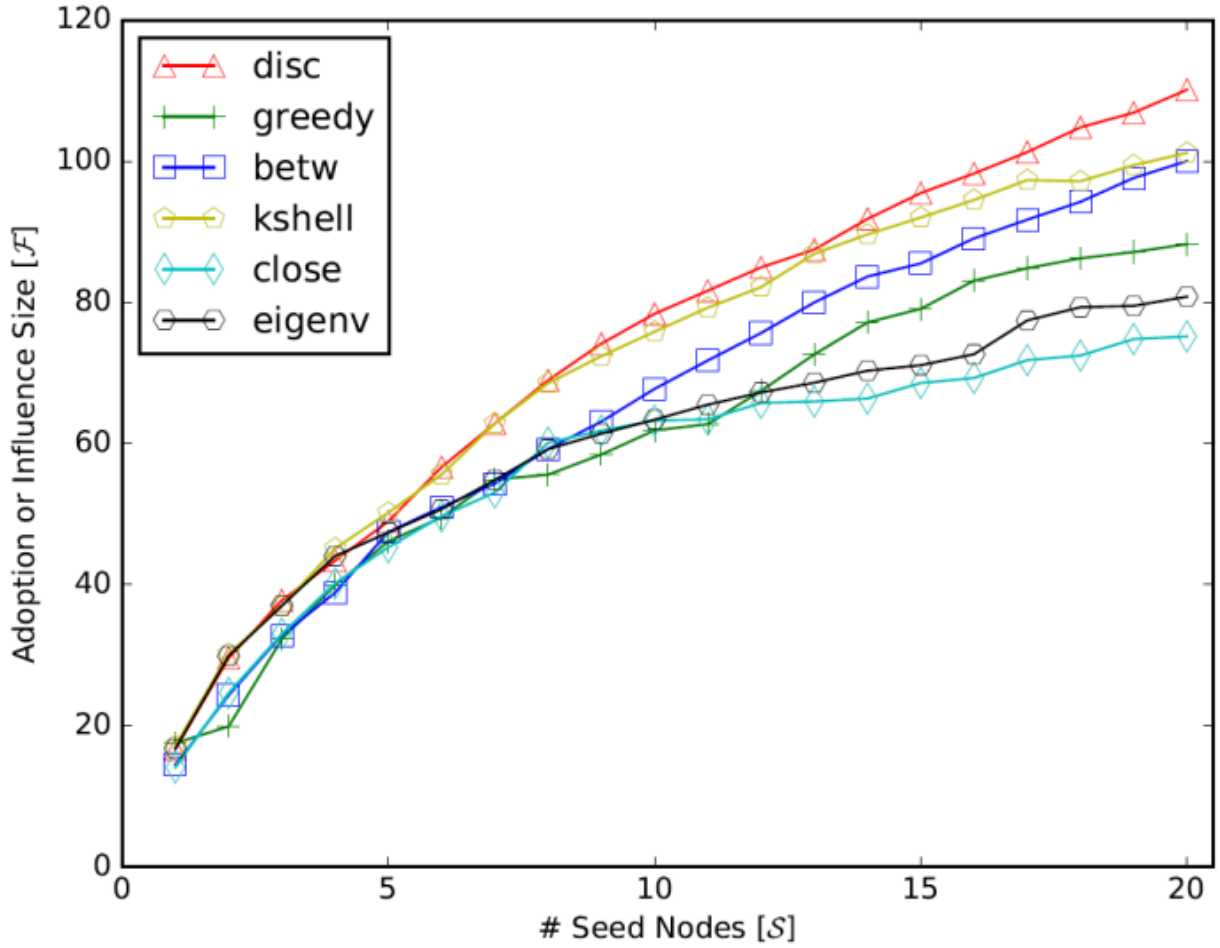

Figure 36: Information diffusion on preferential attachemnt artificial networks with six centralities and heuristics (Disc: degree discount; Greedy: greedy algorithm; Betw: betweeness centrality; kshell: K-shell; close: closeness centrality; eigenv: eigenvector centrality). The total number of nodes is $N = 800$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.3, p_n = 0.2$

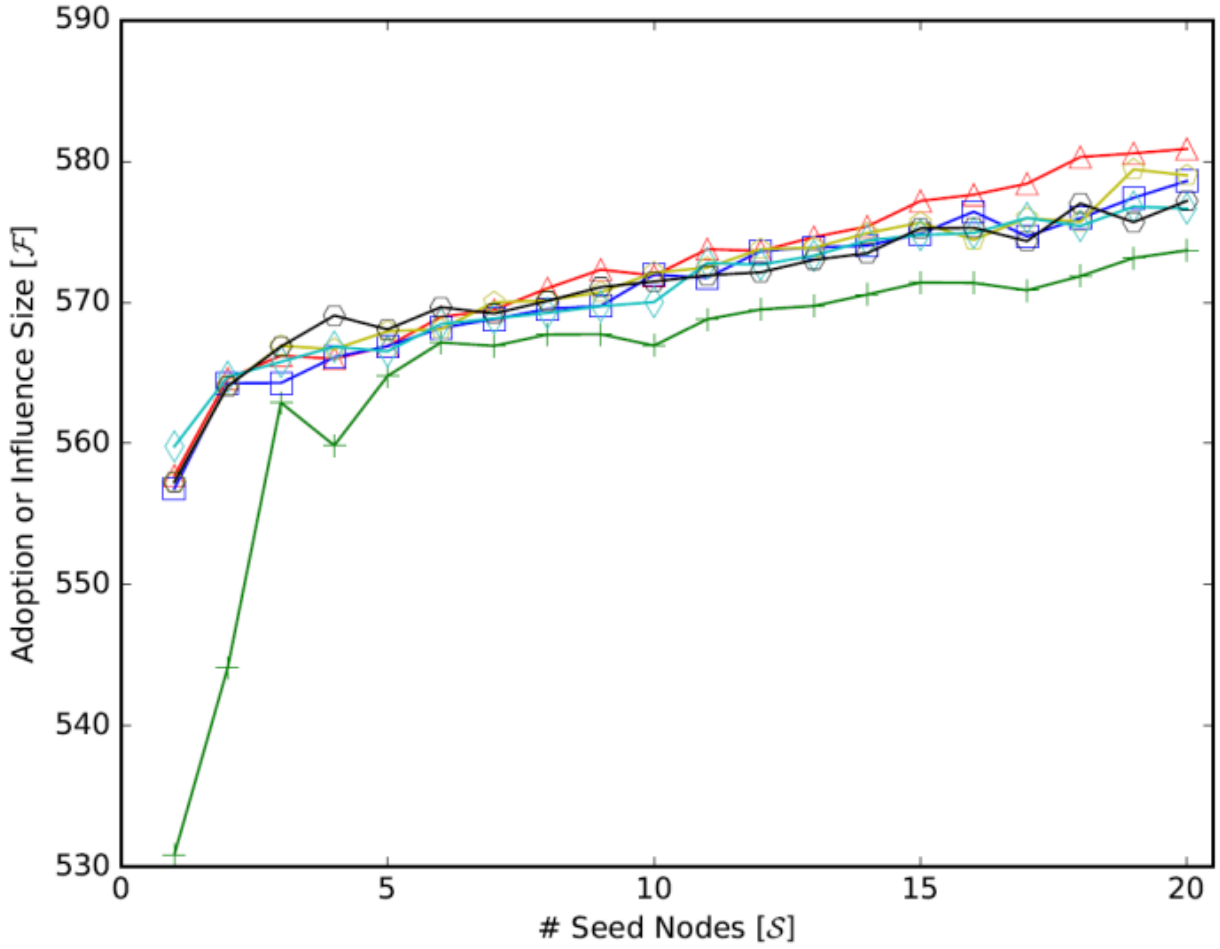

Figure 37: Information diffusion on random artificial networks with six centralities and heuristics. The total number of nodes is $N = 800$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.3, p_n = 0.2$

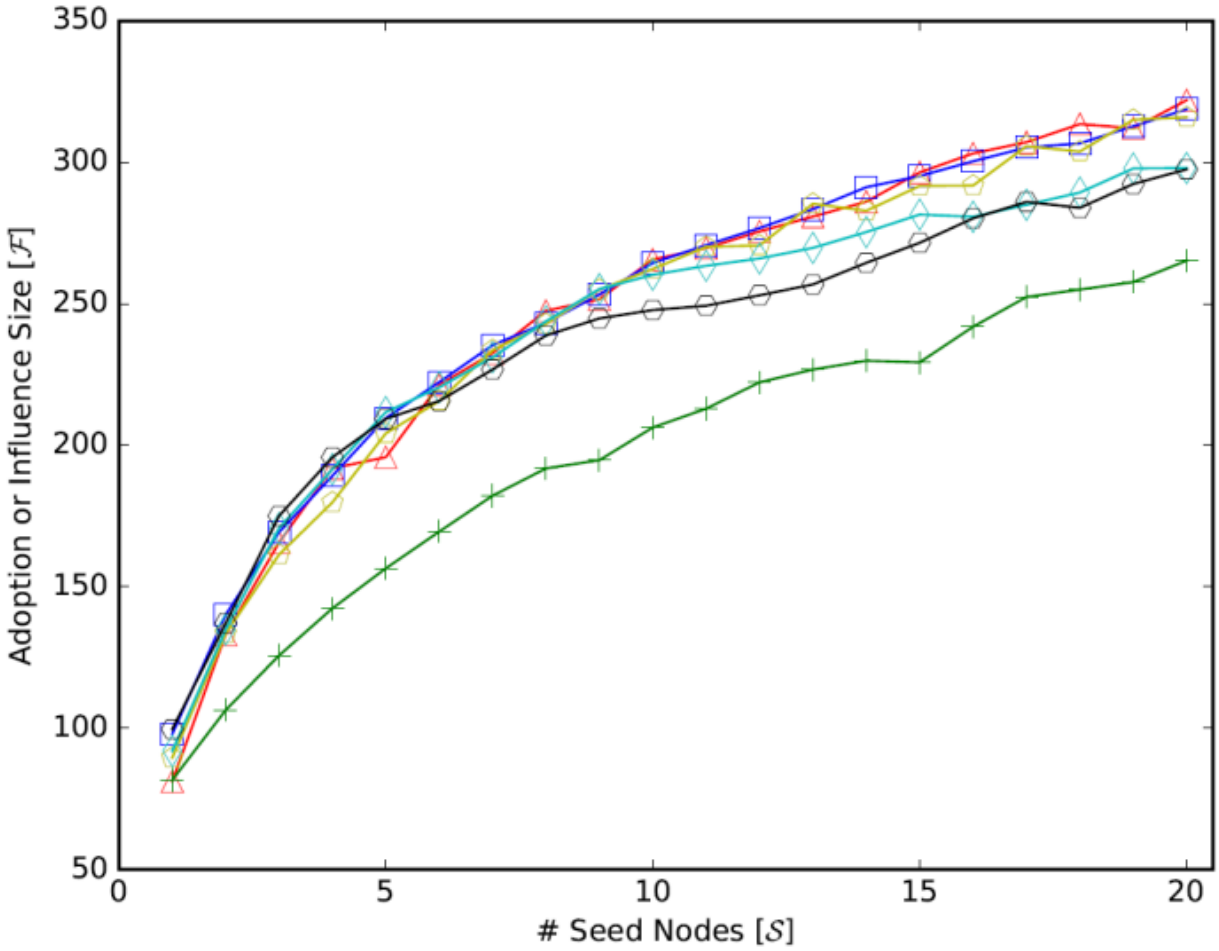

Figure 38: Information diffusion on small-world artificial networks with six centralities and heuristics. The total number of nodes is $N = 800$; Propagation probabilities for opinion leaders and normal people are $p_{op} = 0.3, p_n = 0.2$

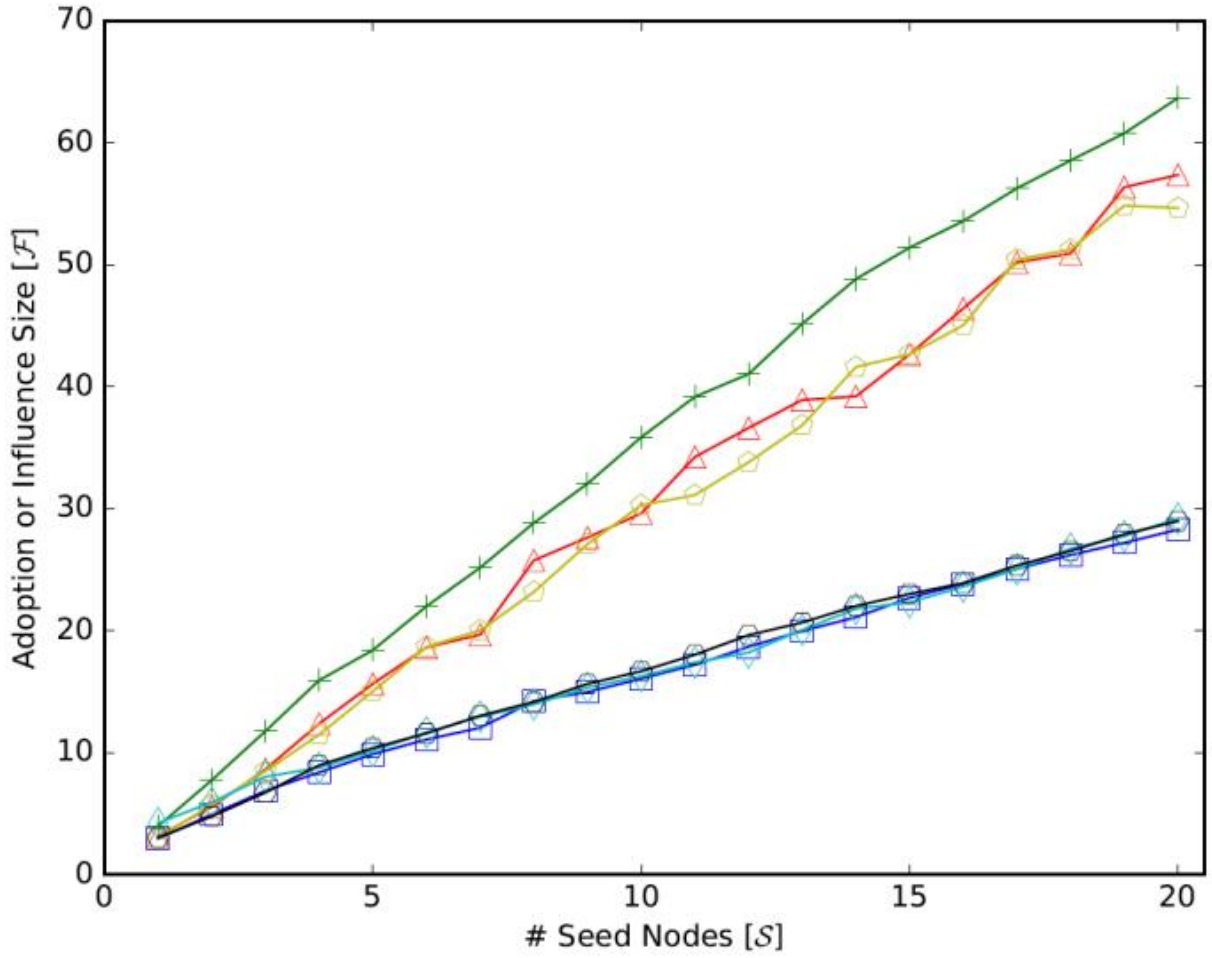

Figure 39: Information diffusion on lattice artificial networks with six centralities and heuristics. The total number of nodes is $N = 800$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.3, p_n = 0.2$

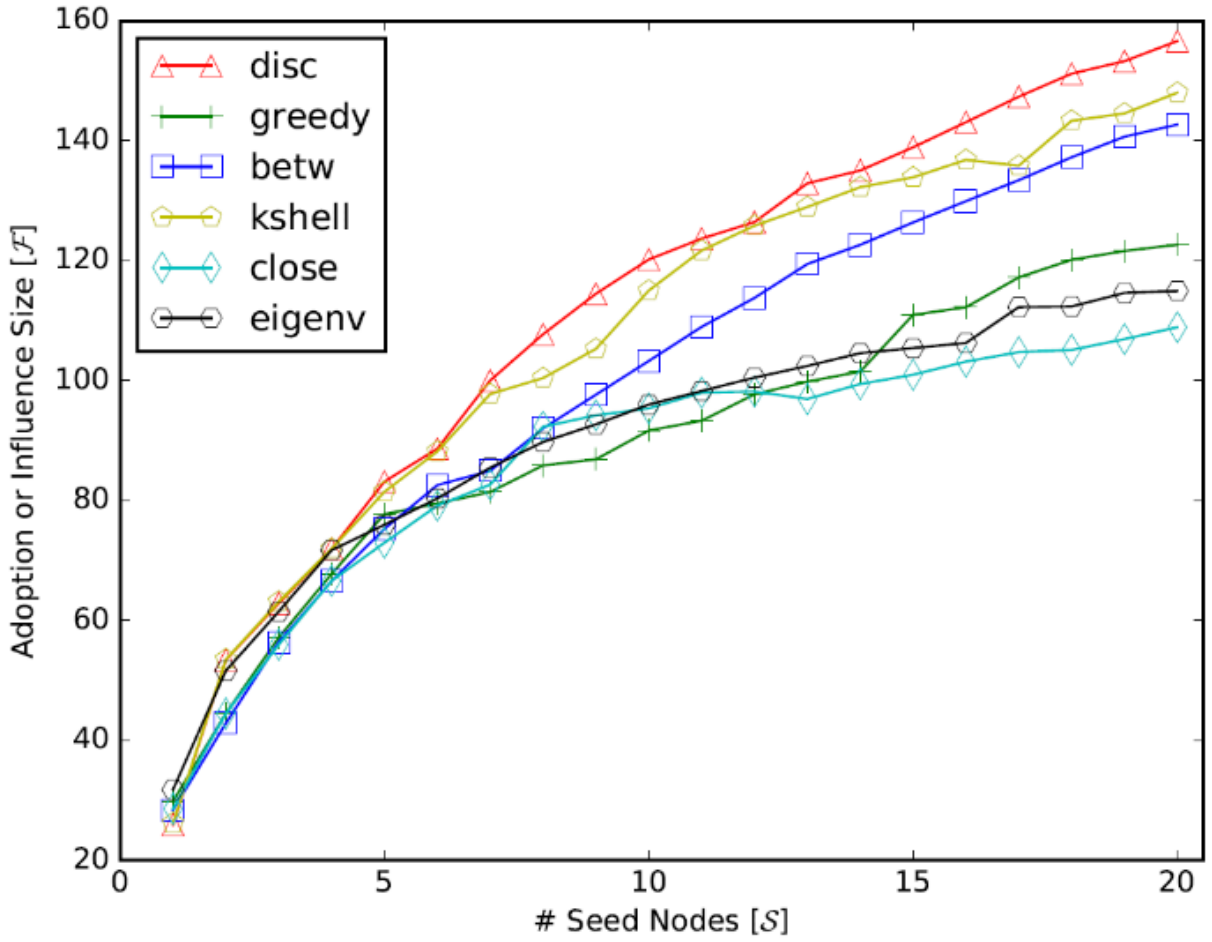

Figure 40: Information diffusion on preferential attachement artificial networks with six centralities and heuristics (Disc: degree discount; Greedy: greedy algorithm; Betw: betweeness centrality; kshell: K-shell; close: closeness centrality; eigenv: eigenvector centrality). The total number of nodes is $N = 800$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.4, p_n = 0.3$

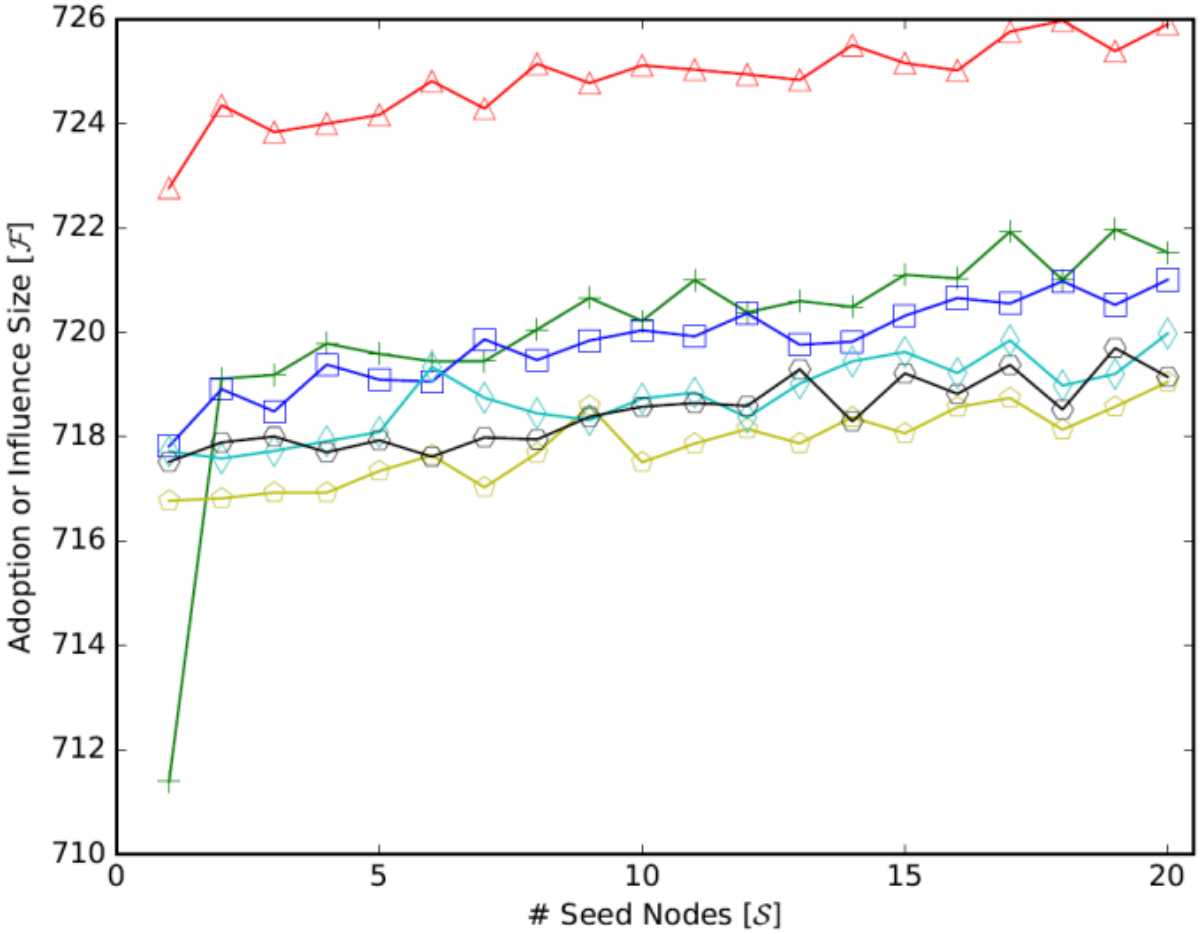

Figure 41: Information diffusion on random artificial networks with six centralities and heuristics. The total number of nodes is $N = 800$*; Propagation probabilities for opinion leaders and normal people are* $p_{op} = 0.4, p_n = 0.3$

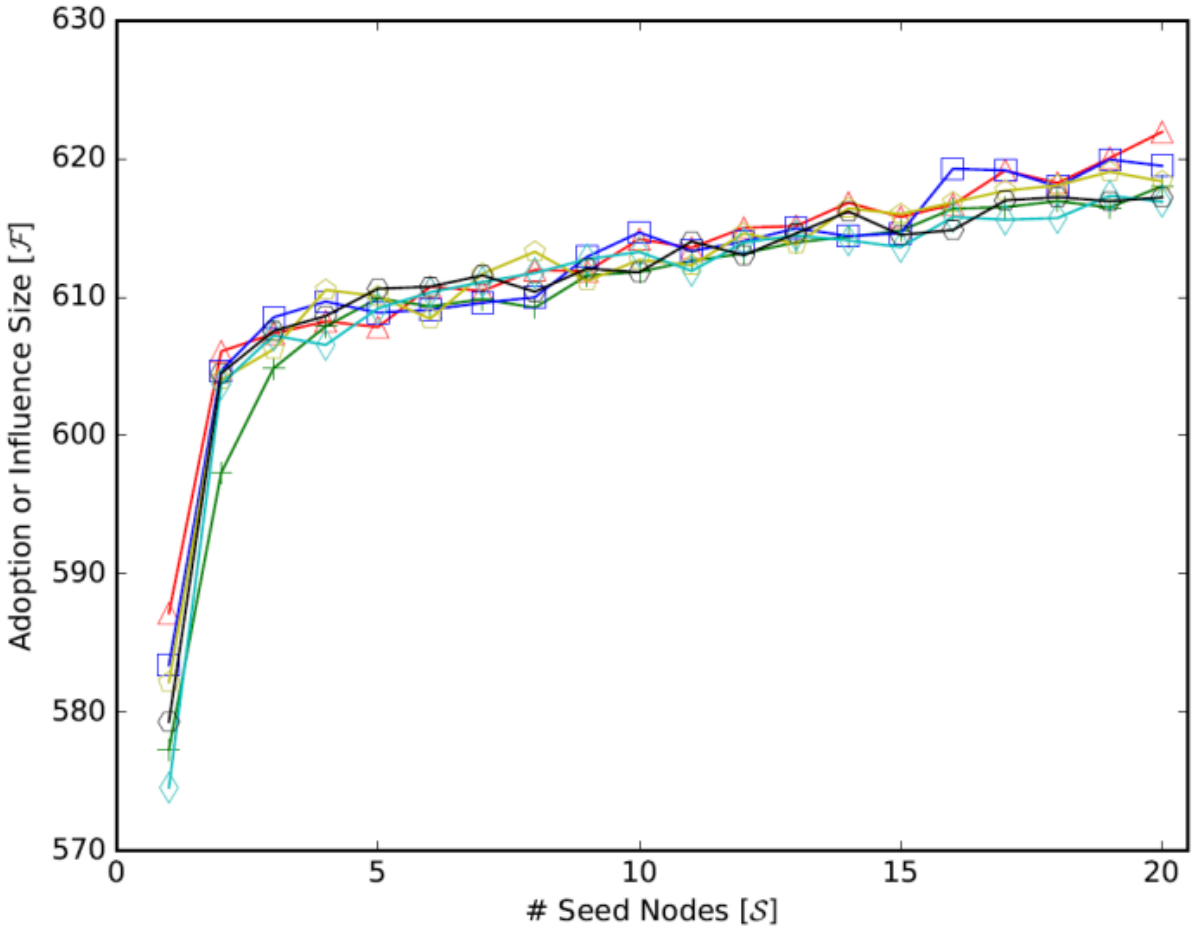

Figure 42: Information diffusion on small-world artificial network with six centralities and heuristics. The total number of nodes is $N = 800$; Propagation probabilities for opinion leaders and normal people are $p_{op} = 0.4, p_n = 0.3$

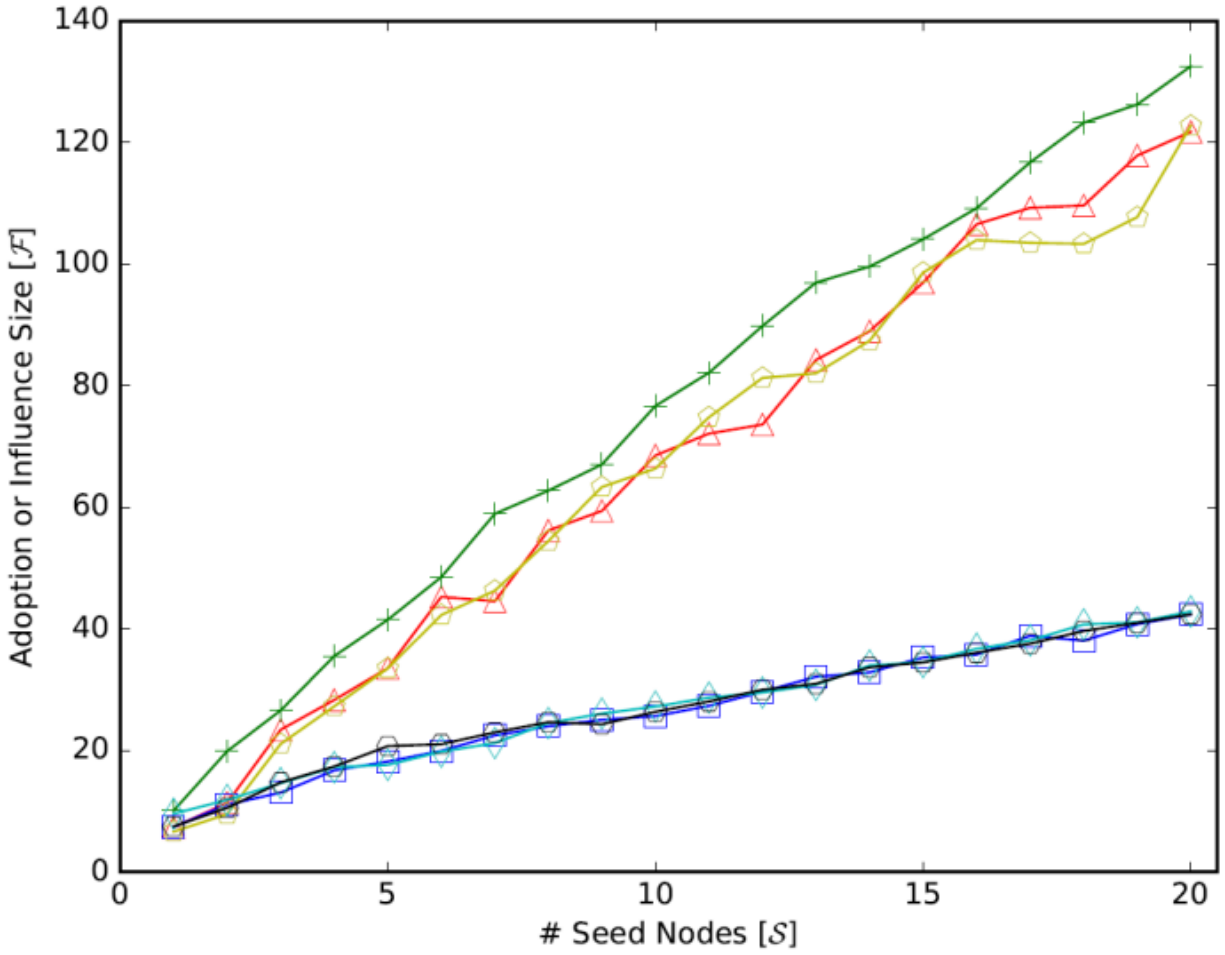

Figure 43: Information diffusion on lattice artificial network with six centralities and heuristics. The total number of nodes is $N = 800$; Propagation probabilities for opinion leaders and normal people are $p_{op} = 0.4, p_n = 0.3$

From these figures (Figure 8 to Figure 43), eigenvector, closeness, and betweeness centralities performed the worst in lattice networks. Degree discount had a good performance in all of these networks and sometimes it performed the best. Greedy algorithms performed not as well as expected, except in lattice networks. Especailly, the first early adopters selected by GA performs the worst in many types of networks. For example, in Figure 18, the first early adopters only influenced about 118 nodes, but the early adopters choosen by other methods could influence about 130 nodes. This may be due to insufficient number of simulations during the selection of early adopters. GA selects early adopters by simulating all possible diffusion results multiple times and chooses the node with the best performance.

The experiments in this thesis simulated the process of diffusion 100 times on each node for the selection of early adopters using GA. While in the experiment in the original paper on GA, the author simulated 20,000 times on each node in order to locate the best candidate of early adopters (Kempe et al., 2003). This is, of course, very time-consuming and not applicable for agent-based models in NetLogo, which is part of the shortcomings in this thesis research. Nevertheless, the heavy workload for greedy algorithm suggested the need for finding another way of locating early adopter. Rather than simulating and tranversing each of the nodes in the network, which becomes a big problem in huge networks, we may need to come up with new methods that efficiently locate early adopters with a good level of precision. For instance, degree discount heuristics may be a good approach.

It can be also noted that a well-connected network, such as the random network in Figure 41 (N=800, E=3,145), would see its propagation probabilities increase to a higher value (e.g. $p_{op} = 0.4, p_n = 0.3$) with the adoption size staying nearly at a constant level even when there were more early adopters added. The influenced population reached up to 90% in random

networks and 78% in small-world networks with high propogatbion probability ($p_{op} = 0.4, p_n = 0.3$). This indicates that there exists a saturation effect or diminishing marginal returns where an increasnig number of early adopters can no longer ensure a larger adoption size. Beacause in such well-conneted network and with high proprogation probabilties, the value interval in y axis is relatively short: the difference between each interval is only 2 in Figure 41. It results in the irregular or jumpy lines in the graph. Similar resul happened to the experiment in Figure 29. with 400 nodes in a random networks. Each early adopter added in to the networks triggered an information diffsuion with minor difference, which causes to small interval in y axis with only 0.5 difference.

Another interesting phenomenon in the experiments was that in lattice networks where node's degrees were evenly distributed, identifying the optimal early adopters seemed to be unapplicable in many situations (e.g. Figure 11, Figure 14, …, Figure 35, Figure 39 or Figure 43). It means that any one node in a lattice network has nearly the same importance of informaiton diffusion so that any node added in to the set of seed nodes tends to trigger similar diffsuion results. In such a network, opinion leaders and early adopters did not have their special function as they had in other networks.

Therefore, in the subsequent analysis of efficient diffusion with early adopters, lattice networks were excluded. To solve the problem of efficient diffusion in the networks, experiments would require finding a set of early adopters that warranted the most productive results of diffusion. Thus, an evaluation measure was developed in terms of finding how many further adoptions after each early adopter was added. The number of further adoptions was calculated by the adoptions after an early adopter was added and minus the adoptions before the early adopter

was added. Since an early adopter was the one adopted manually at the first step of diffusion, it should be excluded from the result of further adoptions.

Degree discount heuristic was chosen in the evaluation as the method of locating early adopters in a network because it had the best performance during the simulation. Table 3 lists further adoptions that every rank of early adopter triggered, from the quantity of one to twenty. The size of further adoption increased dramatically as the propagation probability increased. Especially in the random network and small-world network, the first early adopter triggered adoption more than 50% of the population already, with the propagation probability $p_{op} = 0.4, p_n = 0.3$. As early adopters increased, the size of adoption decreased greatly before it reached 0. Generally speaking, the first early adopter is the most productive seed node that influences the most of the population in the diffusion process. Sometimes, subsequent early adopters often reach a good portion of the population, but more commonly, these subsequent early adopters barely have an effect in a well-connected network.

The experiment shows that the effectiveness of diffusion largely depends on the structures of networks and the propagation probability. With ABM, we have the ability to simulate information diffusion in different network structures, and this ability enables us to find the efficient diffusion with appropriate propagation probabilities, especially in real networks.

Table 3: Evaluation of the efficiency of diffusion on each early adopter in three different networks. Further adoption is grouped by all sets of propagation probabilities: a) $p_{op} = 0.4, p_n = 0.3$, b) $p_{op} = 0.3, p_n = 0.2$ and c) $p_{op} = 0.2, p_n = 0.1$. S denotes the rank of seed nodes (early adopters); PA refers to preferential attachment. The network size $N = 200$

	p_{op}=0.2			p_{op}=0.3			p_{op}=0.4		
S	PA	small-world	random	PA	small-world	random	PA	small-world	random
1	3.204	3.7	5.316	7.427	27.155	37.945	11.554	133.455	121.207
2	3.384	3.847	3.471	4.054	13.608	9.761	10.045	8.394	9.432
3	2.351	2.77	6.314	4.895	7.552	12.217	3.455	0.581	-0.878
4	1.597	3.122	3.014	3.634	10.439	4.619	5.247	1.326	0.447
5	2.441	3.396	2.113	2.065	6.141	2.394	3.201	-0.199	0.46
6	1.576	1.935	3.532	2.708	6.36	3.202	3.021	0.476	0.352
7	1.519	2.868	2.032	2.059	2.15	3.651	2.421	0.384	-1.505
8	0.907	2.965	2.614	1.959	3.263	1.453	0.049	-0.531	0.295
9	1.524	2.355	2.386	-0.193	2.503	2.367	5.168	-0.57	0.526
10	0.235	-0.757	1.337	3.229	-0.601	0.987	1.706	-0.185	-1.54
11	0.546	2.776	1.742	0.791	1.717	1.335	1.628	-0.458	0.355
12	1.443	0.23	2.121	0.745	-0.517	0.992	0.698	-0.547	-0.551
13	0.113	2.315	1.602	0.877	4.669	1.035	0.433	-0.516	-0.355
14	1.183	1.298	0.675	0.38	0.463	-1.547	0.595	-0.042	-0.384
15	-0.165	-1.338	0.271	1.267	-1.206	2.416	0.917	-0.662	-0.904
16	0.797	2.68	0.034	-0.153	0.486	0.081	0.73	0.011	-0.414
17	-0.43	0.617	1.285	0.669	1.698	1.02	1.073	-1.095	-0.603
18	0.783	1.929	0.689	0.722	1.132	-0.669	-0.451	-0.203	-0.463
19	-0.191	1.717	0.001	-0.205	0.447	-0.843	0.915	-1.263	-0.903
20	0.389	-0.194	-0.085	0.361	2.158	1.591	-0.279	-0.166	-0.14

4.4 Real Network Examination

4.4.1 Finding Propagation Probability of Real Network

This thesis used the dataset of Bernardo wildfire tweets as the real networks to explore the information diffusion. The data set was obtained from SMART Dashboard (http://vision.sdsu.edu/hdma/smart/wildfire_ca) designed by HDMA in San Diego State University. The data set helps look for the propagation probabilities of message spreading on social networks for real-world event. The real-time tweets talking about Bernardo fire arose along with the news of wildfire. Therefore, the information diffusion was affected by the status of the event (Bernardo wildfire in this case). It is assumed that different periods of the event featured different propagation probabilities during the diffusion process. Accordingly, it would be more meaningful for investigation on the propagation probabilities during each period of the event.

Table 4: Result of information diffusion in Bernardo wildfire tweets under 5-day partition

Time-range	Seed accounts	Influence
Day 1	KUSI_News, RSF_Fire	622
Day 2	SanDiegoCP, thesandiegonewz, twit_san_diego, sandiegobnews, 10News, ooph, dancohenCBS8	447
Day 3	blufinki	142
Day 4	BlazonLaurels, EdZieralski,jennifercdougla	213
Day 5	thesandiegonewz, AthensMarketSD, KPBSnews	52

Table 4 presents details of information diffusion in Bernardo wildfire tweets in a 5-day partition. The tweets dataset was placed into bins of 1-day intervals using midnight as the dividing mark line. This is because the frequency of tweets always drops at midnight and then increased again in the morning. The Wildfire topic emerged on Day 1 (5/13/2014 10am) and started becoming viral and trigger most of the adoptions on Day 1 and Day 2. It started to lose attention from the public on Day 3 but regained some popularity later on Day 4. This was

because the Bernardo wildfire was reportedly 100% contained on Day 4, which brought back the attention. The seed accounts in the table are those early adopters who first spread the information originally (i.e. not a retweet) at the beginning of the day.

With these early adopters, Bernardo wildfire network was imported into NetLogo to run the information diffusion model (Figure 44) searching for the propagation probabilities that mimic diffusion on real networks. Because we had no previous knowledge about the propagation probabilities, a grid search was conducted to find optimal parameters by simulating the diffusion process and by increasing the propagation probabilities by a fixed increment (0.01 initially and 0.001 in detail) in each simulation. Each pair of parameters were modeled 10 times. The results were evaluated by comparing the simulated influence and the real influence. Figure 44 shows the GUI of grid search. Users need to choose the amount of increment (e.g. 0.1, 0.01 and 0.001) to run the grid search: *find-parameters*. All these functions are based on the assumption that users have imported the real networks into NetLogo model.

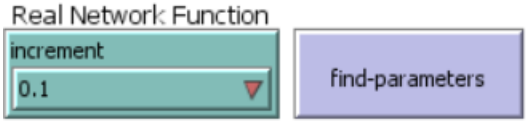

Figure 44: GUI of grid search for propagation probabilities in the model of NetLogo

Using contour maps, the areas of errors (difference between simulated results and real network results) was plotted from low to high. Lower values are light greens, and higher values are darker reds, so areas of light green indicate areas with minimal errors and, thus, closest fits. Depending on the day in the event, the contour map of the day presents quite different insights about the areas of errors and optimal parameters. It is important to notice that the white area in the plots (Figure 45. to Figure 49) is where that p_n should be smaller than p_{op}. This area indicates that opinion leaders should have a greater influence on their neighbors. Thus in the experiment of grid search, the search area where $p_n > p_{op}$ are not being considered.

It is assumed that for each day of the Bernardo wildfire, the propagation probabilities are not the same. In these figures, diffusion on Day 1 (Figure 45) witnessed a larger cascade of information diffusion than other days. This implies larger propagation probabilities on that day. Propagation probabilities then should decay as the time goes on if the topic has no update content -- as shown in Table 4. Some of the simulations yielded a series of optimal parameters while others yielded only a few or even none. This was because of the randomness of the model, which was the same reason why contour lines were so irregular in each figure. Nevertheless, these plots still indicate that there was a range of parameter pairs that produced similar results. Sometime there were strong opinion leaders in the diffusion process and had a larger p_{op}. However, some opinion leaders would only have a slightly more influence (i.e. little larger p_{op}) than the influence by normal people. This explains why we have a range of values that match the real diffusion result fairly well. In addition, the difference of the slope of low error areas (light green) in the contour maps suggests that in some scenarios, normal people are the major factors that propagation the information while in other scenarios, opinion leaders play critical roles. For instance, the slope of light green area on Day 4 (Figure 48) is relatively flat, indicating that opinion leaders are not as important as normal people are. More specifically, the influence of increased or decreased p_{op} is less than the influence of increased or decreased p_n in terms of fitting areas. However, the scenario on Day 5 (Figure 49) tells a very different story. The slope of light green area is very steep, so p_{op} dominates how well the diffusion model fits no matter how p_n changes. If the we can gauge the difference between p_{op} and p_n, the range of optimal parameters could narrow down. Yet unfortunately, the difference between p_{op} and p_n seems to vary with the events and there is no way to forecast it to date.

With the help of ABM simulation, we are able to identify the range of optimal parameters (propagation probabilities) which can be used to mimic the information diffusion in real events. Though sometimes the range could be wide (Figure 45), the extent of the range usually does not exceed 0.1, which still gives us a reference of how information spread. In fact, the precision of optimal parameters relates to the location of early adopters that are identified in the dataset. As we examine Table 3, the seed nodes or early adopters in each day actually are located in very different places. Some of them that are very central in the network seem to be very influential no matter how low we set the diffusion probability to be. Thus the simulation ran smoothly and it produced a narrow range of optimal parameters (Figure 45 for example). In the case of Figure 47., the early adopters on Day 3 were located in peripheries of the network, which barely helped the spread of the information to the public. Meanwhile, because the opinion leaders were usually in the central part of the network, it was difficult for peripheral nodes to reach them. Thus chances are great that opinion leaders do not participate in the diffusion process, which leaves the p_{op} to be inaction. Such unstable diffusion caused the irregular contour lines in the figures and also the wide range of optimal parameters.

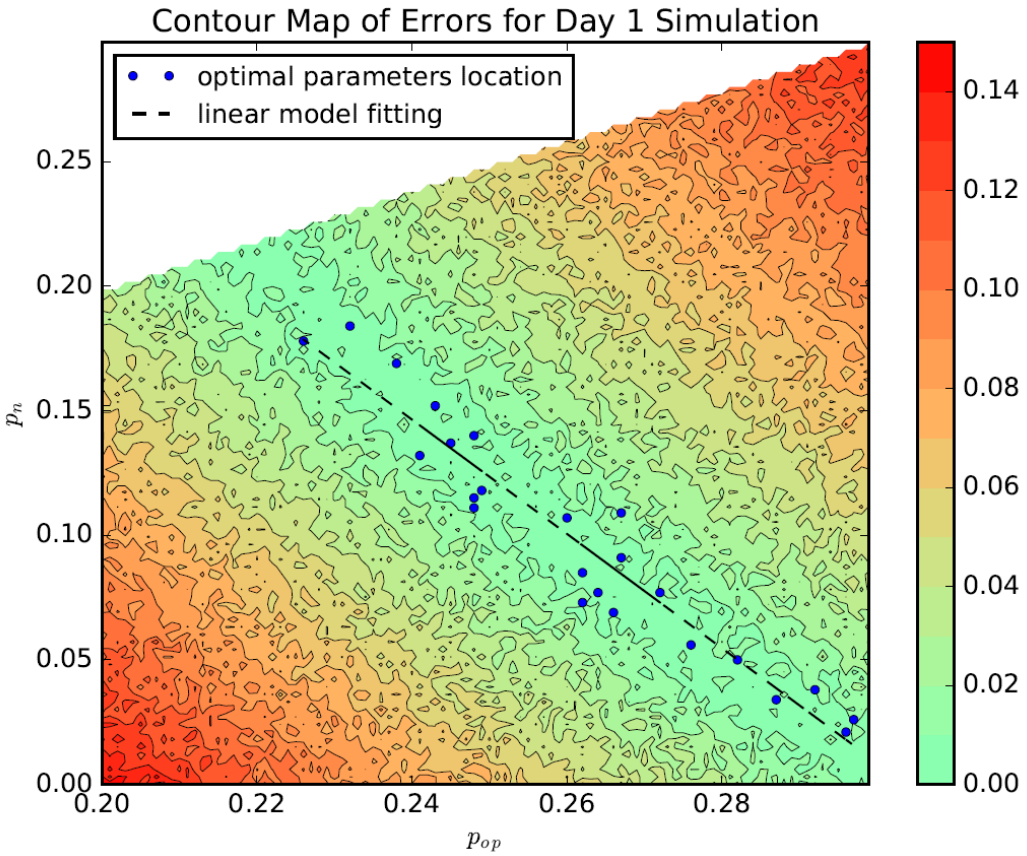

Figure 45: Contour map illustrating the sensitivity of information diffusion model on the Day 1 of Bernardo wildfire data

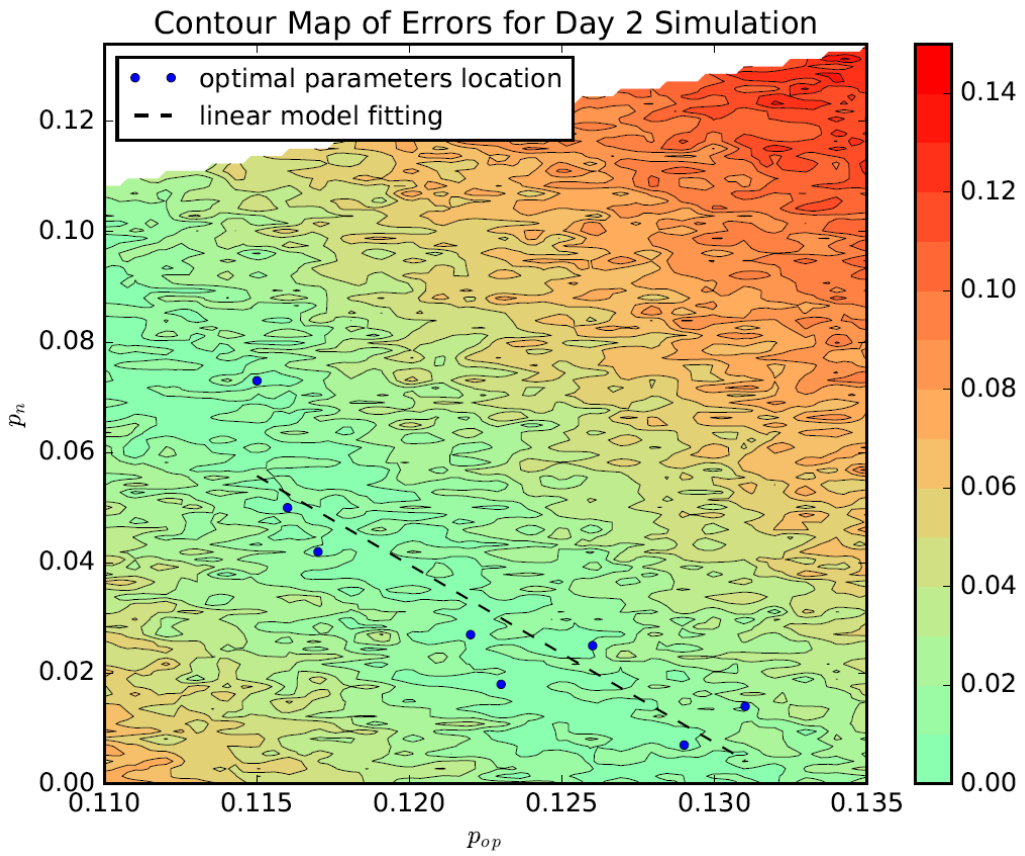

Figure 46: Contour map illustrating the sensitivity of information diffusion model on the Day 2 of Bernardo wildfire data

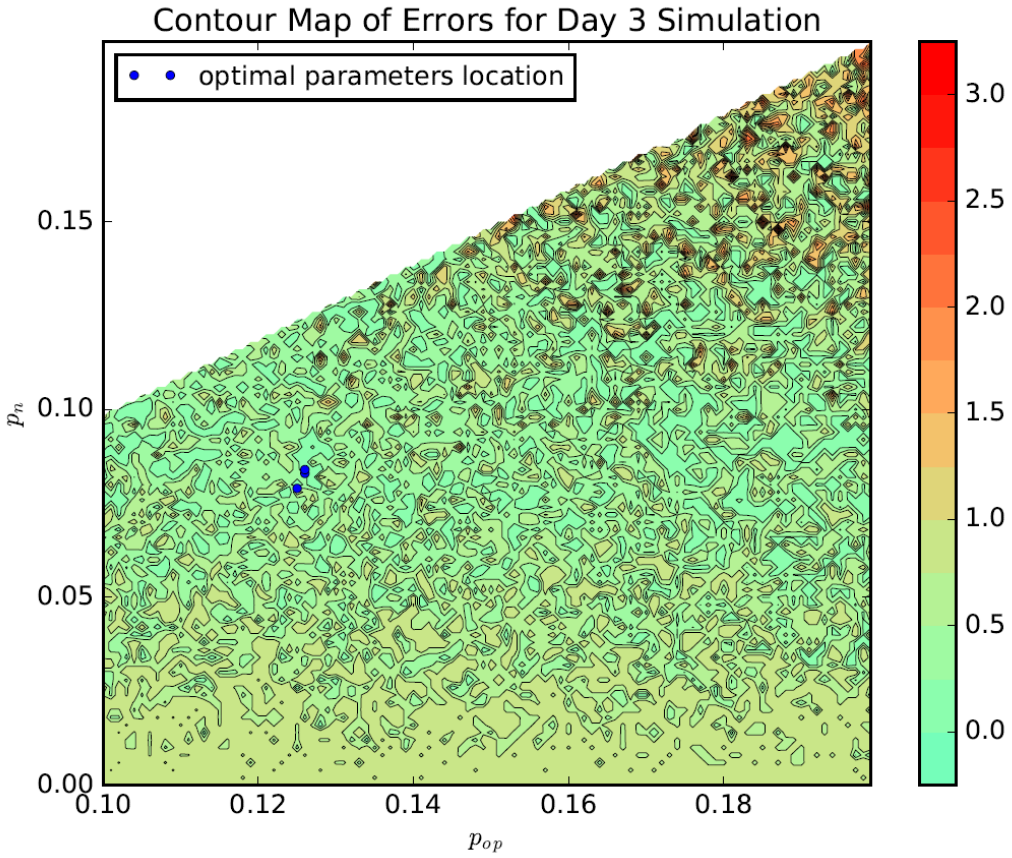

Figure 47: Contour map illustrating the sensitivity of information diffusion model on the Day 3 of Bernardo wildfire data

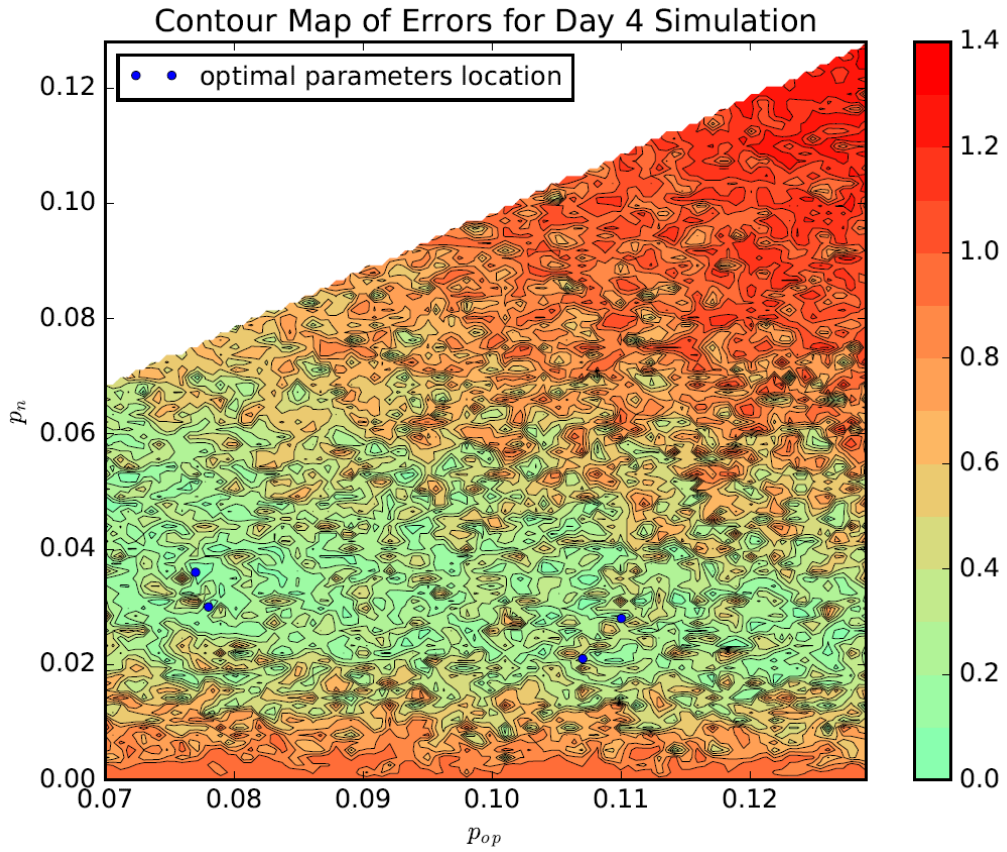

Figure 48: Contour map illustrating the sensitivity of information diffusion model on the Day 4 of Bernardo wildfire data

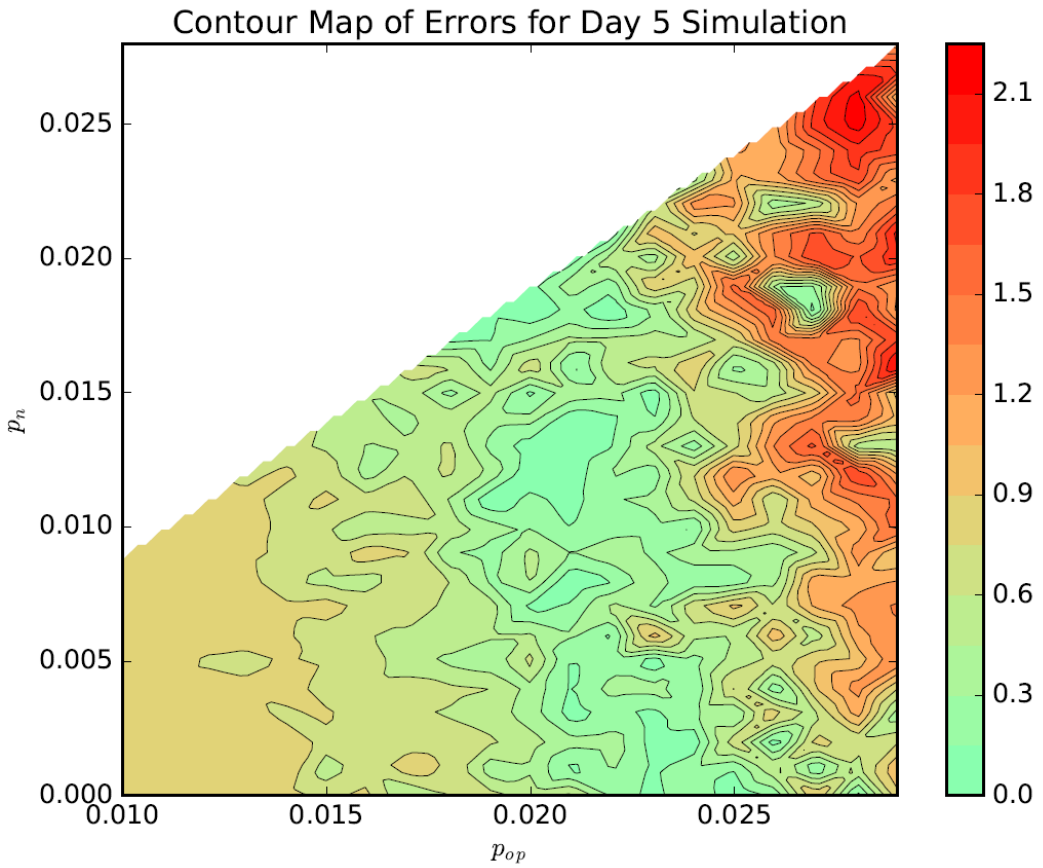

Figure 49: Contour map illustrating the sensitivity of information diffusion model on the Day 5 of Bernardo wildfire data

4.4.2 Social and Diffusion Links

In the world of Twitter, information spread can be observed through retweets in Twitter. The retweets can be seen as the diffusion links in the networks that differentiate social links in that retweets do not require a follow relationship. In this section, both social and diffusion links are examined based on retweets in the dataset of Bernardo wildfire.

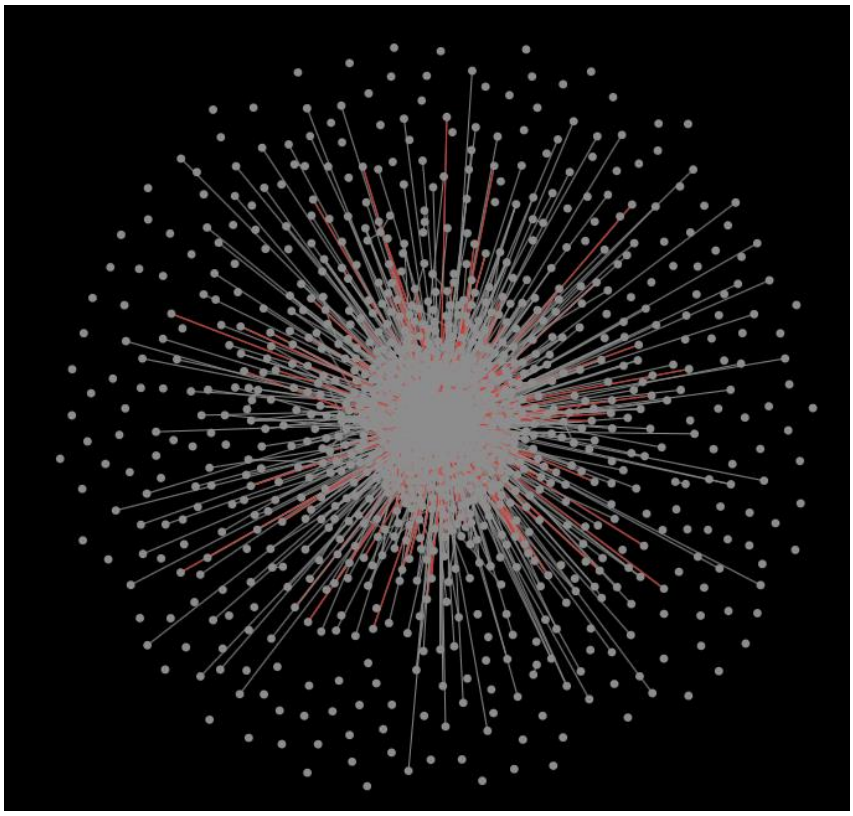

Figure 50: Visualization of Bernardo wildfire tweets diffusion links and social networks among active users in the topic

Figure 50 shows the networks formed by users that participated in Bernardo wildfire tweets. The networks contain both social networks and diffusion networks (1,300 unique users and a total of 8,025 links and 1,075 diffusion links and 6,950 social links, respectively). Red links in the figure represent the links that are both diffusion links and social links. In total, about 34% of the information diffusion links are part of their observable underlying social networks.

There are two reasons why observable underlying social networks accounted for only 34% of the information diffusion in the twitter world. First, Twitter provides search engine for the users who are searching for specific Twitter topics using key words. These users can obtain information related to the interested topic without following the information-source users. Beside, users could opt to receive notifications of current trends (hot topics in Twitter) in their

own homepages after login. The trends are customized based on the location of the users, so for example, Twitter users in San Diego are likely to received Bernardo wildfire topics during the event as opposed to users in Ohio. Second reason pertains to the retweet mechanism that Twitter users have to follow when they retweet.

To give an example, let's have three users: A, B and C. Their relationship in Twitter is that User C follows User B who is follower of User A. If User A posted a tweet and User B, as the follower, retweeted it, User C will then receive the retweet posted by User B. However, if User C also want to retweet the tweet, the retweet will belong to User A who is the original author, although User C does not follow User A. In this process of retweet, the role of User B and its relationship with User C are ignored. Thus, the retweet mechanism indicates that there may exist a good portion of retweets that are actually attributed to their social networks.

From Figure 50, it is easy to notice that there were still a portion of nodes that were not included in either social networks or diffusion networks. They seemed to be standalone individuals that participated in the topic but had no interactions with others in the networks. These individuals accounted for about 19% of total users (241 out of 1,300). Therefore, how do they know about the wildfire? The information source of them can be speculated by examining the content of their text. In general, if a user shared a news from a news website by clicking the Twitter button on the website, his/her tweet will contain "http://...." at the end of the tweet. This pattern implied that their posts were not original and were adopted from a website page. Thus, in the case of Bernardo wildfire, there were 67% of the tweets confirm with such pattern, indicating that a large portion of these standalone tweets were actually from the traditional media such as newspapers.

From the example of Bernardo wildfire, we can see that Twitter is actually a composition of social networking service and information networking service. Although information networks are more noticeable from Twitter's features such as follower-following asymmetrical social relationship, retweet, and twitter search engine while social network is the cornerstone that triggers the large information cascade. Although people can view and share popular tweets even not connecting the authors, without the help of social networks, these popular tweets would not go viral in the first place.

CHAPTER V. CONCLUSION AND DISCUSSION

In the information era, how to disseminate information to the targets in a short time becomes extremely important for different reasons. While sometimes information can be easily disseminated with only minor efforts and resources, it is also the case that information diffusion sometimes dies before it could reach out. Although the content of information itself plays a critical role in terms of determining how far and fast the information can spread, the intermediaries through which the information propagates are also significant factors that influence the information spread. One of the intermediaries is social network.

Social network as the classic channel for people acquiring information is still an essential source of information nowadays. With the prevalence of social media, online social networks become popular and because of the low accessibility, these networks are valuable sources for us in investigating how information diffuses on social networks. This thesis aims at finding efficient information diffusion in online social networks. The studies carried out in the thesis research addressed the problems from two aspects: the structure of online social networks and the early adopters of the information.

More specifically, the analysis of efficient information diffusion is structured and organized into three parts. The first part explored the effect of different structures of social networks on efficient diffusion of information, searching for the relationship between certain characteristics of the networks, and the effectiveness of information diffusion on these networks. The second part experimented with six centralities and heuristics of identifying influential early adopters in the social networks so that more people can adopt the information with less of those

early adopters. The third part examined the diffusion process in real online social networks of Twitter, trying to locate the appropriate propagation probabilities for the agent-based model (ABM) that was proposed and developed technically for the experiment and analysis.

Another important purpose of this thesis research is the intention for demonstrating the strength of ABM. All the functions and experiments of either artificial networks or real networks were implemented by NetLogo which is one of the most widely used platforms for building ABM. What makes ABM powerful is the preciseness of implementing a model, the heterogeneity and interaction among the agents, and more importantly, the interrelations of model parameters which are not permitted in traditional regression models. It is believed that ABM is a promising approach that can help to better understand information diffusion and other more complex human dynamics. ABM becomes more feasible especially with the increasingly powerful computing resources.

The summary of the contributions of the thesis is stated as follows:

- The thesis proposed and developed an ABM in NetLogo for simulating information diffusion on both self-generating artificial networks and real social networks. The NetLogo model has user interface which allows most people to use it without much knowledge about programming. (Sec. 4.1)
- With the same number of nodes and edges, the network having lower average path length or lower average clustering coefficient tended to have wider information diffusion. Such characteristics of network would contribute to the initial prediction of information diffusion without in-depth and time-consuming analyses. (Sec. 4.2)
- Degree discount had the overall best performance on locating influential early adopters in three artificial networks (preferential attachment, random and small-

world). Greedy algorithm only performed the best (among all centralities and heuristic algorithms) in lattice network. In order to assure the comparability of the networks, this research used artificial networks. It should be pointed out that real network samples with exact same number of nodes as artificial networks would greatly help evaluate the performance of centralities and heuristics. However, finding such real network samples is "finding a needle in a hay stack" (Sec. 4.3)

- How to locate optimal early adopters in order to satisfy efficient information diffusion mainly depended on the network structure and propagation probabilities among individuals in the network. The ABM is a powerful approach for handling the such challenges in various networks and propagation probabilities. (Sec. 4.3)
- This thesis examined an information diffusion in a real online social network from Twitter. Using the ABM, it was found that there was a range of optimal propagation parameter pairs that could mimic the information diffusion in real social network. Such range of propagation probabilities usually decrease along with time unless a new update emerged in the topic. This finding contributes to the understanding of the evolution of information diffusion in an event. (Sec. 4.4)
- In the case of dataset in the thesis, there were only 34% of the information diffusion comes from observable underlying social networks. The underlying social network, however, was the essential factor for the information going viral. The problem is that the result was derived from a fraction of the data set in Twitter. A data set that including all the tweets in Twitter would help reinforce the conclusion. (4.4.1)

In short, this thesis research contributes to the field of ABM in simulating efficient information diffusion in online social networks. Such work creates new knowledge on

innovation dissemination and the optimization of decision making for business leaders and policy makers and human dynamics. It also advances the field of geography in that a good understanding of information diffusion provides a hint to understanding social events and incidents and improving people's situation awareness of their surroundings.

Application of this research findings could be used in multiple fields. In academy, the thesis could contribute to the methodological knowledge of information diffusion on networks and solutions of efficient information propagation in social networks. In business, the efficient information diffusion could help companies save money on their advertisement strategies especially in social media. In incidents, particularly in urgent situations, an efficient and effective information diffusion could make up for the shortage of traditional media. For example, in a tsunami, people may not get the updated information through traditional media such as TV or radio. Information in social media could be an essential information source for these people. Through targeting small amount of people and letting the information spread fast through the communities, it could largely reduce people's vulnerability. At las but not the least, this ABM in Netlogo is open-source, which means researchers who have better ideas are able to and encouraged to modify this Netlogo model in order to serve their own ends. In summary, this model will contribute to our knowledge of information both in academic and practical, hoping that it would reveal a tip of secret hiding behind the information diffusion.

As in many other studies of experimenting with simulations, there are some limitations to the work presented here, which also point to directions of the potential future work. First, the networks that can be simulated in the ABM of NetLogo are not large because of the limitations of ordinary PC's computing resources in terms of data storage and computation speed. An ABM

that can handle a network with more than hundreds of thousands of nodes and edges would be more realistic to those networks in real world.

Second, although the ABM could simulate a range of optimal propagation parameters, there is still room for improvement in terms of the accuracy of estimating parametric values. For example, if the difference of propagation probabilities between opinion leaders and ordinary people can be measured, the accuracy of the ABM to mimic information could be improved significantly.

Third, the study used the tweets of Bernardo wild fire as the example, which contains different styles of tweets. These styles include text, image and video. Although these tweets could be operated under the same way (e.g. comment and retweet), there might exist difference in the attractiveness from users. Users, for example, are probably interested in retweeting pictures rather than plain text. Thus, it is believed that analyzing such difference could help better understand the information diffusion on different style of messages.

Fourth, efficiency in the study means propagating the information to the public with least amount of resources. It especially restricted to the budgets of the information spreaders because of the effect of diminishing return. Additionally, there are some differences in defining efficiency among different topics. While this research defined efficiency by targeting most of the people, it does not consider time during which information spread in the networks. In some emergency incidents, time probably would be thought as the essential factor by the consideration of the information spreader. Thus, future research incorporating the factor of time is necessary. Further, the model in this research only consider two type of people: opinion leaders and common people. The propagation probabilities by the same type of people is assumed to be the same. However, strength of tie exists in the social networks that people usually have strong relation to their best

friends or families, resulting in a bigger probability among them. Should this factor to be considered in the model, the simulation will be improved in a large scale. Of course it requires the analysis on a large amount of social network data, which rightly fits to the popularity of big data.

At last but not the least, Twitter provides an excellent resource for researchers studying information diffusion in online social networks. However, because of the special features of Twitter, observing the effect of social network on information diffusion becomes difficult. Thus, more studies are needed using more diverse dataset such as those from Facebook, Google+ and other social networking sites.

REFERENCES

Ahn, Yong-Yeol, James P. Bagrow, and Sune Lehmann. "Link communities reveal multiscale complexity in networks." Nature 466.7307: 761-764. (2010)

Alvarez-Hamelin, J. I., and J. R. Busch. "A low complexity visualization tool that helps to perform complex systems analysis." New Journal of Physics10.12 (2008): 125003.

Aral, Sinan, and Dylan Walker. "Identifying influential and susceptible members of social networks." Science 337.6092 (2012): 337-341.

Bakshy, Eytan, et al. "The role of social networks in information diffusion."Proceedings of the 21st international conference on World Wide Web. ACM, (2012)

Barabási, Albert-László, and Réka Albert. "Emergence of scaling in random networks." science 286.5439 (1999): 509-512.

Barabási, Albert-László, Réka Albert, and Hawoong Jeong. "Scale-free characteristics of random networks: the topology of the world-wide web."Physica A: Statistical Mechanics and its Applications 281.1 (2000): 69-77.

Batagelj, Vladimir, and Matjaž Zaveršnik. "Fast algorithms for determining (generalized) core groups in social networks." Advances in Data Analysis and Classification 5.2 (2011): 129-145.

Bodendorf, Freimut, and Carolin Kaiser. "Detecting opinion leaders and trends in online social networks." Proceedings of the 2nd ACM workshop on Social web search and mining. ACM, (2009)

Bonabeau, Eric. "Agent-based modeling: Methods and techniques for simulating human systems." Proceedings of the National Academy of Sciences 99.suppl 3 (2002): 7280-7287.

Bonacich, Phillip. "Power and centrality: A family of measures." American journal of sociology (1987): 1170-1182.

Brown, Jacqueline Johnson, and Peter H. Reingen. "Social ties and word-of-mouth referral behavior." Journal of

Carmi, Shai, et al. "A model of Internet topology using k-shell decomposition." Proceedings of the National Academy of Sciences 104.27 (2007): 11150-11154.

Centola, Damon. "The spread of behavior in an online social network experiment." science 329.5996 (2010): 1194-1197.

Chen, Wei, Yajun Wang, and Siyu Yang. "Efficient influence maximization in social networks." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009

Dangalchev, Chavdar. "Generation models for scale-free networks." Physica A: Statistical Mechanics and its Applications 338.3 (2004): 659-671.

Dodds, Peter Sheridan, and Duncan J. Watts. "Universal behavior in a generalized model of contagion." Physical review letters 92.21: 218701. (2004)

Doumit, Gaby, et al. "Opinion leaders and changes over time: a survey."Implementation science 6.1 (2011): 1.

Easley, David, and Jon Kleinberg. Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press, (2010).

Erdös, Paul, and Alfréd Rényi. "On random graphs, I." Publicationes Mathematicae (Debrecen) 6 (1959): 290-297.

Fortunato, Santo. "Community detection in graphs." Physics Reports 486.3: 75-174. (2010)

Freeman, Linton C. "Centrality in social networks conceptual clarification." Social networks 1.3 (1978): 215-239.

Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." Proceedings of the National Academy of Sciences 99.12: 7821-7826. (2002)

Goldenberg, Jacob, Barak Libai, and Eitan Muller. "Talk of the network: A complex systems look at the underlying process of word-of-mouth." Marketing letters 12.3 (2001): 211-223.

Grabowicz, Przemyslaw A., et al. "Social features of online networks: The strength of intermediary ties in online social media." PloS one 7.1 (2012): e29358.

Guille, Adrien, et al. "Information diffusion in online social networks: A survey."ACM SIGMOD Record 42.2 (2013): 17-28.

Herr, Paul M., Frank R. Kardes, and John Kim. "Effects of word-of-mouth and product-attribute information on persuasion: An accessibility-diagnosticity perspective." Journal of consumer research 17.4 (1991): 454-462.

Herrmann, Jeffrey, et al. "An agent-based model of urgent diffusion in social media." Robert H. Smith School Research Paper (2013)

Huberman, Bernardo A., Daniel M. Romero, and Fang Wu. "Social networks that matter: Twitter under the microscope." Available at SSRN 1313405 (2008).

Huckfeldt, R. Robert, and John Sprague. Citizens, politics and social communication: Information and influence in an election campaign. Cambridge University Press, (1995)

J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In Proc. of the 7th ACM conference on Electronic commerce. ACM, (2006)

K.Satio, M.Kimura, K.Ohara, and H.Motoda. "Selecting information diffusion models over social networks for behavioral analysis." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, (2010).

Katz, Elihu. "The two-step flow of communication: An up-to-date report on an hypothesis." Public opinion quarterly 21.1 (1957): 61-78.

Kempe, David, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, (2003)

Kimura, Masahiro, Kazumi Saito, and Ryohei Nakano. "Extracting influential nodes for information diffusion on a social network." AAAI. Vol. 7. (2007)

Kleinberg, Jon. "The small-world phenomenon: An algorithmic perspective."Proceedings of the thirty-second annual ACM symposium on Theory of computing. ACM, 2000.

Leskovec, Jure, et al. "Cost-effective outbreak detection in networks."Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007.

Lin, Nan. "Social networks and status attainment." Annual review of sociology (1999): 467-487.

M. Granovetter. The strength of weak ties. American Journal of Sociology, 78(6):1, (1973)

M. Granovetter. Threshold models of collective behavior. American journal of sociology, pages 1420-1443, (1978)

M. Newman and J. Park. Why social networks are different from other types of networks.Physical Review E, 68(3):036122, (2003)

Macal, Charles M., and Michael J. North. "Tutorial on agent-based modeling and simulation." Proceedings of the 37th conference on Winter simulation. Winter Simulation Conference, (2005).

Milgram, Stanley. "The small world problem." Psychology today 2.1 (1967): 60-67.

Miorandi, Daniele, and Francesco De Pellegrini. "K-shell decomposition for dynamic complex networks." Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium on. IEEE, 2010.

Mislove, Alan, et al. "Measurement and analysis of online social networks."Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, (2007)

Moreno, Yamir, Maziar Nekovee, and Amalio F. Pacheco. "Dynamics of rumor spreading in complex networks." Physical Review E 69.6 (2004): 066130.

Newman, Mark EJ. "Mixing patterns in networks." Physical Review E 67.2 (2003): 026126.

Newman, Mark EJ. "Modularity and community structure in networks."Proceedings of the National Academy of Sciences 103.23: 8577-8582. (2006)

Ning, Ma, et al. "Recognition of online opinion leaders based on social network analysis." Active Media Technology. Springer Berlin Heidelberg, (2012): 483-492.

Onnela, J-P., et al. "Structure and tie strengths in mobile communication networks." Proceedings of the National Academy of Sciences 104.18: 7332-7336. (2007)

Papert, S. Mindstorms. NY: Basic Books. (1980)

Pei, Sen, et al. "Searching for superspreaders of information in real-world social media." Scientific reports 4 (2014).

Rand, W., Rust, R.T. “Agent-based modeling in marketing: Guidelines for rigor.” International Journal of Research in Marketing 28(3) (2011) 181-193

Rogers, E. M. Diffusion of innovations (5th ed.). New York: Free Press. (2003).

Rogers, Everett M. Diffusion of innovations. Simon and Schuster, (2010)

Rosvall, M., and C. T. Bergstrom. "Maps of information flow reveal community structure in complex networks." arXiv preprint physics.soc-ph/0707.0609 (2007).

Sabidussi, Gert. "The centrality index of a graph." Psychometrika 31.4 (1966): 581-603.

Seidman, Stephen B. "Network structure and minimum degree." Social networks 5.3 (1983): 269-287.

Steven F. Railsback; Volker Grimm. Agent-Based and Individual-Based Modeling: A Practical Introduction. Cambridge: Princeton University Press. ISBN 978-0-691-13674-5. (2011)

Tisue, Seth, and Uri Wilensky. "Netlogo: A simple environment for modeling complexity." International conference on complex systems. Vol. 21. (2004).

Trusov, Michael, Randolph E. Bucklin, and Koen Pauwels. "Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site." Journal of marketing 73.5 (2009): 90-102.

Valente, Thomas W., and Patchareeya Pumpuang. "Identifying opinion leaders to promote behavior change." Health Education & Behavior (2007).

Valente, Thomas W., and Rebecca L. Davis. "Accelerating the diffusion of innovations using opinion leaders." The Annals of the American Academy of Political and Social Science 566.1 (1999): 55-67.

Van Eck, Peter S., Wander Jager, and Peter SH Leeflang. "Opinion leaders' role in innovation diffusion: A simulation study." Journal of Product Innovation Management 28.2 (2011): 187-203.

Watts, Duncan J. "A simple model of global cascades on random networks."Proceedings of the National Academy of Sciences 99.9: 5766-5771. (2002)

Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of ‘small world’networks." nature 393.6684 (1998): 440-442.

Weng, Lilian, Filippo Menczer, and Yong-Yeol Ahn. "Virality prediction and community structure in social networks." Scientific reports 3 (2013).

Weng, Lilian. “Information diffusion on online social networks.” Doctoral Dissertation Indiana University. (2014)

Wilensky, U. NetLogo Preferential Attachment model. http://ccl.northwestern.edu/NetLogo/models/PreferentialAttachment. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. (2005)

APPENDIX

A: NetLogo Code

```
;; author: Zhuo Chen

;; This model is created by modifying example model from nw extention.
(https://github.com/NetLogo/NW-
Extension/blob/5.x/demo/Network%20Extension%20General%20Demo.nlogo)

;; Generally, it addded importing and saving GraphML format networks, using
six centralities and heuristics to locate early adopters, and information
diffusion function.


extensions [ nw ]


; we have two different kind of link breeds, one directed and one
undirected, just

; to show what the different networks look like with directed vs. undirected
links

directed-link-breed [ dirlinks dirlink ]

undirected-link-breed [ unlinks unlink ]

; below two link breeds are created for importing real networks from
Twitter.

undirected-link-breed [ SocialLinks SocialLink ]

undirected-link-breed [ DiffusionLinks DiffusionLink ]


globals [

  opinion-leaders                ; AGENTSETS to store nodes who are opinion
leaders

  seed-nodes                     ; AGENTSETS to store seed nodes that adopt
information at first

  adopter-size-list              ; For calculate the mean adotion size

  file                           ; FILE to store the output experiment

]


turtles-own [
```

```
  ; attributes of node in real network (tweets)
  ID
  Name
  Identity
  Tweet_Time
  X_COR
  Y_COR

  influence                       ; numeric value to determine how
influential the turtle is
  adopt?
  has-spread?                     ; boolean for independent cascade mode
  opinion-leader?
  dis-degree                      ; discount degree
  ks-degree                       ; k-shell degree
  nb-seed-neigbhors               ; number of neighbors of the turtle that
are already selected as seeds
]



;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;; Clear functions
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

to clear
  clear-all
  set-current-plot "Degree distribution"
  set-default-shape turtles "circle"
  reset-ticks
end

to clear-graph
  ; clear early adopters and opinion leaders but keep the underlying graph.
  set seed-nodes nobody
  ask turtles [
```

```
    set shape "circle"
    set size 1
    set influence 0
    set color grey
    set adopt? false
    set has-spread? false
  ]
end

to reset-graph
  ; used for swithcing the method in set seed nodes function.
  ask turtles [
    set shape "circle"
    set size 1
    set influence 0
    set color grey
    set adopt? false
    set has-spread? false
  ]
  if seed-nodes != nobody and seed-nodes  != 0 [ask seed-nodes [set adopt?
true]]
end

;; Reports the link set corresponding to the value of the links-to-use combo
box
to-report get-links-to-use
  report ifelse-value (links-to-use = "directed")
    [ dirlinks ]
    [ unlinks ]
end



;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;; Layouts
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
```

```
to redo-layout [ forever? ]
  if layout = "radial" and count turtles > 1 [
    layout-radial turtles links ( max-one-of turtles [ count my-links +
count my-out-links + count my-in-links ] )
  ]
  if layout = "spring" [
    let factor sqrt count turtles
    if factor = 0 [ set factor 1 ]
    repeat ifelse-value forever? [ 1 ] [ 50 ] [
      layout-spring turtles links (1 / factor) (14 / factor) (1.5 / factor)
      display
      if not forever? [ wait 0.005 ]
    ]
  ]
  if layout = "circle" [
    layout-circle sort turtles max-pycor * 0.9
  ]
  if layout = "tutte" [
    layout-circle sort turtles max-pycor * 0.9
    repeat 10 [
      layout-tutte max-n-of (count turtles * 0.5) turtles [ count my-links ]
links 12
    ]
  ]
end

to layout-once
  redo-layout false
end

to spring-forever
  set layout "spring"
  redo-layout true
end
```

```
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;; Network Generators
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

to generate [ generator-task ]
  ; we have a general "generate" procedure that basically just takes a task
  ; parameter and run it, but takes care of calling layout and update stuff
  set-default-shape turtles "circle"
  run generator-task
  ask turtles [
    set color grey
    set adopt? false
    set has-spread? false
  ]
  layout-once
  update-plots
end

to preferential-attachment
  generate task [ nw:generate-preferential-attachment turtles get-links-to-
use nb-nodes ]
end

to ring
  generate task [ nw:generate-ring turtles get-links-to-use nb-nodes ]
end

to star
  generate task [ nw:generate-star turtles get-links-to-use nb-nodes ]
end

to wheel
  ifelse (links-to-use = "directed") [
    ifelse spokes-direction = "inward" [
      generate task [ nw:generate-wheel-inward turtles get-links-to-use nb-
nodes ]
```

```
    ]
    [ ; if it's not inward, it's outward
      generate task [ nw:generate-wheel-outward turtles get-links-to-use nb-
nodes ]
    ]
  ]
  [ ; for an undirected network, we don't care about spokes
    generate task [ nw:generate-wheel turtles get-links-to-use nb-nodes ]
  ]
end


to lattice-2d
  generate task [ nw:generate-lattice-2d turtles get-links-to-use nb-rows
nb-cols wrap ]
end


to small-world
  generate task [ nw:generate-small-world turtles get-links-to-use nb-rows
nb-cols clustering-exponent wrap ]
end


to generate-random
  generate task [ nw:generate-random turtles get-links-to-use nb-nodes
connexion-prob ]
end



;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;; Saving and loading of network files
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;


;; save and load GraphML, user could change the name of the to-save graphML
;; and open a GraphML from Windows Explore
to save-graphml
  nw:set-context turtles get-links-to-use
  nw:save-graphml "demo.graphml"
end
```

```
to load-graphml
  set-default-shape turtles "circle"
  nw:set-context turtles get-links-to-use
  let graphmlfile user-file
  if is-string? graphmlfile
  [nw:load-graphml graphmlfile]
  ask turtles [set label ""]
  layout-once
  update-plots
end



;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;; Opinion leaders
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

to set-opinion-leaders
  ;;; first clear existing opinion leaders
  ask turtles [set opinion-leader? false]
  set opinion-leaders nobody
  ;ask turtles [set color white]
  ;; sort turtles in a order of degree by descending
  foreach sort-by [[count link-neighbors] of ?1 > [count link-neighbors]
of ?2] turtles
  [
    ;; if leaders in opinion-leader less than the desired number
    ;; the number of opinon-leaders here are 10% of total population
    ifelse (opinion-leaders = nobody) or (count opinion-leaders < precision
(count turtles / 10) 0)
    [
      ask ?
        [
          if [shape] of ? != "target"; if it has not been selected as seed
nodes
          [
            set shape "square"
```

```
            set opinion-leader? true     ; square is reserved for opinion
leaders in this model
          ]
          set color yellow         ;; yellow is also for opinion ledaers ;
red target means it is
          set size 2               ;; both opinion leader and seed nodes
          set opinion-leader? true
        ]
      ;; add current turtle into opinion-leader agentset
      set opinion-leaders (turtle-set ? opinion-leaders)
    ]
    [stop]
  ]
end



;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;; Seed Nodes / Early Adopters
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;


to set-seed-nodes
  ;; set the seed nodes in the network using the methods chosen from seed-
nodes-preference
  ;; draw the seed nodes as red target after the calculation.
  reset-graph
  if seed-nodes-preference = "Betweenness" [
    draw-seed-nodes task nw:betweenness-centrality
  ]
  if seed-nodes-preference = "Eigenvector" [
    draw-seed-nodes task nw:eigenvector-centrality
  ]
  if seed-nodes-preference = "Closeness" [
    draw-seed-nodes task nw:closeness-centrality
  ]
  if seed-nodes-preference = "Diffusion simulation" [   ; Greedy algorithm
    ; start simulation to find seed-nodes
    simulate-seed-nodes
```

```
    draw-seed-nodes-simple
    ; draw these seed nodes
  ]
  if seed-nodes-preference = "K-Shell" [
    k-shell
    draw-seed-nodes-simple
  ]
  if seed-nodes-preference = "Degree Discount" [
    degree-discount
    draw-seed-nodes-simple
  ]

  ;; clear "Adoption Experiment" plot and the adoption size/influence
  set-current-plot "Adoption Experiment"
  clear-plot
  set adopter-size-list []
end

to draw-seed-nodes [ measure ]
  set seed-nodes nobody
  foreach sort-by [[runresult measure] of ?1 > [runresult measure] of ?2]
turtles
    [
      ifelse (seed-nodes = nobody) or (count seed-nodes < nb-seed-nodes)
      [
        ask ? [
          set shape "target"
          set color red
          set size 2
        ]
        set seed-nodes (turtle-set ? seed-nodes)
      ]
      [stop]
    ]
end
```

```
to draw-seed-nodes-simple
  ask seed-nodes [
    set shape "target"
    set color red
    set size 2
    set adopt? true
  ]
end


to simulate-seed-nodes
;; simulate the diffusion process to find early adopters using greedy
algorithm
  let node-counter 0
  ifelse seed-nodes = nobody or seed-nodes = 0
  [print "there is no seed-nodes" set seed-nodes nobody]
  [set node-counter count seed-nodes]


  while [node-counter < nb-seed-nodes]  ; node-counter means number of seed
nodes
  [
    ask turtles with [adopt? = false]
    [
      let spread-influence 0
      repeat 300             ; the number of simulations on each seed node
      [
        adopt-cascade  ; run diffusion on current turtle with seed-nodes
        set spread-influence count turtles with [adopt? = true] + spread-
influence
        ask turtles with [adopt? = true][set adopt? false set has-spread?
false]  ; reset the status of all turtles
        if seed-nodes != nobody [ask seed-nodes [set adopt? true]]
      ]
      ; recorded the influence with current turtle
      set influence spread-influence / 200
    ]
    ; find the turtle with max influence and add it in seed-nodes turtle-set
    let seed-node one-of turtles with [influence = max [influence] of
turtles]
```

```
    set seed-nodes (turtle-set seed-node seed-nodes)
    ask seed-nodes [set adopt? true]
    print seed-nodes
    set node-counter node-counter + 1
  ]
end


to degree-discount     ; degree discount heuristics developed by Wei Chen et
al. (http://dl.acm.org/citation.cfm?id=1557047)
  set seed-nodes no-turtles
  ask turtles [
    set dis-degree count link-neighbors   ;initiate the discount degree for
all the nodes
    set nb-seed-neigbhors count link-neighbors with [member? self seed-
nodes]
  ]


  let node-counter 0
  while [node-counter < nb-seed-nodes]  ; node-counter means number of seed
nodes
  [
    ; finding the nodes that has the highest discount degree as the seed
nodes/early adopters
    let u max-one-of turtles with [not member? self seed-nodes][dis-degree]
    set seed-nodes (turtle-set u seed-nodes)
    ask u [
      ask link-neighbors with [not member? self seed-nodes]
      [
        set nb-seed-neigbhors nb-seed-neigbhors + 1
        set dis-degree count link-neighbors - 2 * nb-seed-neigbhors - (count
link-neighbors - nb-seed-neigbhors) * p-adoption * nb-seed-neigbhors
      ]
    ]
    set node-counter node-counter + 1
  ]
  print seed-nodes
end
```

```
to k-shell
  ; k-shell heuristics
  ; because we cannot remove nodes in the network as the original methods
suggested,
  ; this methods turned to rank each node using k-shell by multiplying a
large number.
  let my-count 1     ; my-count for the rank of the node using k-shell
  set seed-nodes nobody


  ask turtles [
    ; initializing virable
    set ks-degree 0
    set influence 0
  ]


  ; calculating the influence of all the nodes in the network using k-shell
methods.
  while [any? turtles with [ks-degree = 0]]
    [
      ; if there are any turtles that have ks-degree 0 and the number of
their neighbors that have 0 ks-degree is less than my-count
      ; rank them as the #my-count shell of the network
      while [any? turtles with [ks-degree = 0 and (count link-neighbors with
[ks-degree = 0]) <= my-count]]
      [
        ask turtles with [ks-degree = 0 and (count link-neighbors with [ks-
degree = 0]) <= my-count]
        [
          set ks-degree my-count * 5000  ; 5000 as the multiplier for
calcualting k-shell influence
          set influence ks-degree + count link-neighbors  ; diffirenciate
the nodes that has the same ks-degree by counting link-neighbors
        ]
      ]
      set my-count my-count + 1
    ]
   set seed-nodes max-n-of nb-seed-nodes turtles [influence]
   print seed-nodes
end
```

```
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;; Information Diffusion Process
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;


to adopt-cascade   ; information diffsion. runninng until no one can be
influenced
  ;set color red
  set adopt? true  ;; set current node as innitial adopters


  while [true]  ;; it's an infinite loop
  [
    let initial-adopters turtles with [adopt? = true and has-spread? =
false]
    ifelse any? initial-adopters   ; check if there is any new adopters
    [
      ask initial-adopters    ; ask all new adopters to spread the
inforamtion
      [
        let potential-adopters link-neighbors with [adopt? = false]  ;;
those who have not adopted the inforamtion but are neighbors of adopters
        if any? potential-adopters
        [
          ask potential-adopters
          [
            let adopted-fraction 0
            let adopted-neighbors link-neighbors with [adopt? = true]
            let nb-neibhbors count link-neighbors
            let nb-adopted-leaders count adopted-neighbors with [opinion-
leader? = true]
            set adopted-fraction (count adopted-neighbors) / nb-neibhbors


            let random-thresh random-float 1.0  ;; use random values as the
threshold of adoption.
            if random-thresh < (nb-adopted-leaders * p-op-leader +  (count
adopted-neighbors - nb-adopted-leaders) * p-adoption) / count adopted-
neighbors  ;; p-adoption is the paratmeter controlled by slider in the
interface
```

```
              [
                set adopt? true
                ;set color red
              ]
            ]
          ]
          set has-spread? true
        ]
      ]
      [stop] ;; else (not more new adopters) stop!!!
    ]
end

to spread-information  ; only diiffse information once.
  ask seed-nodes [
    adopt-cascade
    let adopters turtles with [adopt? = true]
    ask adopters
    [
      set color red
    ]
  ]
end

to reset-diffuse  ; reset the diffusion to the starting status.
  ask turtles with [color = red]
  [
   set color grey
   set adopt? false
   set has-spread? false
   ]
  draw-seed-nodes-simple
end

to experiment
```

```
;; an experiment for testing the performance of the methods of choosing
early adopters.
;; it runs 1000 times simulation on each node until the number of seed nodes
reaches 20.
  set-opinion-leaders
  while [nb-seed-nodes <= 20] [
    set-seed-nodes
    let exp-count  0
    while [exp-count < 1000][   ;; 1000 time simulation on each seed node
      spread-information
      update-plots
      set adopter-size-list lput count turtles with [adopt? = true] adopter-
size-list
      print count turtles with [adopt? = true]
      if file != 0 [write-to-file file]  ;; write the results into a csv
file
      reset-diffuse
      set exp-count exp-count + 1
    ]
    set nb-seed-nodes nb-seed-nodes + 1
  ]
end


to find-parameters
;; THIS FUNCTION IS TO FIND OPTIMAL PARAMETERS BY GRID SEARCHING.
;;;;;;;;;;;;;;EXPLANATION OF VARIABLES;;;;;;;;;;;;;;;;;;;;;;;;;;;
;; probability of adopting from opinion leaders   p-op
;; probability of adopting from normal people     p-nm
;; first run: p-op & p-nm range from 0.1-0.4 where p-op > p-nm, increasement
0.1
;; p-op   p-nm   avg.adoption
  set seed-nodes nobody
  set adopter-size-list []
  let seed-list ["A" "B" "C"]  ; set the early adopters mannually based on
the real network data.

  ;; set up the result of experiment file
  let p-file "parameter_experiment.csv"
```

```
  if (file-exists? (p-file)) [carefully [file-delete (p-file)] [print error-
message]]
  file-open(p-file)
     file-type "p-op,"
     file-type "p-nm,"
     file-type "avg_adoption"
  file-close


  ; set opinion leaders and set-seed-nodes
  set p-op-leader 0.01
  set-opinion-leaders
  ask turtles with [member? name seed-list] [set seed-nodes (turtle-set self
seed-nodes)]
  if seed-nodes = nobody [
    user-message ("Please import real network  and set seed-list before
running this function!")
    stop
    ]
  draw-seed-nodes-simple


  ;; grid search to tranverse all the possbile parameters.
  while [p-op-leader <= 0.03] [
    while [p-adoption < p-op-leader] [
      let exp-count 0
      while [exp-count < 10] [
        spread-information
        update-plots
        set adopter-size-list lput count turtles with [adopt? = true]
adopter-size-list
        reset-diffuse
        set exp-count exp-count + 1
      ]
      ;;write result to the file
      file-open (p-file)
      file-print " "
          file-type p-op-leader                                    file-type
","
```

```
          file-type p-adoption                                    file-type
","
          file-type mean  adopter-size-list                       file-type
","
      file-close

      print mean adopter-size-list
      set adopter-size-list []
      set p-adoption p-adoption + increment
    ]
    set p-adoption 0.0 ;; another round of search
    set p-op-leader p-op-leader + increment
  ]
end



;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;; Dealing with files
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

to set-up-file  ;;  run it before hitting experiment button
  ; it initialize the output file for the results of experiment.
  set file user-new-file
  if (file-exists? (word file ".csv")) [carefully [file-delete (word file
".csv")] [print error-message]]
  file-open (word file ".csv")
     file-type "seed-nodes-preference,"
     file-type "population,"
     file-type "seed nodes size,"
     file-type "number of opinon ledaers,"
     file-type "adoption probability,"
     file-type "adoption probability from opinion ledaers,"
     file-type "adoption size"
  file-close
end

to write-to-file [my-file]
```

```
  file-open (word my-file ".csv")
  file-print " "
     file-type seed-nodes-preference                                    file-
type ","
     file-type count turtles                                            file-
type ","
     file-type count seed-nodes                                         file-
type ","
     file-type count turtles with [opinion-leader? = true]              file-
type ","
     file-type p-adoption                                               file-
type ","
     file-type p-op-leader                                              file-
type ","
     file-type count turtles with [adopt? = true]                       file-
type ","
  file-close
end



;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;; Reporters for monitors
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

to-report mean-path-length
  report nw:mean-path-length
end
```