Models of Influence in Online Social Networks

Kanna AlFalahi,[1,*] Yacine Atif,[1,†] Ajith Abraham[2,‡]
[1] *Faculty of Information Technology, UAE University, Al Ain, United Arab Emirates*
[2] *VSB-Technical University of Ostrava, Ostrava, Czech Republic*

Online social networks gained their popularity from relationships users can build with each other. These social ties play an important role in asserting users' behaviors in a social network. For example, a user might purchase a product that his friend recently bought. Such phenomenon is called social influence, which is used to study users' behavior when the action of one user can affect the behavior of his neighbors in a social network. Social influence is increasingly investigated nowadays as it can help spreading messages widely, particularly in the context of marketing, to rapidly promote products and services based on social friends' behavior in the network. This wide interest in social influence raises the need to develop models to evaluate the rate of social influence. In this paper, we discuss metrics used to measure influence probabilities. Then, we reveal means to maximize social influence by identifying and using the most influential users in a social network. Along with these contributions, we also survey existing social influence models, and classify them into an original categorization framework. Then, based on our proposed metrics, we show the results of an experimental evaluation to compare the influence power of some of the surveyed salient models used to maximize social influence. © 2013 Wiley Periodicals, Inc.

1. INTRODUCTION

Online social networks (OSNs) gained huge popularity since they were introduced a decade ago. Millions of people tend to register and participate in OSNs such as Facebook, LinkedIn, Flickr, MySpace, and Twitter. Facebook by itself accounted for more than 800 million active users in 2011.[1] These social networks have a great impact on people's lives at different levels, and in a variety of contexts. One use of OSNs would be in reporting adversities and boost awareness about a situation, especially in places that lack physical communication facilities, due, for example, to nature disasters such as hurricanes and earthquakes, or simply political censorship. People are increasingly using social networks to spread information during crisis because these networks are handy and easy to use. Acar and Muraki[2] studied posts

[*] Author to whom all correspondence should be addressed; e-mail: K.alfalahi@uaeu.ac.ae.
[†] e-mail: Yacine.atif@uaue.ac.ae
[‡] e-mail: ajith.abraham@ieee.org

INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 00, 1–23 (2013)

View this article online at wileyonlinelibrary.com. • DOI 10.1002/int.21631

on Twitter (called tweets) two weeks after the devastating Tohoku earthquake that resulted in the overwhelmingly destructive Tsunami in Japan during March 2011. They found that people in the affected areas had a tendency to post tweets related to their unsafe situation, while people in remote area post tweets to let their followers know that they are safe. Another widespread use of OSNs occurred during political protests in Tunisia and Egypt in January 2011, where massive antigovernmental demonstrations forced dictatorships to fall. What is interesting in these events is that social network bloggers did not use OSNs to advertise their webpages or encourage people to write about their frustrations, but to engage people and motivate them into taking actions not only online but in real world too, which illustrates the influential power and impact of social networking.

In real-life context, an influencer is a person who is followed by many people and has the power to make changes in a community. The same aspect would occur in OSNs context, as they form a large social space where people are engaging together to build relationships and expand their connections with others. These OSNs have the same traits of real-life communications and many people thrive socially in OSNs as they do in their real life. In past years, influential people were those who have many friends. This idea evolved as influencers started not only to have many friends, but also to actively engage their friendship community into actions. In current days, many influencers drive discussion topics about a specific topic or brand.[3] This kind of influence was the main building stone of the interest graph: a network of people who are interested in each other's content.[4] Interest graphs help brand making of products and services by targeting powerful influential people in social networks.[4]

In this paper, we address social influence by evaluating different models used to measure influence probability. Conceptually, OSNs are related to graph theory,[5] computer science, and social science[6] fields. To study and analyze these networks properly, a combination of these disciplines needs to be considered. Social networks can be modeled as a graph that contains nodes representing members and edges corresponding to the relationship type between the nodes (e.g., friendship). Social networks analysis (SNA)[7] can help tracing the sources and distribution of influence power in social networks, based on the structure of the network. The influential power of a user rises with his relationship among other influential users in the network.

Sociologists studied in the past the power of a specific node in the network by addressing the attributes of centrality using SNA such as degree, closeness, and betweenness centralities.[7] Nodes with high degree, high closeness, and high betweenness will have greater influence. Figure 1 shows a sample social network graph and the edges between the nodes. The graph shows clusters and central nodes that can be sources of great influential power when evaluating their social influence probability. One drawback of measuring influence based on SNA is that centrality is based on the structure of the network, while influence should be based on the dynamics and changes that occur in the OSNs connections and links.

A better understanding of the evolution of social networks leads to a better investigation of the community structure and social influence [8] in these networks. The outcomes of this investigation helps in performing different activities around OSNs-based communities such as targeted advertisement, and items

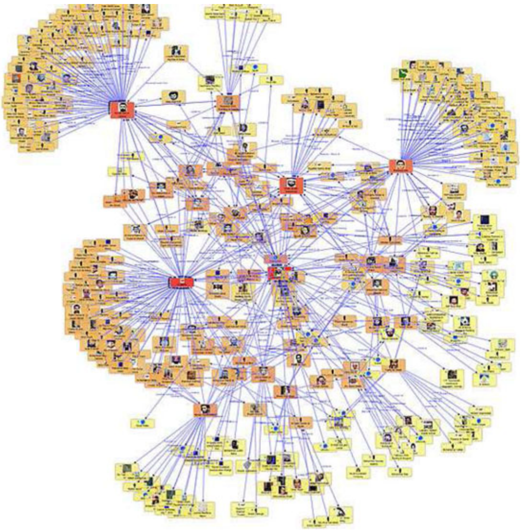


Figure 1. Social network diagram.

recommendation to OSNs users. Research works carried out in this area are few, sparse, and span multiple disciplines. Besides our efforts to conglomerate the state-of-the-art surrounding social influence in OSNs, we also evaluate different approaches and classify them to understand commonalities and distinguish differences, to better maximize influence across OSNs.

The rest of this paper is organized as follows: Section 2 provides basic information about social influence through defining influence concepts in OSNs context and stating some related background. We then discuss some properties of social networks, which are relevant to evaluate influence in OSNs. Section 3 provides a survey on OSNs' social influence related works through an original classification, and then state the problem of measuring influence probability in OSNs. Section 4 reveals a comparison between the different social influence models as well as their limitations, strength, and challenges. Section 5 shows an experiment to compare the influence diffusion in OSNs based on the Linear Threshold and the Independent Cascade propagation models.

2. BACKGROUND OF SOCIAL INFLUENCE

Social influence has been studied by sociologists and social psychologists since the early years of the 20th century.[9] It started in 1898, with the first experiment by Norman Triplett on the phenomenon of social facilitation.[10] This theory implies that people tend to do well in the things they are good at when they are watched by others.[10] One of the main theories of social influence was proposed in 1950 by Leon Festinger called Cognitive Dissonance Theory. The theory is related to how the way of thinking can affect our behavior.[11] In 1959, French and Raven discussed social power and provided formalization for the social influence concept.[12] Research reached more maturity in both theory and methods during the 1980's and 1990's.[9]

Social influence has been studied in different disciplines and has historical roots in sociology through studying opinion formation and the diffusion of innovations;[13,14] and economics, where social influence, represented as theoretical models, shows how individuals are inclined to coordinate their economic decisions.[15,16] Recently, digital social influence research has started to attract more attention due to the availability of many important applications. For example, computer scientists developed models of social influence to support applications such as viral marketing,[17–19] the spread of online news,[20,21] and the growth of online communities.[22]

2.1. Social Influence Definition

Sociologists defined social influence as a "change in an individual's thoughts, feelings, attitudes, or behaviors that results from interaction with another individual or a group".[12] Social influence occurs when an individual changes his/her behavior after interacting with other individuals who tend to be similar or superior.

Social influence involves social correlations, which are divided into three categories as follows:[23]

- **Influence:** where a user performs an action based on his friends' recent actions. For example, when a user purchases a product because one of his friends just bought or recommended that product.
- **Homophily:** a user chooses friends who share the same characteristics;[24,25] this leads to perform the same actions. For example, two persons who have Xbox are more likely to be friends due to the same interest.
- **Confounding factors:** or external influence that affects individuals who are located near each other in the social network. One example would be when two users live in the same city, which makes them perform the same activities like taking the same photos and posting them with the same tags in an online photo-sharing network like Flickr.

Performing social influence across OSNs help diffusing different behaviors, ideas, and new technologies. For example, a fashion company might provide coupons to the most influential users in their social network in exchange of promoting a new product. Different approaches were proposed to leverage social influence[18] considering the effect of influence on business returns growth.[26] These efforts are centered on the process of carefully choosing targets with high influential power as a marketing strategy that leads to high acceptance of a certain product among users of a social network. Social influence is becoming a complex and a subtle force that governs the dynamics of all social networks. Given the high expected returns and the induced complexity, there is a need for methods and techniques to analyze and quantify social influence.

2.2. Properties Of OSNs

Rich properties and components of social networks paved the way for a better analysis of individual user actions, leading to further profiling of users' behavior in OSNs. Through their behaviors, people can influence other users to do specific actions. This is a powerful process which can generate substantial revenues or incite large-scale global actions. Social influence appears as a social correlation pattern where the actions of a user can urge his or her friends to behave in the same way.[27]

OSNs exhibit different properties which make it to study users' actions that could influence communities' behaviors. The availability of rich interactions between users and the large data sets that result from such interactions facilitate social influence analysis. To better understand social influence, we need to describe the social network structure and introduce some properties, which are categorized as follows:

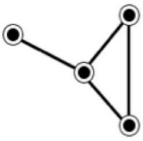
Large scale: Each network has basic properties such as: network order, represented by the number of nodes in the network; the size, that represents the number of edges in the network; and the node degree, which represents the number of edges that are connected to a node. OSNs are large-scale networks with high order and size that may reach millions, and having users with very high degrees. For example, in Facebook, there are over 500 million nodes with an average degree of 130.[1] In Twitter, the nodes of celebrities such as Lady Gaga, Justin Bieber, Britney Spears, and Ashton Kutcher have a degree of more than 6 millions. LinkedIn has more than 90 million nodes, having a new user joining every second.[28]

Network clustering: The idea of clusters or cliques is very common in social networks. Clusters are groups of friends who know each other. This is related to the idea of "friend of your friend is likely to be your friend".[29] The degree by which nodes are able to be clustered together can be measured by the clustering coefficient. In general, the clustering coefficient *C* is based on the number of closed triples in a network (i.e., a set of three nodes connected to each other, "triangles"), and it can be calculated by the following equation:[30]

$$C = \frac{3\ \times \mathit{number\ of\ triangles}}{\mathit{number\ of\ connected\ triples\ of\ vertices}} \tag{1}$$

For example, the clustering coefficient *C* for the network below can be measured as follows:

$$C = \frac{3 \text{ x } 1}{5} = \frac{3}{5}$$



Power law degree distribution: The degree of a node represents the number of edges connected to that node.[5] A distribution function P(K) gives the probability that a selected node at random has a degree K.[5] Plotting the P(K) function for a network generates a histogram of degree distribution of nodes similar to the one shown in Figure 2. Note that the distribution has a long right tail as shown in Figure 3. The long right tail indicates that in social networks, most nodes have a low degree, whereas a small proportion of nodes known as "hubs" have a high degree. And this is fairly true for social networks. Many studies[31–34] showed that OSNs follow the power law degree distribution.

Some of the above properties are common in complex networks, for example, large scale and power law distribution. These OSNs properties are also used to analyze issues pertaining to social influence. Another important property of OSNs that is relevant to social influence analysis as well is the ability to retrieve OSNs data

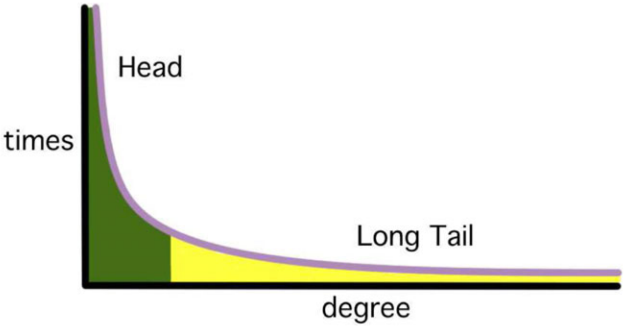


Figure 2. Histogram of degree distribution of nodes.

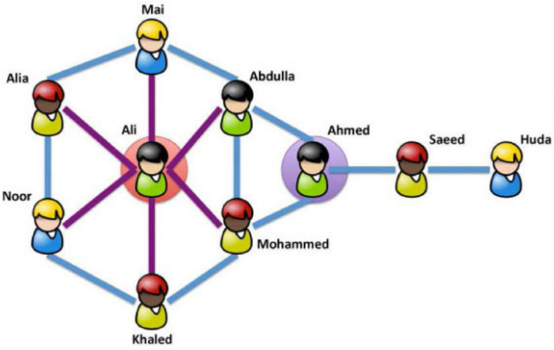


Figure 3. Degree centrality.

easily through APIs, while in real world social networks, a substantial physical effort is needed to collect these data. Application Programming Interface (APIs) represent a set of procedures that help in accomplishing a task or help in interacting with different software components. Many of OSNs APIs are based on SOAP and REST services. Such APIs contain functions for remote access that allow an easier retrieval of information. The opportunity of data extraction and analysis in OSNs encouraged an increased research affluence to model OSNs and their social influence strengths.

2.3. Basic Measurements of Influence Strength

Social networks are modeled as graphs $G = (V, E)$, where V is the set of nodes in the network and E is the set of edges. The nodes are related to the users and the edges represent the relationships between these users in the network. Influence strength can be related to a node or an edge in the network. For example, some nodes in the network might have higher influence than other nodes. Let us say that a node A has high influence and higher edge strength on node B; this high influence will make node B behave similar to node A. Next we present the basic measures of this strength on edge and node levels.

2.3.1. Edge Strength

Edge or tie strength concept was introduced by Granovetter.[35] On edge level there are two different types of ties, strong ties and weak ties. The tie strength depends on the number of overlapping friends or neighbors between two nodes.[35] The larger the overlap the stronger the ties between the nodes. The strength between two nodes A and B can be defined in terms of Jaccard coefficient as follows:[36]

$$S(A, B) = \frac{| n_A \cap n_B |}{| n_A \cup n_B |} \tag{2}$$

where n_A and n_B are the neighbors of nodes A and B, respectively. There are other measurements to determine the tie strength such as embeddedness discussed in Ref. 37 Strong ties represent trust relationship between nodes or simply friendship, whereas Weak ties occur between acquaintances when the friendship overlap is small and restricted information is shared between the nodes such as private profiles, hidden personal details and private posts.

2.3.2. Node Strength

The node importance in OSN is measured through centrality. Nodes with high centrality have higher influence in the network than the nodes with less centrality power. Here, we distinguish three levels of centrality: degree, betweenness, and closeness.

Degree centrality is the number of ties that a node has.[7] In Figure 3, node Ali has the highest degree centrality, because it is the node with the highest number of ties or edges. This means that he is quite active in the network. However, he is not necessarily the most influential person because he is only directly connected within one degree to people in his clique—he has to go through Ahmed to get to other cliques.

Betweenness centrality occurs when a node falls in a favored position between two cliques in the network.[7] In Figure 4, Ahmed has the highest betweenness because he is between Abdulla, Mohammed, and Saeed, who are between other nodes. Abdulla, Mohammed, and Saeed have a lower betweenness because they are essentially within their own cliques. So, Ahmed has potentially more influence in the network. Betweenness represents a single point of failure—when the node with the highest betweenness centrality is removed from the network, the ties between cliques separate apart.

Closeness centrality measures how quickly a node can access more nodes in a network.[7] In Figure 5, Abdulla and Mohammed have the highest closeness centrality because they can reach more entities through shorter paths.

2.4. Social Influence Analysis

There are different considerations for modeling influence in social networks. Edge and node strengths are typical attributes used to analyze influence in social

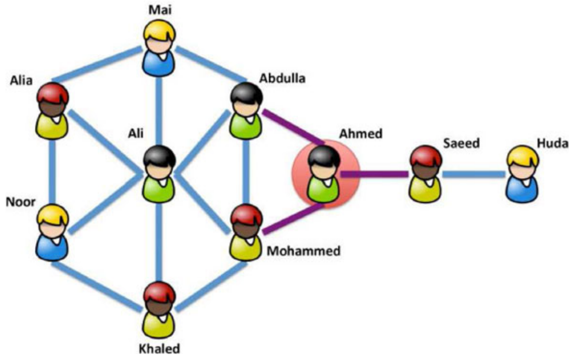


Figure 4. Betweenness centrality.

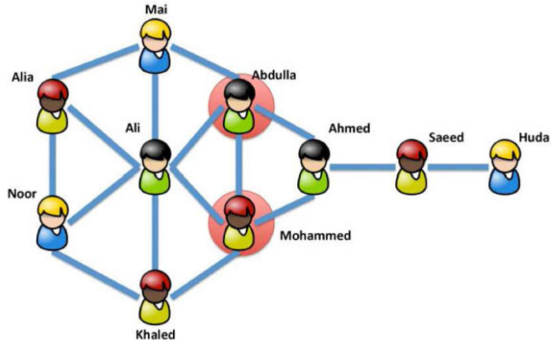


Figure 5. Closeness centrality.

networks. In addition, the followings are additional analytical considerations relevant to social influence analysis:

- ***Multitopics:*** social influence will have different effects on different topics discussed in the social network. For example, assume two neighbors *A*, specialized in data mining and *B*, specialized in programming. *A* will have high influence on *B* when the topic is related to data mining while *B* will have higher influence on *A* when the topic is related to programming.
- ***User actions:*** considering user actions and past behaviors while measuring influence.
- ***Scalability:*** the number of nodes in OSNs increases rapidly. Therefore, there is a need to develop methods that scale well with large data sets.[23]

As far as we know, a limited research has been proposed to study and compare modeling techniques to contrast their limitations and challenges. In our paper, we will address these social influence related contrasts and study them based on performance metrics which we will introduce later. We also propose a categorization of these models to classify them following some common features to distinguish their tradeoffs in a single snapshot.

3. SOCIAL INFLUENCE STRUCTURES AND MODELS

Some theoretical and empirical works have been performed to perceive users' behavior when correlated to their friends' attributes in a social network. Backstorm et al.[22] observed the process of joining an online community and they noticed a correlation between a user joining an online community and the number of friends who are in that community. In another study, Marlow et al.[38] observed the tag usage in Flickr. They noticed a correlation between the tags assigned by a user and those assigned by his friends in his social network. These works provided evidence of influence between users and their friends in OSNs.

The diffusion of influence can be modeled through probabilistic frameworks.[39] While a behavior is spreading through social network users, we need to estimate the probability that a particular individual will embrace the new behavior, given that k of his/her neighbors in the social network have done so. Neighbors refer to people who have a direct edge or tie between them in OSNs. At any point in time t, users would be "adopters" or "nonadopters" of the behavior based on whether they adopt the new behavior at that time.[39]

The properties of social networks enable evaluation of probabilities of users' behaviors in social networks, especially when those behaviors are spread over large populations. For example, the probability of a person to purchase a product given that k of his or her friends recommended that product.[19] Another example would be the probability of joining an online community as a function of the number k of neighbors belonging to the community.[38,40]

If we have a social network with an intention to influence the individual users of this network as we want to introduce a new product, then a viral marketing strategy could start by targeting the most influential users in the network. This will generate a chain reaction of influence-driven advertisement campaign. By applying this method, reaching a very large portion of the network would occur with very small marketing costs.

3.1. Social Influence Structure

The problem of influence maximization can be expressed as follows: "given a network with influence estimates, how to select an initial set of k users such that they eventually influence the largest number of users in the social network".[41]

This influence problem can be formally stated as follows: given a social graph that is undirected $G = (V, E, T)$, where V represents the set of users in the network, E is the set of edges in the network, and T is the matrix of time stamps at which the social ties were created. The social ties in that matrix represent the links and relationships between the nodes in the social network. A tie between users u and v is represented by an undirected edge $(u, v) \in E$. Each edge is labeled with a time stamp at which the edge was created. Assuming that social ties are never broken,[41] the labeling function can be represented by $T : E \rightarrow N$.

A log of actions is maintained where an action could be joining an online community or purchasing a product. This is formulated as *Actions(User, Action, Time)* where a tuple (u, a, t_u) indicates that user u has performed action a at a time

t_u. The log contains all the actions performed by all users in V of the social graph G. Let A represent the actions set, A_u represents the number of actions performed by user u, and $A_{u\&v}$ is the number of actions performed by both users u and v while $A_{u|v}$ represents the number of actions that either u or v performed. This can be shown through the following formula $A_{u|v} = A_u + A_v - A_{u\&v}$. We also use A_{u2v} to denote the number of actions propagated from u to v.[41] Definition 1 formally introduces the action propagation between users in graph G.

DEFINITION 1 (Action propagation). *We say that an action $a \in A$ propagates from user u to v if (i) $(u, v) \in E$; (ii) $\exists(u, a, t_i), (v, a, t_j) \in Actions\,(V, A, T)$ with $t_i < t_j$; and (iii) $T(u, v) \leq t_i$. When this happens, we state the predicate $prop(a, u, v, \triangle t)$ where $\triangle t = t_j - t_i$.*

Definition 2 shows the propagation graph[41] of each action. This leads to a natural notion of a propagation graph, defined next.

DEFINITION 2 (Propagation graph). *For each action a, we define a propagation graph $PG(a) = (V(a), E(a))$ as follows: $V(a) = \{v \,|\, \exists t\,(u, a, t) \in Actions(V, A, T)\}$; there is a directed edge $u \xrightarrow{\triangle t} v$ in E(a) whenever $prop(a, u, v, \triangle t)$.*

The propagation graph of an action is a directed graph, which contains all the users who performed that action, with the edges connecting them according to the direction of propagation.

3.2. Social Influence Models

Although many approaches have been proposed to address the problem of measuring influence probability, there are limited works to contrast their strength and limitations. Sun and Tang[36] introduced a research survey of social influence analysis models and algorithms for measuring social influence. They discussed influence maximization and its application in viral marketing. They focused on the computational aspect of social influence analysis by evaluating the selection of people who are similar to each other (for example, two users who have the same opinion). They also assessed the influence that leads users to adopt behaviors experienced by their neighbors (for example, changing the opinion of a user to agree with one of his neighbors). In addition, they provided methods to measure the weight of influence. Our survey approach categorizes social influence models into four broader categories: 1) static models, 2) dynamic models, 3) diffusion models, and 4) models based on users behaviors. We also contrast different influence models stating their strength and limitations.

OSNs are still new and are still not fully analyzed. OSN models should represent and satisfy some inherent properties introduced in Section 2.2. As a result, modeling social influence studies in OSNs are still in their infancy. There are no standard models for representing influence, which leads to difficulties in analyzing large-scale networks based on social influence. In this section, we study different models of social influence and compare the results of each model to determine the most accurate way to measure the probability ($p_{u,v}$) with which a node u is influenced

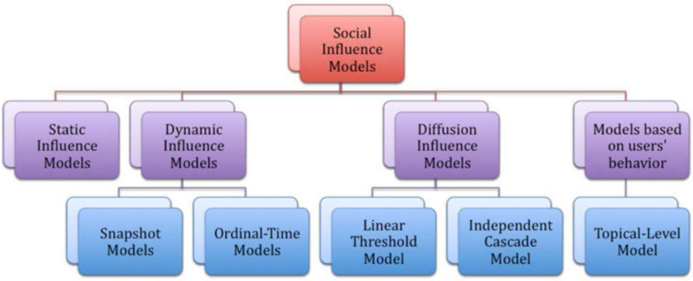


Figure 6. Social influence models.

by its neighbor v. We discuss the strengths and limitations as well as the different challenges the models might have.

Generally, there are two basic categories to represent influence models in social networks. Static influence models are the simplest and easy to assess. In these models, it is assumed that the probability of influence is static and time independent. Only the current state of the network and the most influential nodes at that state are considered. The second category of models is labelled as dynamic influence models, which assume that the influence changes over time. We will see later that the models in this category are the most accurate as they can tell the history of a specific network and identify the most influential nodes for diffusing a behavior/information, but they are very expensive when tested on large data sets as they take long time to execute on large social networks.

Other categories of social network models discussed in this paper are categorized as linear threshold models and independent cascade models. Other models based on greedy algorithms and past user behaviors such as topical affinity propagation models are also addressed in the comparative survey of this paper. Figure 6 shows a hierarchical view of the social influence models which we are discussed in this paper.

The challenge researchers might face is how to compare models that are of different categorization and state the relationship between them. Especially when the relationship between these models is ambiguous. Thus, we aim to clarify this ambiguity by explaining and finding commonalities or categories across these model. Next, we briefly introduce each of these categories.

3.2.1. Static Influence Models

Static influence models are independent of time and used to capture the most influential nodes presently. Therefore, the network size is fixed. One instance of this model is based on ***Bernoulli distribution***. In these social influence models, a specific node u has a fixed probability to influence its inactive neighbor v. If it activates the neighbor, then this is a successful attempt and otherwise, failure. Each attempt can be shown as a Bernoulli trial. Figure 7 shows a sample illustration to explain Bernoulli trials. The influence probability can be estimated using maximum

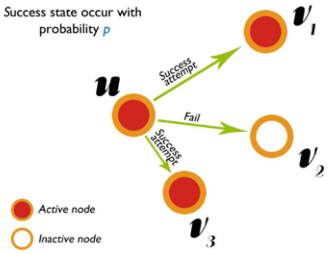


Figure 7. Bernoulli distribution: Node u will have fixed probability to influence its inactive neighbor v_1, v_2 and v_3. If u attempt is “successful”, node v will be activated otherwise node v will remain inactive.

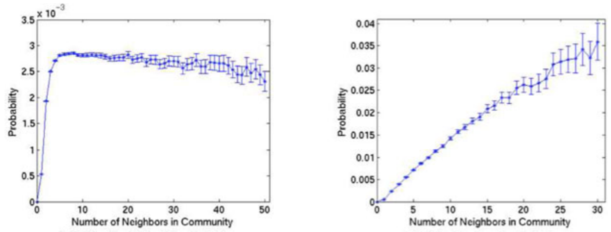
likelihood estimator (MLE)[41] as the ratio of successful attempts over the total number of trials:

$$p_{u,v} = \frac{A_{v2u}}{A_v} \tag{3}$$

3.2.2. Dynamic Influence Models

In real life, influence changes over time and may not stay static. For example, users’ opinions could change over time. When a user is influenced by its neighbors to join a community, she/he is initially excited to join that community, but over time that user might have less excitement to stay in the community. To represent dynamic influence models, we discuss two models of social influence. The first one is based on capturing a small set of “Snapshot” observations of the social network and the second one is based on detailed temporal dynamics. These two models can be represented as a function of the number k of neighbors who have adopted a new behavior.[39] The individual become $k - exposed$ to the behavior at specific time t if it is a nonadopter at time t but surrounded with exactly k neighbors who are all adopters at time t.

a) Snapshot model: To represent this model, we need to consider two snapshots of the social network at different points in time.[39] Consider then the set of all individuals who are $k - exposed$ in the first snapshot. Let $p_s(k)$ be the fraction of individuals in this set who have become adopters by the time of the second snapshot.[39] To further clarify, imagine that all $k - exposed$ nodes in the first snapshot will flip a coin of fixed bias $p_s(k)$ to decide whether to adopt the behavior or not. On the basis of different experiments on Wikipedia (a free, web based, collaborative, multilingual encyclopedia), LiveJournal (a virtual community where Internet users can keep a blog, journal, or diary), and engaging in email correspondence,[22,39,42] the snapshot curve (shown in Figure 8 b) shows that the influence increases with more links, but the marginal influence of each additional link is slowly decreasing.[39] There are studies that used snapshot models to compute the influence probabilities such as Refs. 22, 40, 42 though they used a large number of snapshots requiring substantial computational sources.



(a) *Wikipedia: Ordinal-Time Definition*

(b) *Wikipedia: Snapshot Definition*

Figure 8. The probability of editing an article in Wikipedia.[39]

b) Ordinal-time model: To represent this model, we need to consider a time sequence of a social network as it evolves over the time. A new link is created in the network or a new individual adopts a new behavior. For each k, consider the set of all individuals who were ever $k - exposed$ at any time, and define $p_0(k)$ to be the fraction of this set that became adopters before acquiring a $(k + 1)^{st}$ neighbor who is an adopter.[39]. To clarify, imagine that a nonadapter acquired the kth neighbor who is adopter, by flipping a coin of fixed bias $p_0(k)$, the nonadopter will decide to adopt or not. The curve of ordinal time in Figure 8a shows that the first five links have greater impact, but after some propagation, subsequent links impact stabilizes. This feature is similar to the power of low distribution. In both of the above cases, there is a need to determine the maximum-likelihood values of probabilities $p_0(k)$ and $p_s(k)$.

Comparing different models and their relationships would reveal interesting performance thresholds and application domains. Generally, the snapshot model is widely used as it is more applicable to capture an observation of the network without the need for performing moment-by-moment measurements.[39] Although there is no apparent relationship between the snapshot and the ordinal-time models, the shape of ordinal time can be approximated from data in a single snapshot. Experimental analysis show that accurate result occurred with more snapshots.

3.2.3. Diffusion Influence Models

These models are used when adopting behavior depends on knowing the number of neighbors who adopted the same behavior. In Refs. 17, 43 Domingos and Richardson proposed a framework for the propagation of influence when addressing the problem of identifying influential users. They proposed a probabilistic model of interaction and heuristics to select the influential users in the context of viral marketing, and confirmed their approach through an empirical study. Their idea is based on how to find the most influential individuals and target them to advertise a new innovation or a product. In a large cascade, they will influence their friends, and friends of friends. Market customers are represented as nodes in social networks and customers influence is modelled as a Markov random field. These diffusion models can be used to optimize marketing decisions. Kempe et al.[18] revealed that in general the problem of selecting influential sets of individuals is NP-complete. This means

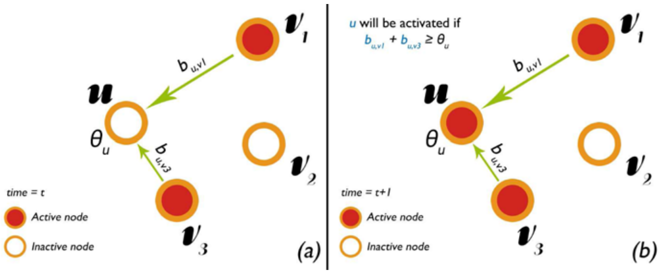


Figure 9. Example of linear threshold model process.

that a solution for such problems can be found if we can find subsets that individually can be solved to provide a good solution to the whole problem. This set of individuals should be chosen to generate the maximum influence during the influence-diffusion process. Approximation algorithms are used to solve the problem of influence maximization. In some influence models, the greedy algorithm will select the set of individuals with approximation $(1 - 1/e - \varepsilon)$,[18] where e is the base of the natural logarithm and ε is any positive real number. In their work, Kempe et al. focused on two influence-diffusion models: linear threshold model and independent cascade model.

a) Linear threshold model. Granovetter and Schelling[44] were among the first to propose the threshold approach to capture influence. In linear threshold model, a weight $b_{u,v}$ is used to measure the tendency of a node u to be influenced by each neighbor v such that $\sum_{v\, neighbor\, of\, u} b_{u,v} \leq 1$. Starting with the initial set of active nodes A_0, the influence propagation resumes as follows: each node u is assigned a threshold θ_u randomly from the interval [0, 1]; the threshold represents the weight fraction of u's neighbors that must adopt the behavior (be active) in order for u to become active and adopt the same behavior. At time stamp t, all nodes that were active in time $t - 1$ remain active, and any node u for which the total weight of its active neighbors is at least θ_u gets activated; where:

$$\sum_{v\ active\, neighbor\, of\ u} b_{u,v} \geq \theta_u \tag{4}$$

The thresholds θ_u represent the tendency of nodes to adopt the new behavior when their neighbors do.[18] Figures 9a and 9b show an example of the process involved in linear threshold model.

In their experiment, Kempe et al.[18] compared their greedy algorithm with nodes' degrees and centrality within the network, as well as incorporating random nodes. On the basis of their experiment, their greedy algorithm outperforms the degree and distance centralities because these two features do not consider the dynamics of social networks and focus on the structure of the network to emphasize influence. Random nodes do not generate good results in linear threshold model. Figure 10 shows the result of these experiments.

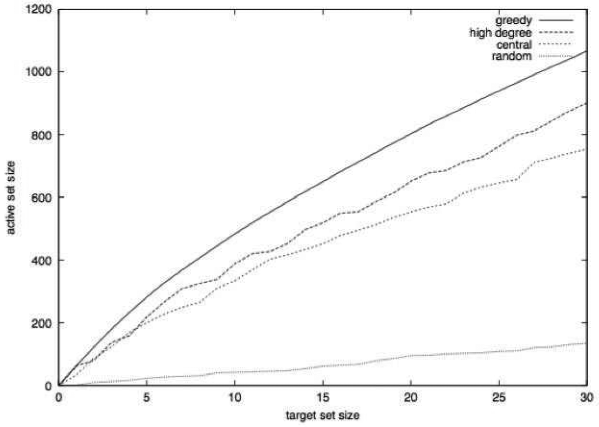


Figure 10. Results for the linear threshold model[18].

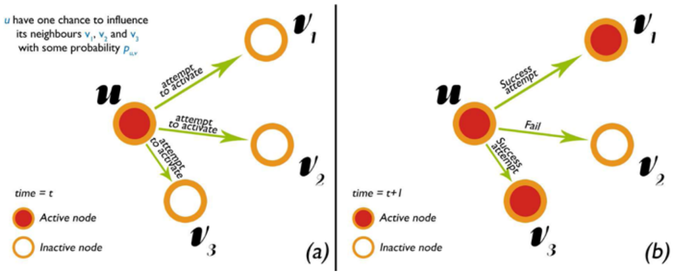


Figure 11. Example of independent cascade model process.

b) Independent cascade model. An independent cascade model starts with an initial set of active nodes A_0. This set of individuals should be chosen the generate the maximum influence during the cascade diffusion process. The process occurs in discrete steps as follows: when node u becomes active for the first time in time step t, its provided with one chance to activate each of its currently inactive neighbor v; in that case u is called *contagious, which* means that it has the ability to affect other nodes as shown in Figure 11a. Node u succeeds to influence its neighbor v with a probability $p_{u,v}$ independent of past history. If u succeeds, then v will become active in time step $t + 1$ as shown in Figure 11b; but whether or not u succeeds, it cannot make any further attempts to activate v in future rounds.[18] The same process continues until us communicate with all neighbors for influence attempts and there are no more contagious nodes.

On the basis of Kempe et al. experiments on independent cascade model, the greedy algorithm still outperforms degree and centrality methods within the network. Interestingly, random nodes performed well on independent cascade models, as shown in Figure 12.

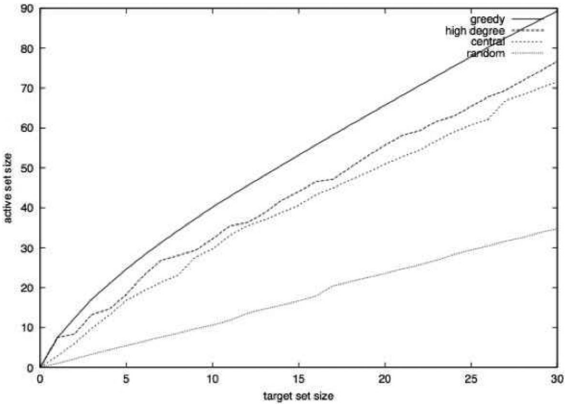


Figure 12. Results of independent cascade model.[18]

3.2.4. Models of Influence based on Users' Behavior

The models discussed above are based on the influence model proposed in Ref. 18, where the influence probabilities are provided in advance as input. Other models proposed in the literature compute the probabilities through mining past users' behavior. In doing so, Tang et al.[23] studied topic-based social influence. In these social networks, discussion topics are distributed across users. The problem is then to find topic-specific subnetworks, and topic-specific influence weights between members of the subnetworks. Then they propose a graphical probabilistic model called topical factor graph (TFG) to unify the information in one probabilistic model. Then they proposed the topical affinity propagation (TAP) model which uses TFG to infer the influence graph. They also dealt with the efficiency problem by devising a distributed implementation of TAP.

Saito et al.[45] have studied the problem of building influence from past users' actions. They focused on the independent cascade model of influence. They formally defined the likelihood maximization problem and then applied Expectation Maximization (EM) algorithm to solve it. Their formulation dose not, however, scale to huge data sets like in social networks. This is due to the fact that in each iteration, *EM* algorithm must update the influence probability.

3.2.5. Other Influence Models

There are many influence models that are based on greedy algorithms. Nemhauser et al.[46] show a greedy approximation algorithm to address the problem of finding a maximal set of individuals. Kempe et al.[18] also proposed a greedy algorithm, but it suffered from the efficiency problem because their proposed model needs to execute Monte Carlo simulation several times until it provides accurate results, which leads to very long computational times. There are studies to improve the efficiency of the greedy algorithms to maximize the influence, such as Refs. 47, 48.

Leskovec et al.[47] studied the influence problem from a different perspective. The main question in their study was: how to select nodes in a network to detect the spread of virus as soon as possible?; this was called outbreak detection. They developed an efficient algorithm based on "lazy-forward" optimization. The algorithm was optimal and 700 times faster than the simple greedy algorithm; but the approach still faces problems related to scalability. In Ref. 48, Chen et al. improved the efficiency of the greedy algorithm.

4. COMPARISON AND IMPLICATIONS

In this section, we compare the influence models discussed in the previous section following the snapshot compilation shown in Table I. Static influence models are based on capturing influence in the current moment. They are time-independent models that do not diffuse over time. They assume that influence probabilities are fixed (static) and do not change over time. Different techniques are used in static influence models; one of them is Bernoulli distribution. Static influence models are easy to apply and test, which makes them one of the easiest ways to measure influence in a network. But since social networks are dynamic, where new links are built/removed regularly, the assumption of static influence models will not make them the best choice to measure influence in social networks. Dynamic influence models were introduced to address static influence probability deficiencies. On the basis of these dynamic models, influence probability changes over time. Snapshot and ordinal-time models are instances of dynamic influence models that are time dependent. Snapshot models take different snapshots of the networks and generate an observation about the network. The snapshot technique is widely used to model social networks because it can capture large-scale data. To get a better observation of the network, we need to take many snapshots for large data sets, which is time consuming and needs a lot of space. Ordinal-time models provide detailed temporal dynamics of the network. These models provide more accurate results since they measure influence moment-by-moment. There is no direct implementation of ordinal-time models on large data sets (such as large social networks), which makes it difficult to draw a conclusion about these models on social networks. Diffusion influence models such as linear threshold and independent cascade models were introduced to address the issue of influence propagation. In linear threshold, every node contributes with a certain weight to its adopting neighbors. If the sum of these weights is greater than a given threshold, the node becomes an adopter too. The weight depends on the edge strength between the node and its neighbors. Using the weight as a measurement between the node and its neighbors will show the strength of the influence. Independent cascade models use cascade processes to measure influence propagation. Each node has two states: to adopt or not to adopt. The adopters will have influence on their neighbors and the adopting neighbors will have influence on their neighbors too and so on, the influence spreads over the network. Each adopting node has one chance to influence its neighbor to adopt the same behavior with some probability that depends on the edge strength between the nodes. These models have the advantage of fast spreading an information/behavior

Table I. Comparison of social influence models.

Models		Based on	Techniques	Algorithms	Strength	Limitation
Static influence models[41]		Capturing the most influential node in the current time	Bernoulli distribution	MLE	Easy to apply and test	Do not conceder the dynamic of the social network
Dynamic influence models	Snapshot[22,40,42]	Taking various number of snapshots for the network	Snapshots	Observations of the social network	– Can capture large-scale data and provide them for analysis – More applicable and easily handled	The need to take many snapshots for large size data
	Ordinal-time	Detailed temporal dynamics of the network	Moment-by-moment measurement	EM	Provide more accurate influence results	No direct implementation of the ordinal-time definition on large-scale data
Diffusion influence models	Linear threshold[17,18,43]	Choosing a threshold at random.	Threshold	– Approximation algorithms – greedy algorithm	– Adopting new behavior based on fraction weight	– Does not consider the correlation between users' actions. – Ignores the attributes that are associated with each user node
	Independent cascade model[18]	Activating nodes based on discrete steps	Cascading processes	– Approximation algorithms – greedy algorithm	– Fast spreading of behaviors	– Does not consider the correlation between users' actions. – Ignores the attributes that are associated with each user node
Models of influence based on users' behavior	Topical Affinity Propagation (TAP)[23]	Computing the probabilities through mining the users' behavior	Graphical probabilistic model	– Affinity propagation algorithm. – Distributed learning algorithms	Devise a distributed learning algorithm under the Map-reduce programming model (deals with the efficiency problem)	TFG model can not capture the social influence between users while building the unified probabilistic model

across the nodes of a network, especially when determining the initial optimal set of the most influential nodes in the network. Both linear threshold and independent cascade models have the same limitations since they both ignore the attributes that are associated with each node and do not consider the correlation between user actions.

Other models based on user behavior were introduced to measure influence. The TAP model uses TFG to build the influence probability model based on the user's topics. This model employs a distributed learning algorithm to deal with the efficiency problem, but it cannot capture the social influence between users while building the unified probabilistic model. We also discussed models that are based on greedy algorithms and we found that they outperform influence measures that are based on the structure of the social network such as degree and distance centralities. But on the other hand, their efficiency is low since they take long times to execute repeated tests to provide accurate results.

5. EXPERIMENTS AND RESULTS

In this section, we will compare the influence diffusion in OSNs based on two famous influence propagation models that are the linear threshold model and the independent cascade model. We implemented the two algorithms using Matlab. Then we performed the experiments using Apple iMac with Mac OS X version 10.6.8, processor 2.66 GHz intel Core i5, and 4GB memory.

We applied the experiment on two actual OSNs. The first social network is Flickr, which is a photo-sharing social network. On Flickr, users can share and embed photographs on their own blogs. The data set of Flickr network consists of 2,570,535 nodes and 33,140,018 links between the nodes. We used the data set provided by Cha et al. in Ref. 49 We selected 500 nodes to run our experiments. Nodes are associated with their actions to favor a photo.

The second social network we used as a testbed for our experiments is Last.fm[a] social network, which is a popular Internet radio to stream music. The data set provided by Ref. 50 contains 1892 users who assigned tags to artists during different time stamps. A tag could be any word related to the artist like rock, POP, sad, and touching, etc. The time stamp shows when the tag assignments were done. We selected 99 nodes to apply our experiment. Each user is associated with his/her action of tagging an artist.

To compare the two influence propagation models, we used four methods to assign edges probabilities in the social graph:

1. Jaccard coefficient based on common actions: in this method, we calculate similarity between two nodes based on the common actions they have. The formula used is $JC_{u,v} = \frac{A_{u,v}}{A_u+A_v-A_{u,v}}$, where A_u is the number of actions performed by node u, A_v is the number of actions performed by node v, and $A_{u,v}$ is the number of common actions performed by nodes u and v.

[a] http://www.lastfm.com

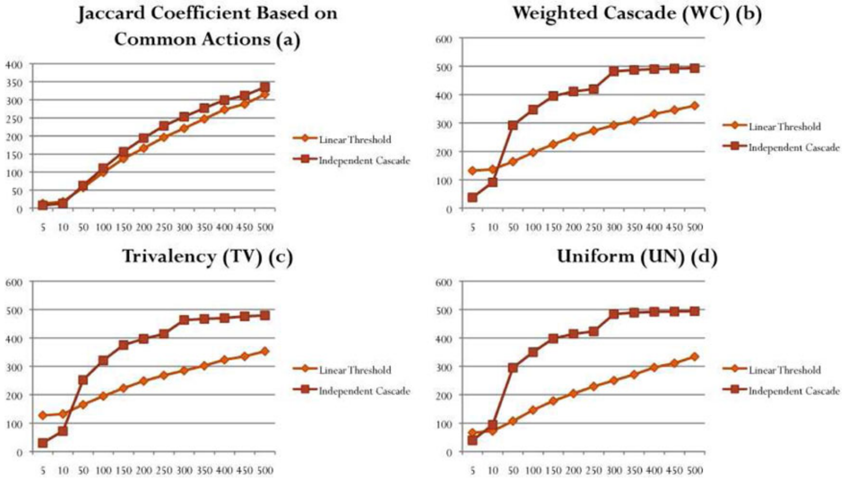


Figure 13. Comparing influence spread For Flickr social network.

2. Weighted cascade: which is a special case of the independent cascade model, where each edge from node u to v is assigned a probability $\frac{1}{d_v}$ of activating v.[18]
3. Trivalency (TV): where edge probabilities are selected uniformly at random from the set $\{0.1, 0.01, 0.001\}$
4. Uniform (UN): where all edges have the same probability (e.g., $p = 0.01$)

Using Flickr social network, we notice that independent cascade model outperforms the linear threshold model in all probability assignment methods when the seed set size becomes larger. The common actions (Figure 13a) probability assignment methods is more steady and both propagation models provide similar curves although it dose not activate as much nodes as the other methods shown in Figure13b–13d. The results could be different for different sittings applied to run the algorithms such as the threshold value or the number of nodes in the seed set.

Using Last.fm social network, we applied the same probability assignment methods used above on the actual network data sets. In these experiments, we notice that in the trivalency and uniform methods, the independent cascade model outperforms the linear threshold model by activating more nodes during the propagating process Figure 14.

6. CONCLUSIONS

In this paper, we defined social influence and stated its importance in evolving social networks. We introduced some analytics used when measuring centrality in social networks such as centrality measurements. We also surveyed measure models, which address the objective of influence maximization in social networks. We stated the strength and limitation of each model through a comparative study. We compared two propagation models empirically to state which is better for social

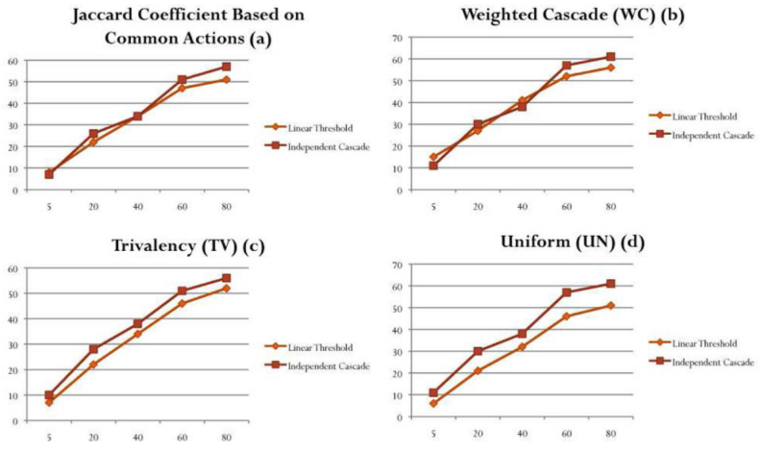


Figure 14. Comparing influence spread For Last.fm.

influence diffusion. We also revealed new research directions in social networks mining toward the prospects of further analyzing social influence.

There are many possible future directions to extend these social influence models to address perceived limitations such as scalability and efficiency. A combination of these models may generate more accurate results and help addressing the individual limitations.

References

1. Facebook. Facebook statistics. Available at http://www.facebook.com/press/info.php?statistics; 2011. Last Accessed: 10-Nov-2011.
2. Acar A, Muraki Y. Twitter for crisis communication: Lessons learned from Japan's tsunami disaster. Int J Web Based Communities 2011;7(3):392–402.
3. Dumenco S. A very brief (cartoon) history of social influence. Ad Age Digital, Available at http://bit.ly/hmXCif; 2011. Last Accessed: 10-Nov-2011.
4. Solis B. The interest graph on twitter is alive: Studying starbucks top followers. Available at http://bit.ly/geE0i4; 2011. Last Accessed: 10-Nov-2011.
5. Newth D. Complex science for a complex world: Exploring human ecosystems with agents, Chapter (5). The structure of social networks. Canberra, Australia: ANU Press; 2006.
6. Scott J. Social network analysis: A handbook. SAGE; 2000. SAGE Publications, Thousand Oaks, California.
7. Hanneman RA, Riddle M. Introduction to social network methods, chapter (10). Centrality and Power. Riverside, CA: Department of Sociology at the University of California; 2005.
8. Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S. Feedback effects between similarity and social influence in on-line communities. In: Proc 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining; 2008. pp. 160–168.
9. Wikipedia. Social psychology. Available at http://wikipedia.org/wiki/Social_Psychology; 2011. Last Accessed: 28-Dec-2011.
10. Triplett N. The dynamogenic factors in pacemaking and competition. Am J Psychol 1898;9(4):507–533.
11. Festinger L. A theory of cognitive dissonance. Palo Alto, CA: Stanford University Press; 1957.

12. Rashotte L. Social influence. In Blackwell Encyclopedia of Sociology. Blackwell Publishing, Malden, Massachusetts, pp. 4426–4429, 2007.
13. Rogers E. Diffusion of Innovations. Free Press; 4 edition; 1995. New York, NY.
14. Strang D, Soule S. Diffusion in organizations and social movements: From hybrid corn to poison pills. Ann Rev Sociol 1998;24:265–290.
15. Blume L. The statistical mechanics of strategic interaction. Game Econ Behav 1993;5:387–424.
16. Young HP. Individual strategy and social structure: An evolutionary theory of institutions. Princeton University Press; 1998. Princeton, New Jersey.
17. Domingos P, Richardson M. Mining the network value of customers. In Proc Seventh ACM SIGKDD Int Conf Knowledge Discovery and Data Mining; 2001. pp 57–66.
18. Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence through a social network. In Proc Ninth ACM SIGKDD Int Conf Knowledge Discovery and Data Mining; 2003. pp 137–146.
19. Leskovec J, Adamic L, Huberman B. The dynamics of viral marketing. In Proc Seventh ACM Conf on Electronic Commerce; 2007. pp. 228–237.
20. Gruhl D, Liben-Nowell D, Guha RV, Tomkins A. Information diffusion through blogspace. In Proc. 13th Int World Wide Web Conf; 2004. pp 491–501.
21. Leskovec J, McGlohon M, Faloutsos C, Glance N, Hurst M. Cascading behavior in large blog graphs. In Proc of the Seventh SIAM Int Conf Data Mining; 2007.
22. Backstrom L, Huttenlocher D, Kleinberg J, Lan X. Group formation in large social networks: Membership, growth, and evolution. In Proc 12th ACM SIGKDD Int Conf Knowledge Discovery and Data Mining; 2006.
23. Tang J, Sun J, Wang C, Yang Z. Social influence analysis in large-scale networks. In Proc 15th ACM SIGKDD Int Conf Knowledge Discovery and Data Mining (KDD ’09); 2009. pp 807–816.
24. Lazarsfeld P, Merton RK. Friendship as a social process: A substantive and methodological analysis. Freedom and control in modern society; Van Nostrand, New York, pp 18–66, 1954.
25. McPherson M, Smith-Lovin1 L, Cook JM. Birds of a feather: Homophily in social networks. Ann Rev Sociol 2001;27:415–444.
26. Young HP. The diffusion of innovations in social networks. In The economy as a complex evolving system. Oxford University Press, Oxford, New York, N.Y. volume III, 2003.
27. Anagnostopoulos A, Kumar R, Mahdian M. Influence and correlation in social networks. Proc 14th ACM SIGKDD Int Conf Knowledge Discovery and Data Mining (KDD ’08); 2008. pp 7–15.
28. Linkedin. Linkedin press center. Available at http://press.linkedin.com/; 2011.
29. Li C-T. Eight properties of social network (1-4). The Odd World, Available at http://bit.ly/uhRGsf; 2007. Last Accessed: 11-June-2010.
30. Wikipedia. Clustering coefficient, Wikipedia the free encyclopedia. Available at: http://wikipedia.org/wiki/Clustering_coefficient; 2010. Last Accessed 11-Sep-2010.
31. Kumar R, Novak J, Tomkins A. Structure and evolution of on-line social networks. Proc 12th ACM SIGKDD Int Conf Knowledge Discovery and Data Mining; 2006. pp 611–617.
32. Golder S, Wilkinson D, Huberman B. Rhythms of social interaction: Messaging within a massive on-line network. Proc Third Int Conf Communities and Technologies; 2007.
33. Mislove A, Marcon M, Gummadi K, Druschel P, Bhattacharjee B. Measurement and analysis of on-line social networks. Proc Seventh ACM SIGCOMM Conf Internet Measurement; 2007. pp 29–42.
34. Java A, Song X, Finin T, Tseng B. Why we twitter: Understanding microblogging usage and communities. Proc Joint Ninth WEBKDD and First SNA-KDD Workshop; 2007. pp 56–65.
35. Granovetter M. The strength of weak ties. Am J Sociol 1973;78(6):1360–1380.
36. Sun J, Tang J. A Survey of Models and Algorithms for Social Influence Analysis. In Social Network Data Analytics. Springer US, New York NY; pp. 177–214, 2011.
37. Granovetter M. Economic action and social structure: The problem of embeddedness. Am J Sociol 1985;91(3):481–510.

38. Marlow C, Naaman M, Boyd D, Davis M. Ht06, tagging paper, taxonomy, flickr, academic article, toread. Proc Seventeenth Conf Hypertext and Hypermedia HYPERTEXT '06; 2006. pp 31–40.
39. Cosley D, Huttenlocher DP, Kleinberg JM, Lan X, Suri S. Sequential influence models in social networks. Proc Fourth Int Conf Weblogs and Social Media; 2010.
40. Shi X, Zhu J, Cai R, Zhang L. User grouping behavior in online forums. Proc 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining; 2009. pp 777–786.
41. Goyal A, Bonchi F, Lakshmanan LVS. Learning influence probabilities in social networks. Proc Third ACM Int Conf Web Search and Data Mining, WSDM '10; 2010. pp 241–250.
42. Kossinets G, Watts DJ. Empirical analysis of an evolving social network. Science 2006;311(5757):88–90.
43. Richardson M, Domingos P. Mining knowledge-sharing sites for viral marketing. Proc Eighth ACM SIGKDD Int Conf Knowledge Discovery and Data Mining (KDD'02); 2002. pp 61–70.
44. Granovetter M. Threshold models of collective behavior. Am J Sociol 1978;83(6):1420–1443.
45. Saito K, Nakano R, Kimura M. Prediction of information diffusion probabilities for independent cascade model. Proc 12th Int Conf Knowledge-Based Intelligent Information and Engineering Systems, Part III (KES'08); 2008.
46. Nemhauser GL, Wolsey LA, Fisher ML. An analysis of approximations for maximizing submodular set functions. Mathematical Programming 1978;14(1):265–294.
47. Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance NS. Cost-effective outbreak detection in networks. Proc 13th ACM Int Conf Knowledge Discovery and Data Mining (KDD'07); 2007. pp 420–429.
48. Chen W, Wang Y, Yang S. Efficient influence maximization in social networks. Proc 15th ACM Int Conf on Knowledge Discovery and Data Mining (KDD'09); 2009. pp 199–208.
49. Cha M, Mislove A, Gummadi KP. A measurement-driven analysis of information propagation in the flickr social network. Proc 18th Int World Wide Web Conference (WWW'09); 2009.
50. Cantador I, Brusilovsky P, Kuflik T. Second workshop on information heterogeneity and fusion in recommender systems (hetrec 2011). Proc Fifth ACM conference on Recommender systems, RecSys 2011, New York, NY; 2011. ACM.