# Identifying Super-Mediators of Information Diffusion in Social Networks

Kazumi Saito[1], Masahiro Kimura[2], Kouzou Ohara[3], and Hiroshi Motoda[4]

[1] School of Administration and Informatics, University of Shizuoka
k-saito@u-shizuoka-ken.ac.jp
[2] Department of Electronics and Informatics, Ryukoku University
kimura@rins.ryukoku.ac.jp
[3] Department of Integrated Information Technology, Aoyama Gakuin University
ohara@it.aoyama.ac.jp
[4] Institute of Scientific and Industrial Research, Osaka University
motoda@ar.sanken.osaka-u.ac.jp

**Abstract.** We propose a method to discover a different kind of influential nodes in a social network, which we call "super-mediators", *i.e.*, those nodes which play an important role in receiving the information and passing it to other nodes. We mathematically formulate this as a difference maximization problem in the average influence degree with respect to a node removal, *i.e.*, a node that contributes to making the difference large is influential. We further characterize the property of these super-mediators as having both large influence degree, *i.e.*, capable of widely spreading information to other recipient nodes, and large reverse-influence degree, *i.e.*, capable of widely receiving information from other information source nodes. We conducted extensive experiments using three real world social networks and confirmed that this property holds. We further investigated how well the conventional centrality measures capture super-mediators. In short the in-degree centrality is a good measure when the diffusion probability is small and the betweenness centrality is a good measure when the diffusion probability is large, but the super-mediators do depend on the value of the diffusion probability and no single centrality measure works equally well for a wide range of the diffusion probability.

**Keywords:** Information diffusion, super-mediator, influence degree, reverse-influence degree

## 1 Introduction

The emergence of Social Media such as Facebook, Digg and Twitter has provided us with the opportunity to create large social networks, which play a fundamental role in the spread of information, ideas, and influence. Such effects have been observed in real life, when an idea or an action gains sudden widespread popularity through "word-of-mouth" or "viral marketing" effects. This phenomenon has attracted the interest of many researchers from diverse fields [12], such as sociology, psychology, economy, computer science, etc.

A substantial amount of work has been devoted to the task of analyzing and mining information diffusion (*i.e.*, cascading) processes in large social networks [16, 14, 2, 1, 18, 25, 3]. Widely used information diffusion models in these studies are *independent cascade (IC)* [4, 6, 7], *linear threshold (LT)* [26, 27] and their variants [8, 19, 5, 10, 21, 22]. These two models focus on different aspects of information diffusion. IC model is sender-centered (information push) and each active node *independently* influences its inactive neighbors with given diffusion probabilities. LT model is receiver-centered (information pull) and a node is influenced by its active neighbors if their total weight exceeds the threshold for the node. Basically the former models diffusion process of how a disease spreads and the latter models diffusion process of how an opinion or innovation spreads.

The main focus of research using these models over the past decade has been on optimization problems in which the goal is to maximize the spread of information through a given network, either by selecting a good subset of nodes to initiate the cascade [6, 9] or by applying a broader set of intervention strategies such as node and link additions [17, 24]. In particular the former problem is well studied as the *influence maximization problem*, *i.e.*, finding a subset of nodes of size $K$ that maximizes the expected influence degree with $K$ as a parameter. In [23] we proposed a new type of influence maximization problem which we called "Target selection problem" (to avoid confusion, we called the original influence maximization problem as "Source selection problem"). The difference is that the new problem does not assume that the information is guaranteed to start spreading from the selected target nodes. Rather we send the same information from outside of the network to the selected targets as a probabilistic process. This is closer to a situation in which we send a direct mail to selected customers expecting that they spread the received information to others. What we found very interesting is that the nodes selected as the solution of the target selection problem were substantially different from the nodes selected as the solution of the source selection problem, especially in case of LT model. We attributed the difference to the fact that the target nodes must not only be able to be influential, *i.e.*, capable of widely spreading information to other recipients, but also be able to be reverse-influential, *i.e.*, capable of widely receiving the information from other sources. In a separate context we studied another type of influential nodes which we called "super-mediators", *i.e.*, nodes which play an important role in receiving the information and passing it to other nodes [20]. There, we empirically[1] defined the super-mediators to be the nodes that frequently appear in the long information diffusion sequences that start from a node and are shared by many of these sequences that starts from different nodes[2]. The biggest difference of the present work from [20] is that the present work is model-driven while our previous work is data-driven, *i.e.*, [20] does not assume any diffusion model but it requires that abundant observed diffusion sequences are available.

The work in this paper is motivated by these studies. "Source selection problem" only cares the ability of nodes to spread information. "Target selection problem" cares also the ability of the nodes to receive information in addition to the ability to spread the

[1] We call it empirical in the sense that the characterization is qualitative and there is no mathematically defined objective function to be optimized.

[2] We assume that there are many sequences of different length for each starting node.

information, but only in the first step of information diffusion chain. Super-mediators share the same concept as "target selection problem", but they can be any nodes in the chain of information diffusion process. From this observation, we can mathematically define the super-mediators as the solution of an optimization problem and rank them. The influence degree $\sigma(v)$ of a node $v$ is defined to be the expected number of active nodes at the end of diffusion process, *i.e.*, nodes that have become influenced due to information diffusion (see Sec.2 for a more rigorous definition). The average influence degree of the whole network is defined to be the average of $\sigma(v)$ over all nodes in the network. If a node $v$ is a super-mediator, removing this node would substantially decrease the average influence degree. Thus, the importance of each node as a super-mediator can be quantified as the difference of the average influence degree with respect to the node removal.

Here in this paper we use IC model as the information diffusion model and only consider a single node removal, *i.e.*, $K = 1$, but this optimization problem carries the same problem of computational complexity of estimating influence degree[3]. We devised the bond percolation [9] and pruning [8] algorithms to efficiently estimate the influence degree. In this paper, we further improved these techniques and reduced the computation time drastically (but this is not our focus in this paper).

We wanted to characterize the property of the super-mediators returned as the solution of the optimization problem. As mentioned above, there are two important factors: the ability to spread information and the ability to receive information. The former is captured by the influence degree. The latter is captured by the reverse-influence degree, which is a new concept born in this study, *i.e.*, the expected number of initial source nodes from which the information reaches a node at the end of information diffusion. Our hypothesis is that the super-mediators should be ranked high in terms of both of them. We have tested our hypothesis using three real world networks (Enron, Blog and Wikipedia), and confirmed that this property holds. In case of Enron e-mail network, the nodes identified as super-mediators are interpretable in the light of open literature. We further investigated whether the conventional centrality measures can serve as a good measure to identify the super-mediators. What we found is that the super-mediators depend on the value of the diffusion probability and in short the in-degree centrality is a good measure when the diffusion probability is small and the betweenness centrality is a good measure when the diffusion probability is large, and no single centrality measure works equally well for a wide range of the diffusion probability. It can be said that the measure we proposed in this paper is a new centrality that can be added to the existing pool, but the difference is that this measure explicitly considers information diffusion process while the existing centrality considers only network structure.

The paper is organized as follows. Section 2 gives a brief description of the independent cascade model. Section 3 defines super-mediators and gives an algorithm to find and rank them. Section 4 characterizes the super-mediators and introduces a new concept "reverse influence degree" and gives an efficient way to compute it. Section 5 reports experimental results and shows that the hypothesis we made holds for the three

[3] If we consider $K > 1$, the problem becomes more difficult, but we can still use the sub modular property and the same greedy algorithm as is used in "Source selection problem" with various tactics, *e.g.*, burnout [19].

networks. Section 6 summarizes what has been achieved in this work and addresses the future work.

## 2 Information Diffusion Model

We consider a network represented by a directed graph $G = (V, E)$, where $V$ and $E$ ($\subset V \times V$) are the sets of all the nodes and links, respectively. Below we revisit the definition of IC model according to the literatures [6, 11]. The diffusion process proceeds from an initial active node in discrete time-step $t \geq 0$, and it is assumed that nodes can switch their states only from inactive to active (*i.e.*, the SIR setting, see Section 3).

IC model has a *diffusion probability* $p_{u,v}$ with $0 < p_{u,v} < 1$ for each link $(u, v)$ as a parameter. Suppose that a node $u$ first becomes active at time-step $t$, it is given a single chance to activate each currently inactive child node $v$, and succeeds with probability $p_{u,v}$. If $u$ succeeds, then $v$ will become active at time-step $t + 1$. If multiple parent nodes of $v$ first become active at time-step $t$, then their activation trials are sequenced in an arbitrary order, but all performed at time-step $t$. Whether $u$ succeeds or not, it cannot make any further trials to activate $v$ in subsequent rounds. The process terminates if no more activations are possible.

For an initial active node $v \in V$, let $\varphi(v; G)$ denote the number of active nodes at the end of the random diffusion process. It is noted that $\varphi(v; G)$ is a random variable. We denote the expected value of $\varphi(v; G)$ by $\sigma(v; G)$, and call it the *influence degree of* $v$.

## 3 Identifying Super-Mediators

As stated earlier, we conjecture that if a node $w$ is a super-mediator, removing this node would substantially decrease the average influence degree. In order to mathematically formulate this notion, we first define the following graph $G \setminus \{w\}$, which is constructed by removing a node $w$ from a directed graph $G = (V, E)$:

$$G \setminus \{w\} = (V \setminus \{w\}, E \setminus \{w\}), \quad E \setminus \{w\} = \{(u, v) \in E \mid u \neq w, v \neq w\}. \tag{1}$$

Then, we can quantify the super-mediator degree of each node $w$, denoted by $medt(w)$, as the difference in the average influence degree with respect to the node removal, *i.e.*,

$$medt(w) = \sum_{v \in V} \sigma(v; G) p(v) - \sum_{v \in V \setminus \{w\}} \sigma(v; G \setminus \{w\}) p(v), \tag{2}$$

where $p(v)$ stands for the probability that the node $v$ becomes *an information source node*, that is, an initial active node. Of course, we want to identify the nodes that have large values of super-mediator degree.

We apply our bond percolation technique [9] to efficiently calculate the super-mediator degree $medt(w)$ for each node $w \in V$. Note first that the IC model on $G$ can be identified with the so-called susceptible/infective/recovered (SIR) model [15, 27] for the spread of a disease on $G$, where the nodes that become active at time $t$ in the IC model correspond to the infective nodes at time $t$ in the SIR model. Recall that in the

SIR model, each individual occupies one of the three states, "susceptible", "infected" and "recovered", where a susceptible individual becomes infected with a certain probability when it encounters an infected patient, and subsequently recovers at a certain rate. It is known that the SIR model on a network can be exactly mapped onto a bond percolation model on the same network [15, 6]. Thus, the IC model on $G$ is equivalent to a bond percolation model on $G$, that is, these two models have the same probability distribution for the final set of active nodes. Our bond percolation technique [9] exploits this relationship. Here, we present the algorithm for calculating $medt(w)$ based on the bond percolation technique. A bond percolation process on $G$ is the process in which each link of $G$ is randomly designated either "occupied " or "unoccupied" according to some probability distribution in which the occupation probability over each link $(u, v)$ is set to the diffusion probability $p_{u,v}$. Now, we consider $M$ times of bond percolation processes. Let $E_m$ denote the set of occupied links at the $m$-th bond percolation process and let $G_m$ denote the graph $(V, E_m)$, then for a large $M$, we can calculate the estimated influence degree $\bar{\sigma}(u; G)$ with a reasonable accuracy as follows:

$$\bar{\sigma}(u; G) = \frac{1}{M} \sum_{m=1}^{M} |R(u; G_m)|, \tag{3}$$

where $R(u; G_m)$ stands for a set of reachable nodes from $u$ on $G_m$ such that there is a path from $u$ to $v$ for $v \in R(u; G_m)$, and $|R(u; G_m)|$ is the number of nodes in $R(u; G_m)$. Here note that our bond percolation technique decomposes each graph $G_m$ into its SCCs, where SCC (strongly connected component) is a maximal subset $C$ of $V$ such that for all $u, v \in C$ there is a path from $u$ to $v$. Namely, $R(u; G_m) = R(v; G_m)$ if $u, v \in C$. Thus, we can obtain $R(u; G_m)$ for any node $u \in V$ by calculating $R(u; G_m)$ for only one node $u$ in each component $C$.

We obtain the following estimation formula by substituting Equation (3) into Equation (2):

$$medt(w) = \frac{1}{M} \sum_{v \in V} \sum_{m=1}^{M} |R(v; G_m)| p(v) - \frac{1}{M} \sum_{v \in V \setminus \{w\}} \sum_{m=1}^{M} |R(v; G_m \setminus \{w\})| p(v). \tag{4}$$

In order to efficiently calculate $R(v; G_m \setminus \{w\})$ for each pair of nodes, $v$ and $w$, we consider a set of reverse reachable nodes defined by

$$R^{-}(w; G_m) = \{v \in V \mid w \in R(v; G_m)\}.$$

Then, we can easily see that

$$v \notin R^{-}(w; G_m) \implies R(v; G_m \setminus \{w\}) = R(v; G_m).$$

Namely, for the $m$-th bond percolation process and a fixed node $w$, we can obtain $R(v; G_m \setminus \{w\})$ for any node $v \in V$ by calculating $R(v; G_m \setminus \{w\})$ only for $v \in R^{-}(w; G_m)$. Here, as described above, we can further improve the efficiency by applying SCC decomposition for a subgraph consisting of nodes in $R^{-}(w; G_m)$. Below we can summarize our proposed algorithm for calculating the super-mediator degree $medt(w)$ for each node $w \in V$.

1. Perform bond percolation process $M$ times ($m = 1, \cdots, M$);
   (a) For the $m$-th bond percolation process, calculate $R(v; G_m)$ by applying SCC decomposition;
   (b) For each $w \in V$, compute $R^-(w; G_m)$, and for each $v \in V$, set $R(v; G_m \setminus \{w\}) = R(v; G_m)$ if $v \notin R^-(w; G_m)$; otherwise calculate $R(v; G_m \setminus \{w\})$ by applying SCC decomposition;
2. Calculate the super-mediator degree $medt(w)$ according to Equation (4).

## 4 Characterizing Super-Mediator

As mentioned earlier, we attempt to characterize the property of the super-mediators by two important factors for each node $v$: the influence degree $\sigma(v; G)$ and the *reverse-influence degree* denoted by $\sigma^-(v; G)$. First of all, in order to quantify the relationships between these two factors, we define the probability $\sigma(u, v; G)$ that the node $v$ becomes active when $u$ is an information source node. Then, we can calculate $\sigma(v; G)$ by $\sum_{u \in V} \sigma(v, u; G)$. On the other hand, if we define the *reverse-influence degree* as the expected number of initial source nodes from which the information reaches the node $v$ at the end of information diffusion, we can define $\sigma^-(v; G)$ by

$$\sigma^-(v; G) = \sum_{u \in V} \sigma(u, v; G).$$

In order to further quantify the relationships, we consider the following reverse graph $G^-$, which is constructed by reversing any link $(u, v) \in E$ for a directed graph $G = (V, E)$.

$$G^- = (V, E^-), \quad E^- = \{(v, u) \mid (u, v) \in E\}. \tag{5}$$

Then, we can show that the reverse-influence degree of each node $v$ is equal to the influence degree of node $v$ on the reverse graph $G^-$, *i.e.*,

$$\sigma^-(v; G) = \sigma(v; G^-). \tag{6}$$

To confirm this fact, we introduce a function $R(u, v; G_m)$ of $v \in V$ such that $R(u, v; G_m) = 1$ if there is a path from $u$ to $v$ on $G_m$, and $R(u, v; G_m) = 0$ otherwise, where $G_m$ is the graph obtained by the $m$-th bond percolation process in Section 3. Noting that $R(u, v; G_m) = R(v, u; G_m^-)$, it is straightforward to show that Equation (6) holds as shown below:

$$\begin{aligned} \bar{\sigma}^-(v; G) &= \frac{1}{M} \sum_{m=1}^{M} \sum_{u \in V} R(u, v; G_m) \\ &= \frac{1}{M} \sum_{m=1}^{M} \sum_{u \in V} R(v, u; G_m^-) \\ &= \bar{\sigma}(v; G^-), \end{aligned} \tag{7}$$

where $G_m^-$ is the reverse graph of $G_m$. As a natural conjecture, we can expect that the super-mediator nodes are influential on both a given graph $G$ and its associated reverse graph $G^-$, which respectively corresponds to the influence degree $\sigma(v; G)$ and

the reverse-influence degree $\sigma^-(v; G)$. Thus, our hypothesis is that the super-mediators should be ranked high in terms of both of them. We empirically evaluate this hypothesis using three real world networks since exploring this analytically seems difficult.

## 5 Experiments

### 5.1 Datasets and Settings

We employed three datasets of large real networks. The first one is the Enron network, which is derived from the Enron Email Dataset [13]. We regarded each email address as a node, and constructed a link from email address $u$ to email address $v$ only if $u$ sent an email to $v$. The Enron network is a directed network which has $19,603$ nodes and $210,950$ directed links. The second one is the Blog network, which is a trackback network of Japanese blogs used by Kimura et al [11]. The Blog network is also a directed network which has $12,047$ nodes and $53,315$ directed links. The third one is the Wikipedia network, which is a network of people derived from the "list of people" within Japanese Wikipedia, also used by Kimura et al [11]. The Wikipedia network is a bidirectional network having $9,481$ nodes and $245,044$ directed links.

Below we explain the parameter settings of IC model. We first assume a generative model according to the beta distribution with a mean of $\mu$ for the diffusion probability $p_{v,w}$ for any link $(v, w) \in E$. Note that the beta distribution is the conjugate prior probability distribution for the Bernoulli distribution corresponding to a single toss of a coin. We further suppose that each diffusion probability is independently generated from the beta distribution with respect to each information diffusion process. Then the average occupied probability of the bond percolation process over each link reduces to $\mu$. Actually, this formulation is equivalent to assigning a uniform value $\mu$ to the diffusion probability $p_{v,w}$ for any link $(v, w) \in E$, that is, $p_{v,w} = \mu$. According to [6], we set the value of $\mu$ to a value that is less than or equal to $1/\bar{d}$, where $\bar{d}$ is the mean out-degree of a network. Thus, we investigate $\mu = r/\bar{d}$, where $r$ is a parameter with $0 < r \leq 1$. The parameter $M$ to estimate the expectation is set to 10,000 for all experiments. The probability that the node $v$ becomes an information source node was assumed to be uniform, *i.e.*, $p(v) = 1/|V|$.

### 5.2 Centralities

We also investigated whether or not super-mediators can be identified by heuristic methods based on the three well-known centrality measures, *degree centrality*, *closeness centrality*, and *betweenness centrality* that are commonly used as the influence measure in sociology. Let $G = (V, E)$ be a directed network for our analysis, and let $G^- = (V, E^-)$ be the reverse network of $G$. For the degree centrality, we consider the *out-degree* of node $v \in V$, $deg^+(v)$, defined as the number of links from $v$, and the *in-degree* of node $v \in V$, $deg^-(v)$, defined as the number of links to $v$; i.e.,

$$deg^+(v) = |\{(v, w) \in E\}|, \quad deg^-(v) = |\{(w, v) \in E\}| = |\{(v, w) \in E^-\}|.$$

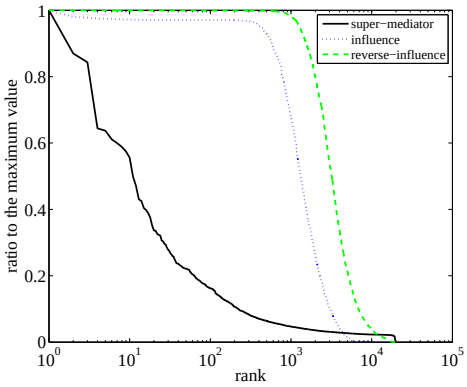



Fig. 1: Distribution of the super-mediator/influence/reverse-influence degree for the Enron network in case of $r = 1.0$.

For the closeness centrality, we consider the *closeness* of node $v \in V$, $close(v)$, defined as

$$close(v) = \frac{1}{|V|} \sum_{w \in V} \frac{1}{dist(v, w)},$$

and the *reverse closeness* of node $v \in V$, $close^{-}(v)$, defined as

$$close^{-}(v) = \frac{1}{|V|} \sum_{w \in V} \frac{1}{dist^{-}(v, w)},$$

where $dist(v, w)$ stands for the graph distance (shortest path length) from node $v$ to node $w$ in the network $G$, and $dist^{-}(v, w)$ stands for the graph distance from node $v$ to node $w$ in the reverse network $G^{-}$. For the betweenness centrality, we consider the *betweenness* of node $v \in V$, $betw(v)$, defined as the total number of shortest paths between pairs of nodes that pass through $v$. We consider detecting super-mediators by ranking the nodes in decreasing order with respect to a centrality measure. We refer to the detection methods by centrality measures $deg(v)$, $deg^{-}(v)$, $close(v)$, $close^{-}(v)$, and $betw(v)$ as the *out-degree*, *in-degree*, *closeness*, *reverse closeness*, and *betweenness* methods, respectively.

### 5.3 Results

**Confirmation of Properties of Super-Mediators** First, we investigated the distributions of the three measures, *i.e.*, the super-mediator, influence, and reverse-influence degree in the Enron network to see how much they differ to each other in terms of characterizing each node. In Fig. 1, the values of "ratio to the maximum value" in each degree are depicted as a function of node rank. Note that node rank is different for each degree. It can be observed that the curves for the influence and reverse-influence degree are similar to each other, while the curve for the super-mediator degree is quite different from the other two. Each curve is almost flat for the first two. The one for the

Table 1: Top 3 email accounts (nodes) in the super-mediator, influence, and reverse-influence degree ranking for the Enron network ($r = 1.0$).

| | account name (ID: ratio to the maximum degree value) | | |
|---|---|---|---|
| rank | super-mediator | influence | reverse-influence |
| 1 | jeff.skilling (642: 1.000) | bob.ambrocik (16734: 1.000) | tom.alonso (5510: 1.000) |
| 2 | kenneth.lay (471: 0.870) | technology.enron (17219: 0.979) | jeff.richter (1768: 0.999) |
| 3 | sally.beck (535: 0.843) | outlook.team (10779: 0.978) | chris.mallory (5933: 0.999) |

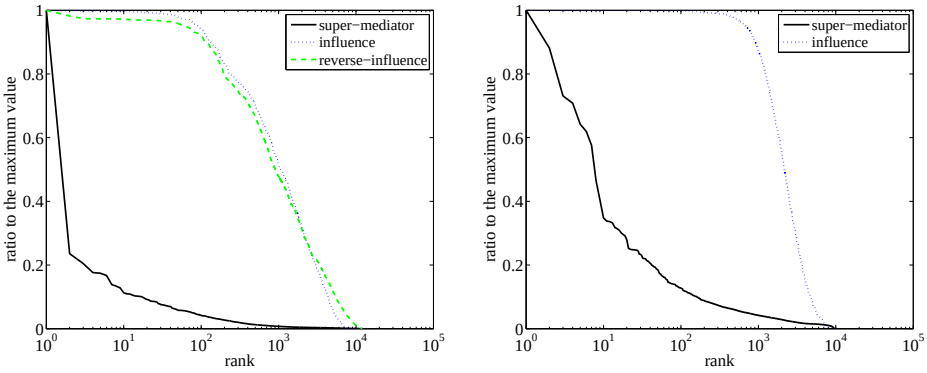
Table 2: The rank of the top 3 super-mediators for the Enron network in the influence and reverse-influence degree ranking ($r = 1.0$).

| | rank (ratio to the maximum degree value) | | |
|---|---|---|---|
| ID | super-mediator | influence | reverse-influence |
| 642 | 1 (1.000) | 441 (0.947) | 642 (0.999) |
| 471 | 2 (0.870) | 122 (0.970) | 374 (0.999) |
| 535 | 3 (0.843) | 126 (0.970) | 377 (0.999) |

influence degree maintains a relatively high ratio close to 1.0 up to approximately top 300 nodes and the one for the reverse-influence degree up to approximately top 1,000 nodes. This means that there is very little difference among these top ranked nodes as far as the influence is concerned. On the other hand, the distribution curve rapidly decreases to the top 1,000 nodes for the super-mediator degree. We can conclude that the super-mediator ranking can characterize each node by far clearly than the influence and reverse-influence ranking.

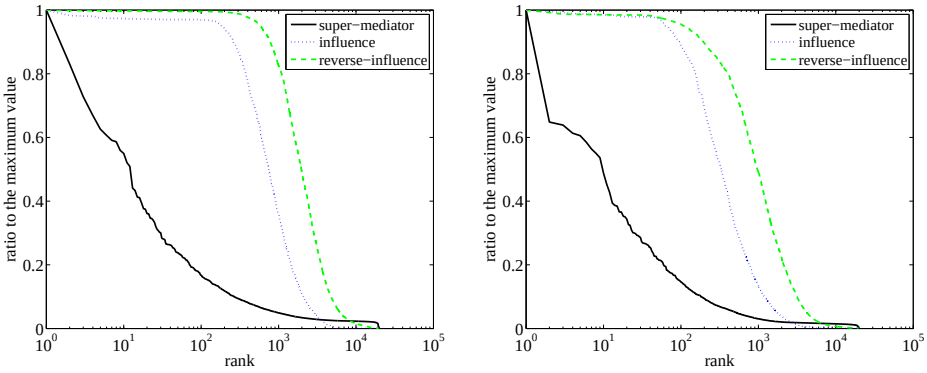
We further examined the top 3 nodes in each ranking for the Enron network in case of $r = 1.0$, and summarized them in Table 1. Again, we can observe that there is a clear difference in the values of the super-mediator degree among the top 3 email accounts (nodes), but the difference is not clear for the other two degree, especially the reverse-influence degree. In addition, these top 3 ranked super-mediators are different from the top 3 ranked nodes for the other two: the influence degree and the reverse-influence degree. It is notable that "Jeffrey Skilling" (the top ranked) and "Kenneth Lay" (the second ranked) in the super-mediator degree are key persons of the Enron scandal: "Jeffrey Skilling" is the former president of Enron and "Kenneth Lay" was the CEO of Enron. Both of them do not appear in the top 3 in both the influence and reverse-influence degree ranking. "Jeffrey Richter", the second ranked in the reverse-influence degree, is known as a trader of Enron, but is not as well-known as the former two executives. These observations suggest the super-mediator degree can be a promising measure to identify nodes that actually play an important role in a given network.

Next, we investigated how the top 3 super-mediators for the Enron network rank in terms of the influence and the reverse-influence degree values. Our conjecture is that the super-mediators should be ranked high in these two measures. The results are summarized in Table 2. It is found that these super-mediators are ranked relatively high, at least they are in the top 5% nodes. This confirms our conjecture. However, because their curves are flat, there are many other nodes that are ranked high in these two measures. This means that the reverse is not necessarily true, *i.e.*, super-mediators have high




(a) Blog network (b) Wikipedia network

Fig. 2: Distribution of the super-mediator/influence/reverse-influence degree for the Blog and Wikipedia networks in case of $r = 1.0$.




(a) $r = 0.5$ (b) $r = 0.25$

Fig. 3: Distribution of the super-mediator/influence/reverse-influence degree for the Enron network in cases of $r = 0.5$ and $r = 0.25$.

values for these two measures but having high values for these two measures are not necessarily super-mediators as defined in Equation (2). Indeed, we observed the same tendencies for the Blog and Wikipedia networks. Here, we show only the distributions of the three measures for these networks in case of $r = 1.0$ in Figs. 2a and 2b, respectively. Note that since the Wikipedia network is bidirectional, the reverse-influence degree is equivalent to the influence degree, so it is not shown in Fig. 2b.

**Further Analysis** Using the Enron network, we further analyzed the properties of super-mediators. First, we investigated the effect of diffusion probability by varying the value of $r$. Figures 3a and 3b, and Tables 3 and 4 show the results for the case of $r = 0.5$ and $r = 0.25$, respectively. Here, each table indicates ranks and the values of "ratio to the maximum value" with respect to super-mediator, influence and reverse influence

Table 3: The rank of the top 3 super-mediators for the Enron network in the influence and reverse-influence degree ranking ($r = 0.5$).

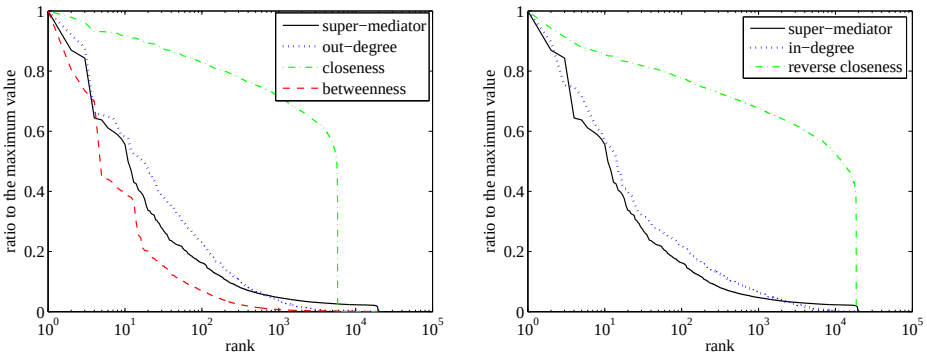
| | rank (ratio to the maximum degree value) | | |
|---|---|---|---|
| ID | super-mediator | influence | reverse-influence |
| 535 | 1 (1.000) | 66 (0.970) | 94 (0.996) |
| 471 | 2 (0.831) | 114 (0.969) | 128 (0.995) |
| 642 | 3 (0.728) | 426 (0.742) | 341 (0.985) |

Table 4: The rank of the top 3 super-mediators for the Enron network in the influence and reverse-influence degree ranking ($r = 0.25$).

| | rank (ratio to the maximum degree value) | | |
|---|---|---|---|
| ID | super-mediator | influence | reverse-influence |
| 535 | 1 (1.000) | 52 (0.970) | 46 (0.979) |
| 6 | 2 (0.648) | 185 (0.734) | 1 (1.000) |
| 471 | 3 (0.639) | 154 (0.799) | 144 (0.931) |

degree for the top 3 super-mediators. It is obvious that the distribution curves shown in Fig. 3a and 3b share the same tendency as those in Fig. 1. The notable difference is that the flat region of each curve shrinks for the influence and reverse-influence degree as the diffusion probability becomes smaller. This is because both $\sigma(v; G)$ and $\sigma^{-}(v; G)$ become smaller for every node $v$ in accordance with the decrease of the diffusion probability. Also from Tables 3 and 4, we can see the same tendency as for the case of $r = 1.0$ although the top 3 nodes and their rankings for $r = 0.5$ and $r = 0.25$ are not exactly the same as for $r = 1.0$. Further we notice that all the values for the influence and reverse-influence degree are not very high due to the aforementioned shrink of the flat region, *i.e.*, third rank for $r = 0.5$ and the second and the third rank for $r = 0.25$, but overall both the influence degree and the reverse-influence degree are high for the high ranked super-mediators. Indeed, in the influence and reverse-influence ranking, these nodes are within the top 3% nodes at $r = 0.5$, and within the top 1% nodes at $r = 0.25$.

Next, we investigated whether the conventional centrality measures can serve as a good measure to identify the super-mediators. Figure 4 displays the values of "ratio to the maximum value" as a function of node rank with respect to out-degree $deg^{+}(v)$, in-degree $deg^{-}(v)$, closeness $close(v)$, reverse closeness $close^{-}(v)$, and betweenness $betw(v)$. We observe that the distributions of out-degree $deg^{+}(v)$, in-degree $deg^{-}(v)$ and betweenness $betw(v)$ are similar to the distribution of the super-mediator degree, while the distributions of closeness $close(v)$ and reverse closeness $close^{-}(v)$ are similar to the distributions of the influence and reverse-influence degree. Here note that the value of "ratio to the maximum value" of a node with respect to the super-mediator degree is less than 0.2 for nodes ranked after the top 100. Thus, we focused on the top 100 nodes, and examined the similarity between the super-mediator ranking and the other ranking, *i.e.*, the out-degree, in-degree, closeness, reverse closeness, and betweenness ranking. Here, we measured the similarity between the top $k$ nodes for one ranking method, denoted as a set $A_k$, and those for the other ranking method, denoted as a set $A'_k$, by the $F$-measure




(a) Out-degree, closeness, and betweenness (b) In-degree and reverse closeness

Fig. 4: Distributions of conventional centrality measures for the Enron network.

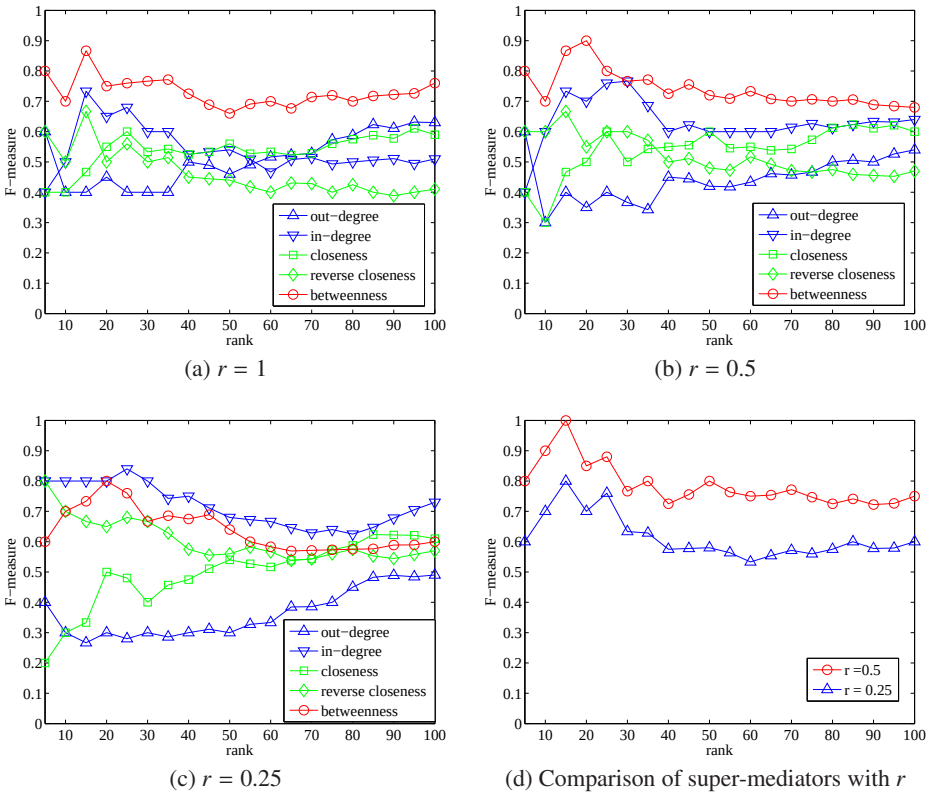
$F(k)$ defined by

$$F(k) = \frac{|A_k \cap A'_k|}{k}.$$

Figures 5a, 5b and 5c show the results for the cases of $r = 1.0$, $r = 0.5$ and $r = 0.25$, respectively. Figure 5d displays the similarities between the super-mediator ranking for the case of $r = 1.0$ and that of $r = 0.5$ and $r = 0.25$. We notice that the super-mediators depend on the value of the diffusion probability from Fig. 5d. We further notice that the betweenness centrality is best when the diffusion probability is large (Fig. 5a, 5b) and the in-degree centrality becomes better when the diffusion probability gets smaller (Fig. 5c). It is interesting that the out-degree centrality is not as good as the in-degree centrality. Further investigation is needed to understand this phenomenon. Table 5 shows the top 3 nodes for the out-degree, in-degree, closeness, reverse-closeness, and betweenness centrality in case of $r = 1.0$. These should be compared with the node IDs in Table 2, *i.e.*, 642, 471 and 535. Two nodes (642, 471) for the betweenness centrality match them and one node (535) for the out-degree, in-degree and closeness centrality matches them. This supports the above observation. In summary no single centrality measure works equally well for a wide range of the diffusion probability. The betweenness centrality is a good measure when the diffusion probability is large and in-degree centrality is a good measure when the diffusion probability is small. This is intuitively understandable. When the diffusion probability is large, there are many long diffusion sequences, in which case the betweenness plays a key role, whereas the diffusion probability is small, many of the diffusion sequences are short, in which case node degree plays a key role.

## 6 Conclusion

We addressed a problem of identifying and characterizing influential nodes in a social network which we call "super-mediators" (nodes which play a role of mediator), *i.e.*, nodes that play an important role in receiving the information and passing it to other




(a) $r = 1$ (b) $r = 0.5$

(c) $r = 0.25$ (d) Comparison of super-mediators with $r$

Fig. 5: Relation between conventional centrality and super-mediator degree for the Enron network.

Table 5: Top 3 nodes for conventional centrality measures for the Enron network for $r = 1.0$

| rank | out-degree | in-degree | closeness | reverse-closeness | betweenness |
|---|---|---|---|---|---|
| 1 | 451 | 6 | 535 | 6 | 6 |
| 2 | 10779 | 203 | 10779 | 203 | 642 |
| 3 | 535 | 535 | 451 | 684 | 471 |

nodes. This notion of influential nodes is different from the conventional one in which a node is said to be influential if the information starting from that node spreads to many other nodes. We quantified the degree of importance as a super-mediator degree and formulated this as the difference of the average influence degree with respect to the node removal. If a node is a super-mediator, removal of this node from the network will substantially decrease the average influence degree. Thus finding the most influential super-mediator is finding a node that maximizes this difference. We can rank the super-mediators according to the amount of difference. This computation requires to estimate influence degree of each node, which is defined to be the expected number of

active nodes at the end of information diffusion process, and is very time consuming. We used our bond percolation approach to simulate an individual diffusion process and the expectation is approximated by the empirical mean of many trials of diffusion process. We conjectured that super-mediators would have both large influence degree, *i.e.*, capable of widely spreading information to other recipient nodes, and large reverse-influence degree, *i.e.*, capable of widely receiving information from other information source nodes. In fact reverse-influence degree of a node in a graph is the same as the influence degree of the same node of the graph in which the edge direction is reversed for all edges. We conducted extensive experiments using three real world social networks (Enron, Blog and Wikipedia) with different diffusion probability assuming independent cascade model, and confirmed that this conjecture is correct, but the reverse is not correct, *i.e.*, nodes that have both large influence degree and large reverse-influence degree are not necessarily super-mediators. The performance of super-mediator degree is tested in the Enron network. The top three super-mediators identified by our method are confirmed to be actually influential. We further investigated how well the conventional centrality measures (in-degree, out-degree, closeness, reverse-closeness and betweenness) capture super-mediators. In short the in-degree centrality is a good measure when the diffusion probability is small and the betweenness centrality is a good measure when the diffusion probability is large, but the super-mediators do depend on the value of the diffusion probability and no single centrality measure works equally well for a wide range of the diffusion probability. Our immediate future work is to investigate the generality of the findings reported in this paper for a variety of networks and elucidate why the out-degree centrality is not as good a measure as the in-degree centrality.

## Acknowledgments

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research under Grant No. AOARD-13-4042, and JSPS Grant-in-Aid for Scientific Research (C) (No. 23500194).

## References

1. Bakshy, E., Hofman, J., Mason, W., Watts, D.: Everyone's an influencer: Quantifying influences on twitter. In: Proceedings of the 4th International Conference on Web Search and Data Mining (WSDM2011). pp. 65–74 (2011)
2. Bakshy, E., Karrer, B., Adamic, L.A.: Social influence and the diffusion of user-created content. In: Proceedings of the 10th ACM conference on Electronic Commerce. pp. 325–334 (2009)
3. Dow, P., Adamic, L., Friggeri, A.: The anatomy of large facebook cascades. In: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM) (2013)
4. Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters 12, 211–223 (2001)
5. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. SIGKDD Explorations 6, 43–52 (2004)
6. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003). pp. 137–146 (2003)

7. Kimura, M., Saito, K., Motoda, H.: Blocking links to minimize contamination spread in a social network. ACM Transactions on Knowledge Discovery from Data 3, 9:1–9:23 (2009)
8. Kimura, M., Saito, K., Motoda, H.: Efficient estimation of influence functions fot sis model on social networks. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09) (2009)
9. Kimura, M., Saito, K., Nakano, R.: Extracting influential nodes for information diffusion on a social network. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI-07). pp. 1371–1376 (2007)
10. Kimura, M., Saito, K., Nakano, R., Motoda, H.: Finding influential nodes in a social network from information diffusion data. In: Proceedings of the 2nd International Workshop on Social Computing, Behavioral Modeling and Prediction (SBP09). pp. 138–145 (2009)
11. Kimura, M., Saito, K., Nakano, R., Motoda, H.: Extracting influential nodes on a social network for information diffusion. Data Mining and Knowledge Discovery 20, 70–97 (2010)
12. Kleinberg, J.: The convergence of social and technological networks. Communications of ACM 51(11), 66–72 (2008)
13. Klimt, B., Yang, Y.: The enron corpus: A new dataset for email classification research. In: Proceedings of the 2004 European Conference on Machine Learning (ECML'04). pp. 217–226 (2004)
14. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. In: Proceedings of the 7th ACM Conference on Electronic Commerce (EC'06). pp. 228–237 (2006)
15. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45, 167–256 (2003)
16. Newman, M.E.J., Forrest, S., Balthrop, J.: Email networks and the spread of computer viruses. Physical Review E 66, 035101 (2002)
17. Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002). pp. 61–70 (2002)
18. Romero, D., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In: Proceedings of the 20th International World Wide Web Conference (WWW2011). pp. 695–704 (2011)
19. Saito, K., Kimura, M., Motoda, H.: Discovering influential nodes for sis models in social networks. In: Proceedings of the Twelfth International Conference of Discovery Science (DS2009). pp. 302–316. Springer, LNAI 5808 (2009)
20. Saito, K., Kimura, M., Motoda, H.: Discovery of super-mediators of information diffusion in social networks. In: Proceedings of the Thirteenth International Conference of Discovery Science (DS2010). pp. 144–158. Springer, LNAI 6332 (2010)
21. Saito, K., Kimura, M., Ohara, K., Motoda, H.: Learning continuous-time information diffusion model for social behavioral data analysis. In: Proceedings of the 1st Asian Conference on Machine Learning (ACML2009). pp. 322–337. LNAI 5828 (2009)
22. Saito, K., Kimura, M., ohara, K., Motoda, H.: Behavioral analyses of information diffusion models by observed data of social network. In: Proceedings of the 2010 International Conference on Social Computing, Behavioral Modeling and Prediction (SBP10). pp. 149–158. LNCS 6007 (2010)
23. Saito, K., Kimura, M., Ohara, K., Motoda, H.: Which targets to contact first to maximize influence over social network. In: Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP 2013). pp. 359–367. LNCS 7812 (2013)
24. Sheldon, D., Dilkina, B., Elmachtoub, A., Finseth, R., Sabharwal, A., Conrad, J., Gomes, C., Shmoys, D., Allen, W., Amundsen, O., Vaughan, W.: Maximizing the spread of cascades using network design. In: Proceedings of the Twenty-Sixth Conference Annual Conference on

Uncertainty in Artificial Intelligence (UAI-10). pp. 517–526. AUAI Press, Corvallis, Oregon (2010)
25. Steeg, G.V., Ghosh, R., Lerman, K.: What stops social epidemics? In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM). pp. 377–384 (2011)
26. Watts, D.J.: A simple model of global cascades on random networks. Proceedings of National Academy of Science, USA 99, 5766–5771 (2002)
27. Watts, D.J., Dodds, P.S.: Influence, networks, and public opinion formation. Journal of Consumer Research 34, 441–458 (2007)