Information Sciences 329 (2016) 985–1000

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Super mediator – A new centrality measure of node importance for information diffusion over social network

Kazumi Saito[a], Masahiro Kimura[b,*], Kouzou Ohara[c], Hiroshi Motoda[d,e]

[a] School of Administration and Informatics, University of Shizuoka, Japan
[b] Department of Electronics and Informatics, Ryukoku University, Japan
[c] Department of Integrated Information Technology, Aoyama Gakuin University, Japan
[d] Institute of Scientific and Industrial Research, Osaka University, Japan
[e] School of Computing, University of Tasmania, Australia

ARTICLE INFO

Article history:
Received 1 April 2014
Revised 28 November 2014
Accepted 12 March 2015
Available online 23 March 2015

Keywords:
Information diffusion
Super-mediator
Influence degree
Reverse-influence degree

ABSTRACT

We propose an efficient method to discover a new type of influential nodes in a social network, which we name "super-mediators", *i.e.*, those nodes which, if removed, decrease information spread. It is formulated mathematically as a problem of difference maximization of the average influence degree with respect to removal of a node, *i.e.*, a node that contributes to making the difference large is influential. This definition requires use of information diffusion model for their identification and thus is "model-driven". The other definition which is more empirical is that super-mediators are those nodes that appear frequently in long diffusion sequences but much less frequently in short diffusion sequences. This definition does not require any model but does require abundant information diffusion data and thus is "data-driven". We attempt to characterize the property of super-mediators from various angles: how the resulting super-mediators are different between these two definitions and which is more reasonable, how super-mediators are compared with nodes identified by other centralities, *e.g.*, betweenness, degree, closeness, etc., how super mediators are different from the solution of well-studied influence maximization problem, *i.e.*, nodes capable of widely spreading information to other recipient nodes, and the solution of reverse-influence maximization problem, *i.e.*, nodes capable of widely receiving information from other information source nodes. We conducted extensive experiments using three real world social networks. The major findings are (1) model-driven super-mediator degree has the best discrimination capability, while influence degree, reverse-influence degree, and data-driven super-mediator degree are much less discriminative (all flat for high ranked nodes), (2) model-driven super-mediators have high scores for either influence degree or reverse-influence degree, while data-driven super-mediators have high scores for both, and (3) model-driven super-mediators are closely correlated with betweenness centrality, but the strength of the correlation depends on the value of diffusion probability.

[*] Corresponding author.
E-mail addresses: k-saito@u-shizuoka-ken.ac.jp (K. Saito), kimura@rins.ryukoku.ac.jp (M. Kimura), ohara@it.aoyama.ac.jp (K. Ohara), motoda@ar.sanken.osaka-u.ac.jp, hmotoda@utas.edu.au (H. Motoda).

http://dx.doi.org/10.1016/j.ins.2015.03.034
0020-0255/© 2015 Elsevier Inc. All rights reserved.

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 spreading 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. Theses phenomena have attracted interest of many researchers from diverse fields [14], such as sociology, psychology, economy, and computer science, and a substantial amount of work has been carried out to analyze and mine information diffusion (*i.e.*, cascading) processes in large social networks [17,16,1,2,20,28,4].

Widely used information diffusion models in these studies are *independent cascade (IC)* [5,7,10], *linear threshold (LT)* [31,32] and their variants [9,21,6,11,22,23]. 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 a 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 [7,8,30] or by applying a broader set of intervention strategies such as node and link additions [19,27]. 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.

Nodes that maximize information spread are certainly influential, but this is not the only way to define influential nodes. In this paper we introduce a new type of influential nodes, which we call super-mediators, *i.e.* nodes, if removed, decrease information spread, and we explore their characteristics. Super-mediators, by intuition, would play an important role in both receiving information and passing it to others. Also, such super-mediators must be key persons in social recommender system [29].

In [25] 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 selected target nodes are guaranteed to start spreading information. Rather we send the same information from outside of the network to the selected targets as a probabilistic diffusion 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. “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 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.

In [24] we empirically[1] studied nodes that frequently appear in the long information diffusion sequences but much less frequently in short diffusion sequences,[2] and called them also super-mediators. Nodes in the long sequences that are shared by many other long sequences should play an important role in both receiving information and passing it to other nodes. This is another definition of super-mediators. This approach does not require any model for information diffusion. What is required is that there are abundant observed information diffusion sequences. There is no need to know the network structure, either. Thus, we call this approach “data-driven” to distinguish it from what we propose in this paper.

The work in this paper is motivated by these two studies [25,24] and is an extended version of what we presented in [26]. We take a “model-driven” approach that requires to compute influence degree, and mathematically define the super-mediators as the solution of an optimization problem and rank them according to the value of the objective function. The influence degree $\varphi(v)$ 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 Section 2 for a more rigorous definition). The average influence degree of the whole network is defined to be the average of $\varphi(v)$ 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.

The main objective of this paper is to explore the properties of model-driven super-mediators returned as the solution of the optimization problem, comparing them with nodes identified influential by other measures including data-driven super-mediator studied earlier. 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. We want to answer the following questions: how model-driven super-mediators are different from data-driven super-mediators and which is more reasonable, how super mediators are

[1] We call it empirical in 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.

different from the solution of influence maximization problem and the solution of reverse-influence maximization problem, and how super-mediators are compared with nodes identified by other centralities, *e.g.*, betweenness, degree, closeness, etc.

We use the 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 [8] and pruning [9] algorithms to efficiently estimate the influence degree. We further recently improved these techniques and reduced the computation time drastically [13] (but this is not our focus in this paper).

We have performed extensive experiments using three real world networks (Enron, Blog and Wikipedia). Probabilistic information diffusion process is simulated by the IC model. The Enron network is useful because we can judge whether or not the returned nodes (people) are in powerful positions in the Enron company by different measures in light of open literature. We conclude that (1) model-driven super-mediator degree has the best discrimination capability, *i.e.*, rank–score curves for influence degree, reverse-influence degree, and data-driven super-mediator degree are all flat for the top 1000 nodes, (2) we originally anticipated that super-mediators defined either way (data-driven or model-driven) have always high scores for both influence degree and reverse-influence degree. There is a difference between data-driven super-mediators and model-driven super-mediators: model-driven super-mediators have high scores for either influence degree or reverse-influence degree, while data-driven super-mediators have high scores for both. Thus, data-driven super-mediator is useful in that it can identify nodes with both high influence degree and high reverse influence degree without knowing network topology, provided the information diffusion data are available, and (3) model-driven super-mediators are closely correlated with betweenness centrality, but the strength of the correlation depends on the value of diffusion probability. When the value is small, model-driven super-mediators are closely correlated with in-degree centrality. In short, no single conventional centrality measure which is defined solely by network topology works equally well for a wide range of the diffusion probability. The model-driven super-mediator we proposed in this paper explicitly considers information diffusion process, but it is different from influence degree and reverse-influence degree, and serves as a new useful centrality measure that can be added to the existing pool.

The paper is organized as follows. Section 2 gives a brief description of the independent cascade model. Section 3 defines super-mediators (both data-driven and model-driven) and gives respective algorithm to find and rank them. Section 4 defines other centrality measures that we used for comparative study (influence degree, reverse-influence degree and major conventional centrality degrees). Section 5 reports experimental results and various findings. 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 [7,12]. 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 $D(v; G)$ denote a set of active nodes at the end of the random diffusion process, and $A(v; G)$ be the number of nodes in $D(v; G)$, *i.e.*, $A(v; G) = |D(v; G)|$. It is noted that $A(v; G)$ is a random variable. We denote the expected value of $A(v; G)$ by $\varphi(v; G)$, and call it the *influence degree of* v. Henceforth, we use $D(v)$ instead of $D(v; G)$ if G is obvious from the context.

3. Super-mediator identification

Super-mediators are nodes which play an important role in information diffusion over a social network that they belong to. To identify these nodes, in this paper, we introduce a new centrality measure referred to as the *super-mediator degree* that is derived from influence degree computed according to the underlying information diffusion model. We distinguish this from the one we studied in our past work [24] which is the one empirically derived from observed information diffusion results without assuming any diffusion model. We first explain *the data-driven super-mediators* that correspond to those identified by the super-mediator degree based on observed data, and then discuss *the model-driven super-mediators* that correspond to those identified by the super-mediator degree based on the diffusion model.

[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 [21].

3.1. Data-driven super-mediator

First, we revisit the data-driven super-mediator [24]. Fig. 1(a) shows an example of information diffusion where the existence of such super-mediator nodes is suggested. In this figure, we plotted the number of activated nodes as the information spreads with time from a source node using the IC model.[4] The plot includes the results of 5000 independent simulations starting from the same node. From this figure, we can observe that (1) diffusion samples exhibit large variations due to their stochastic nature and (2) some curves clearly exhibit sigmoidal behavior with a big jump during the course of diffusion process. From the latter observation, we can conjecture that there exist super-mediator nodes which play an important role to pass the information to other nodes.

In Fig. 1(b), we plotted the distribution of the final influence degree for the above 5000 simulations. From this figure, we can observe that there exist a few bell-shaped curves (which can be approximated by quadratic equations) in a logarithmic scale for each axis, which suggests that the influence degree distribution consists of a small number of lognormal-like distributions. Combining the observation from Fig. 1(a), we conjecture that super-mediators appear as a limited number of active nodes in some lognormal components with high influence degree. Therefore, in order to discover these super-mediator nodes from information diffusion samples, we attempt to divide the diffusion samples in two groups (lower and upper), each assuming some probability distribution, find the best split by maximizing the likelihood, and rank the nodes in the upper samples by the *F*-measure.

Now we assume that an arbitrary node $v \in V$ in a network G can serve as an information source node multiple times, and we consider *a diffusion sample* $D(v)$ that is a set of nodes remaining active at the end of diffusion process, each initiated at node v. Let $\mathcal{S}(v) = \{1, 2, \ldots, M(v)\}$ be a set of indices with respect to information diffusion samples for node v, *i.e.*, $\{D_1(v), D_2(v), \ldots, D_{M(v)}(v)\}$, where $D_m(v)$ stands for the mth diffusion sample. Then, the procedure to identify the data-driven super-mediators is divided into two steps: the clustering step and the ranking step. The former step divides $\mathcal{S}(v)$ into two groups, $\mathcal{S}_1(v)$ and $\mathcal{S}_2(v)$, which are the upper group that includes samples with relatively high influence degree and the lower group that includes the rests, respectively.[5] Namely, $\mathcal{S}_1(v) \cup \mathcal{S}_2(v) = \mathcal{S}(v)$, $\mathcal{S}_1(v) \cap \mathcal{S}_2(v) = \varnothing$, and $\min_{m \in \mathcal{S}_1(v)} |D_m(v)| > \max_{m \in \mathcal{S}_2(v)} |D_m(v)|$. By assuming that each sample is independently drawn from either the upper or the lower group, the following likelihood function can be considered.

$$\mathcal{L}(\mathcal{S}(v); \mathcal{S}_1(v), \Theta) = \prod_{q \in \{1,2\}} \prod_{m \in \mathcal{S}_q(v)} p(m; \theta_q), \tag{1}$$

where $p(m; \theta_q)$ denotes a probability distribution with the parameter set θ_q for the mth diffusion sample, and $\Theta = \{\theta_1, \theta_2\}$. If it is assumed that the influence degree distribution consists of lognormal components, $p(m; \theta_q)$ can be expressed as follows:

$$p(m; \theta_q) = \frac{1}{\sqrt{2\pi\sigma_q^2}|D_m(v)|} \exp\left(-\frac{(\log|D_m(v)| - \mu_q)^2}{2\sigma_q^2}\right), \tag{2}$$

where $\theta_q = \left\{\mu_q, \sigma_q^2\right\}$. Then, based on the maximum likelihood estimation, the optimal upper group $\hat{\mathcal{S}}_1(v)$ can be identified by the following equation.

$$\hat{\mathcal{S}}_1(v) = \underset{\mathcal{S}_1(v)}{\operatorname{argmax}} \left\{\mathcal{L}(\mathcal{S}; \mathcal{S}_1(v), \hat{\Theta})\right\}, \tag{3}$$

where $\hat{\Theta}$ is the set of maximum likelihood estimators.

$\hat{\mathcal{S}}_1(v)$ is efficiently obtained by focusing on the case where $p(m; \theta_q)$ is the lognormal distribution defined in Eq. (2) though the applicability of the method is not limited to this case. Noting that the following equations hold for the maximum liklihood estimate,

$$\hat{\mu_q} = \frac{1}{|\mathcal{S}_q(v)|} \sum_{m \in \mathcal{S}_q(v)} \log|D_m(v)|, \quad \hat{\sigma}_q^2 = \frac{1}{|\mathcal{S}_q(v)|} \sum_{m \in \mathcal{S}_q(v)} \left(\log|D_m(v)| - \hat{\mu_q}\right)^2, \tag{4}$$

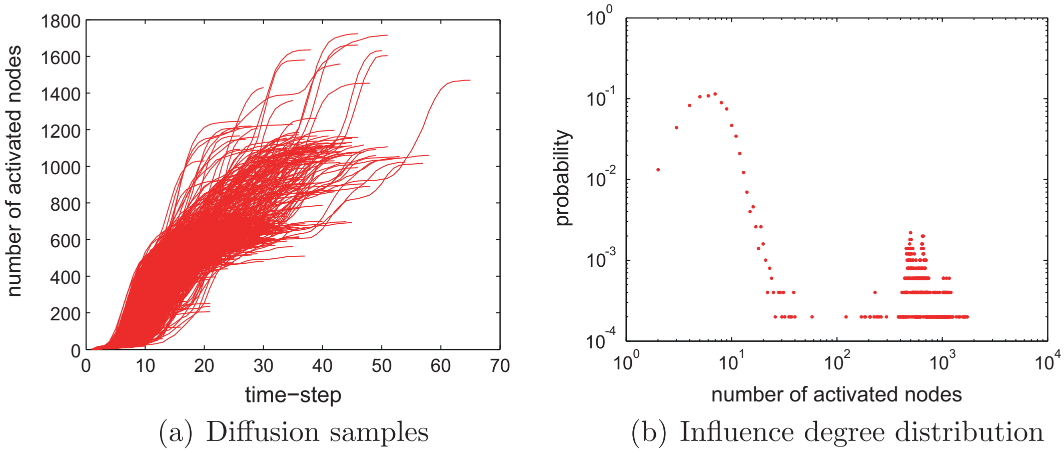
Eq. (3) for a candidate upper group $\mathcal{S}_1(v)$ can be transformed as follows.

$$\hat{\mathcal{S}}_1(v) = \underset{\mathcal{S}_1(v)}{\operatorname{argmax}} \left\{2\log\mathcal{L}(\mathcal{S}(v); \mathcal{S}_1(v), \hat{\Theta})\right\} = \underset{\mathcal{S}_1(v)}{\operatorname{argmax}} \left\{-\sum_{q \in \{1,2\}} |\mathcal{S}_q| \log\left(\hat{\sigma}_q^2\right)\right\}. \tag{5}$$

Therefore, $\hat{\mathcal{S}}_1(v)$ can be efficiently obtained by simply updating the sufficient statistics for calculating the maximum likelihood estimators by successively shifting its boundary between $\mathcal{S}_1(v)$ and $\mathcal{S}_2(v)$. Here, since there might exist more than one diffusion sample with the same length, the following operation is defined to obtain the set of elements with the maximum diffusion length:

[4] The network used to generate these data is the blog network (see Section 5.1).

[5] This approach can be straightforwardly extended in case of q-groups division although here we consider the simplest case ($q = 2$) for the sake of simplicity.



(a) Diffusion samples (b) Influence degree distribution

Fig. 1. Information diffusion from a source node in the blog network for the IC model ($p = 0.1$).

$$\eta(\mathcal{S}(v)) = \left\{ m;\ |D_m(v)| = \max_{m \in \mathcal{S}(v)} \{|D_m(v)|\} \right\}. \tag{6}$$

Then, this clustering algorithm is summarized as follows.

```
D1. Initialize $\mathcal{S}_1(v) \leftarrow \eta(\mathcal{S}(v))$, $\mathcal{S}_2(v) \leftarrow \mathcal{S}(v) \backslash \eta(\mathcal{S}(v))$, and $\hat{L} \leftarrow -\infty$.
D2. Iterate the following procedure:
    D2-1. Set $\mathcal{S}_1(v) \leftarrow \mathcal{S}_1(v) \cup \eta(\mathcal{S}_2(v))$, and $\mathcal{S}_2(v) \leftarrow \mathcal{S}_2(v) \backslash \eta(\mathcal{S}_2(v))$.
    D2-2. If $\mathcal{S}_2(v) = \eta(\mathcal{S}_2(v))$, then terminate the iteration.
    D2-3. Calculate $L = -\sum_{q \in \{1,2\}} |\mathcal{S}_q(v)| \log(\hat{\sigma}_q^2)$.
    D2-4. If $\hat{L} < L$ then set $\hat{L} \leftarrow L$ and $\hat{\mathcal{S}}_1(v) \leftarrow \mathcal{S}_1(v)$
D3. Output $\hat{\mathcal{S}}_1(v)$, and terminate the algorithm.
```

It is noted that this algorithm always produces the optimal result with the computational complexity of $O(M(v))$, while the standard EM algorithm requires much more computational cost and is easily trapped in a local optimum. See [24] for the more detailed discussions.

Next, we describe the ranking step that measures the importance of nodes based on the groups obtained by the first clustering step. The intuition behind this empirical measure is that a node is important for information diffusion if it frequently appears in the long diffusion sequences, but much less in the short ones. To quantify this and define the super-mediator degree of a node, the following F-measure $F(w; v)$, a widely used measure in information retrieval, is employed, which is the harmonic average of *recall* and *precision* of a node w for the node v. The recall means the number of samples that include the node w in the upper group divided by the total number of samples in the upper group, and the precision means the number of samples that include a node w in the upper group divided by the total number of the node w in the samples.

$$F(w; v) = \frac{2|\{m;\ m \in \hat{\mathcal{S}}_1(v), w \in D_m(v)\}|}{|\hat{\mathcal{S}}_1(v)| + |\{m;\ m \in \mathcal{S}(v), w \in D_m(v)\}|}. \tag{7}$$

Indeed, the super-mediator degree of a node w based on observed data, $SMD_{data}(w)$, is defined as the following expected F-measure:

$$SMD_{data}(w) = \sum_{v \in V} F(w; v)\kappa(v), \tag{8}$$

where $\kappa(v)$ stands for the probability that the node v becomes an information source node, that is, an initial active node, which can be empirically estimated by $\kappa(v) = M(v)/\sum_{v \in V} M(v)$.

3.2. Model-driven super-mediator

Next, we describe the model-driven super-mediators and the measure to identify them. Here, we conjecture that if a node w is a super-mediator, removing it would substantially decrease the average influence degree derived based on the underlying information diffusion model. In order to mathematically formulate this notion, we first define the following graph $G \backslash \{w\}$, which is constructed by removing a node w from a directed graph $G = (V, E)$:

$$G\backslash\{w\} = (V\backslash\{w\}, E\backslash\{w\}), \quad E\backslash\{w\} = \{(u,v) \in E | u \neq w, v \neq w\}. \tag{9}$$

Then, we can quantify the model-based super-mediator degree of a node w, denoted by $SMD_{model}(w)$, as the difference in the average influence degree with respect to the node removal, *i.e.*,

$$SMD_{model}(w) = \sum_{v \in V} \varphi(v; G)\kappa(v) - \sum_{v \in V\backslash\{w\}} \varphi(v; G\backslash\{w\})\kappa(v). \tag{10}$$

Here note that, unlike in the case of the data-driven super-mediator degree, we cannot empirically estimate the value of $\kappa(v)$, the probability that node v becomes an initial active node, because observed diffusion samples are not available. Thus, we assume a uniform value for any v, that is, $\kappa(v) = 1/|V|$ to compute SMD_{model} in this paper.

We apply our bond percolation technique [8] to efficiently calculate $SMD_{model}(w)$ 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 [18,32] 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 [18,7]. 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 [8] exploits this relationship. Here, we present the algorithm for calculating $SMD_{model}(w)$ 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 a certain 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 mth bond percolation process and let G_m denote the graph (V, E_m), then for a large M, we can approximate the estimated influence degree $\overline{\varphi}(u; G)$ with a reasonable accuracy as follows:

$$\overline{\varphi}(u; G) = \frac{1}{M}\sum_{m=1}^{M} |R(u; G_m)|, \tag{11}$$

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 Eq. (11) into Eq. (10):

$$SMD_{model}(w) = \frac{1}{M}\sum_{v \in V}\sum_{m=1}^{M} |R(v; G_m)|\kappa(v) - \frac{1}{M}\sum_{v \in V\backslash\{w\}}\sum_{m=1}^{M} |R(v; G_m\backslash\{w\})|\kappa(v). \tag{12}$$

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

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

Then, we can easily see that

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

Namely, for the mth bond percolation process and a fixed node w, we can obtain $R(v; G_m\backslash\{w\})$ for any node $v \in V$ by calculating $R(v; G_m\backslash\{w\})$ 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 $SMD_{model}(w)$ for each node $w \in V$.

```
M1. Perform bond percolation process $M$ times ($m = 1, \ldots, M$);
    M1-1. For the $m$th bond percolation process, calculate $R(v; G_m)$ by applying SCC decomposition.
    M1-2. For each $w \in V$, compute $R^{-}(w; G_m)$, and for each $v \in V$, set $R(v; G_m\backslash\{w\}) = R(v; G_m)$ if $v\neg \in R^{-}(w; G_m)$; otherwise
          calculate $R(v; G_m\backslash\{w\})$ by applying SCC decomposition.
M2. Calculate the super-mediator degree $SMD_{model}(w)$ according to Eq. (12).
```

Here we consider the computational complexity of the above algorithm. Let $H(G)$ be the computation cost of SCC decomposition for a graph $G = (V, E)$, whose computational complexity typically amounts to $O(|E|)$. The step M1-1 requires the computational complexity of $O(H(G_m))$ for SCC decomposition and at most $O(\sum_{v \in V} |R(v; G_m)|)$ to calculate $R(v; G_m)$ for each $v \in V$. On the other hand, let $G_m(v, w)$ be the induced subgraph from $R(v; G_m\backslash\{w\})$; then the step M1-2 requires the computational complexity of $O(\sum_{w \in V}\sum_{v \in R^{-}(w;G_m)} H(G_m(v, w))$ for SCC decomposition and at most $O(\sum_{w \in V}\sum_{v \in R^{-}(w;G_m)} |R(v; G_m\backslash\{w\})|)$ to calculate $R(v; G_m\backslash\{w\})$ for each $v \in R^{-}(w; G_m)$. Here note that $R^{-}(w; G_m)$ is calculated just by reversely following links of the SCC quotient graph. In summary, the computational complexity roughly amounts to $O(MV\langle|R(v)|\rangle^2)$, where $\langle|R(v)|\rangle$ ($\approx \langle|R^{-}(v)|\rangle$) means the average number of reachable nodes.

4. Other centrality measures for comparative analysis

In this section we pick up several other centralities that were recognized as useful measures for node importance. These are divided into two categories, one defined by information diffusion process and the other defined only by network topology. Reverse-influence in the former category is a new measure which we introduced recently [26].

4.1. Influence degree and reverse influence degree

One typical centrality measure based on the information diffusion process is the influence degree $\varphi(v; G)$ for node $v \in V$. In addition to this, in this paper, we introduce a new measure, the *reverse-influence degree* denoted by $\varphi^{-}(v; G)$ that is defined to be the expected number of initial source nodes from which the information reaches the node v at the end of information diffusion. Formally, the reverse-influence degree $\varphi^{-}(v; G)$ for node v is defined as follows:

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

where $\varphi(u, v; G)$ is the probability that the node v becomes active when u is an information source node. Then, it is noted that $\varphi(v; G)$ can be calculated by $\sum_{u \in V} \varphi(v, u; G)$.

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

$$G^{-} = (V, E^{-}), \quad E^{-} = \{(v, u) | (u, v) \in E\}. \tag{13}$$

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.*,

$$\varphi^{-}(v; G) = \varphi(v; G^{-}). \tag{14}$$

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 mth bond percolation process in Section 3. Noting that $R(u, v; G_m) = R(v, u; G_m^{-})$, it is straightforward to show that Eq. (14) holds as shown below:

$$\overline{\varphi}^{-}(v; G) = \frac{1}{M} \sum_{m=1}^{M} \sum_{u \in V} R(u, v; G_m) = \frac{1}{M} \sum_{m=1}^{M} \sum_{u \in V} R(v, u; G_m^{-}) = \overline{\varphi}(v; G^{-}), \tag{15}$$

where G_m^{-} is the reverse graph of G_m.

4.2. Conventional centralities

Next, we show the three well-known conventional centrality measures based only on the network topology, *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.*,

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

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

$$betw(v) = \sum_{u \in V} \sum_{w \in V} \frac{spath_{u,w}(v)}{spath_{u,w}},$$

where $spath_{u,w}$ is the total number of the shortest paths between node u and node v and $spath_{u,w}(v)$ is the number of the shortest paths between node u and node v that passes through node v. Here, efficiently computing $betw(v)$ is still an active research subject [3].

We consider identifying super-mediators by ranking the nodes in decreasing order with respect to a respective centrality measure in our experiments. We refer to the identification methods by the corresponding 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. 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 [15]. 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. [12]. 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. [12]. The Wikipedia network is a bidirectional network having 9481 nodes and 245,044 directed links.

Below we explain the parameter settings of the IC model. We first assume a generative model according to the beta distribution with a mean μ 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 occupation probability of the bond percolation process over each link reduces to μ. Actually, this formulation is equivalent to assigning a uniform value μ to the diffusion probability $p_{v,w}$ for any link $(v, w) \in E$, that is, $p_{v,w} = \mu$. According to [7], we set the value of μ 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 $0 < r \leq 1$.

Under these parameter settings of the IC model, the parameter M to estimate the expectation is set to 10,000 for all experiments for computing the model-driven super-mediator degree SMD_{model}. Similarly, we generated 10,000 diffusion samples for each node $v \in V$ by conducting information diffusion simulations 10,000 times to calculate the data-driven super-mediator degree SMD_{data}, which means $M(v) = 10,000$ for every node v. Note that, in this case, $\kappa(v)$ for computing SMD_{data}, the probability that the node v becomes an information source node, which is estimated by $\kappa(v) = M(v)/\sum_{v \in V} M(v)$, coincides with the one for computing SMD_{model} mentioned above, *i.e.*, $\kappa(v) = 1/|V|$.

5.2. Analysis of discrimination capability

First, we investigated the distributions of the four centrality measures based on the information diffusion process, *i.e.*, the model-driven super-mediator degree SMD_{model}, data-driven super-mediator degree SMD_{data}, influence degree φ, and reverse-influence degree φ^{-} in the Enron network to see how much they differ to each other in characterizing importance of each node. In Fig. 2, the values of “ratio to the maximum value” in each degree are depicted as a function of node rank. Note that rank of each node is different for each degree. It is evident that the curve for the model-driven super-mediator degree SMD_{model} is quite different from the other three, while the curves for these three are similar to each other and almost flat for top ranked nodes. The one for the influence degree maintains a relatively high ratio close to 1.0 up to approximately top 300 nodes and those for the other two degree are very close to 1.0 up to approximately top 1000 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 1000 nodes for SMD_{model}.

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 model-driven super-mediator degree among the top 3 email accounts (nodes), but the difference is not clear for the other three degree values, especially the data-driven super-mediator degree and reverse-influence degree. In addition, the top 3 ranked model-driven super-mediators are different from the top 3 ranked nodes for the other three. It is notable that “Jeffrey Skilling” (the top ranked) and “Kenneth Lay” (the second ranked) in the model-driven 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 the other three. kimura “Jeffrey Richter”, the second ranked in the reverse-influence degree, is known as a trader of Enron, but is in a lower position than the former twoexecutives in the Enron company. These observations suggest the model-driven super-mediator degree can be a promising measure *i.e.*, it identified people of powerful positions in the Enron company. In fact, there are many nodes that have the same high score for each of the data-driven super-mediator degree, influence degree and reverse-influence degree, and the node returned based on ranking with the same score does not make much sense (our algorithm returns the first one on the list). However, this does not mean that the other three are useless as shown below.

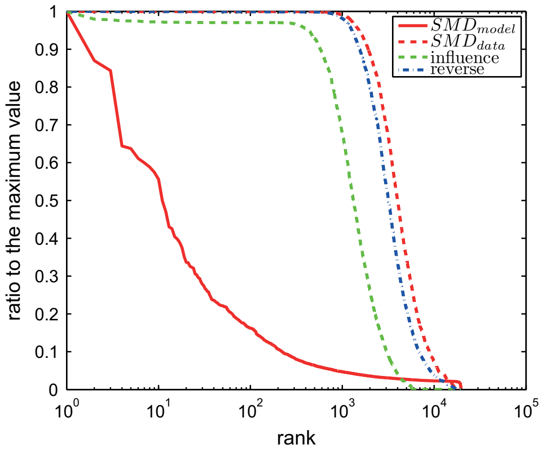


Fig. 2. Distribution of the model-driven super-mediator/data-driven super-mediator/influence/reverse-influence degree for the Enron network in case of $r = 1.0$.

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

Rank	Account name (ID: ratio to the maximum degree value)			
	SMD_{model}	SMD_{data}	Influence	Reverse-influence
1	jeff.skilling (642: 1.000)	dana.davis (485: 1.000)	bob.ambrocik (16,734: 1.000)	tom.alonso (5510: 1.000)
2	kenneth.lay (471: 0.870)	kate.symes (7623: 0.999)	technology.enron (17,219: 0.979)	jeff.richter (1768: 0.999)
3	sally.beck (535: 0.843)	don.baughman (4854: 0.999)	outlook.team (10,779: 0.978)	chris.mallory (5933: 0.999)

Next, we investigated how the top 3 model-driven super-mediators for the Enron network rank in terms of the other three degree values. The results are summarized in Table 2. It is found that these model-driven super-mediators are ranked relatively high, at least they are in the top 5% nodes in all of the other three measures. It is noted that the ranks by the data-driven super-mediator degree are higher than those by the other two degree measures. This implies that the data-driven super-mediator degree can be a better approximation of the model-driven super-mediator degree compared to the other remaining two (influence and reverse-influence) when the network structure and/or information diffusion model are not available.

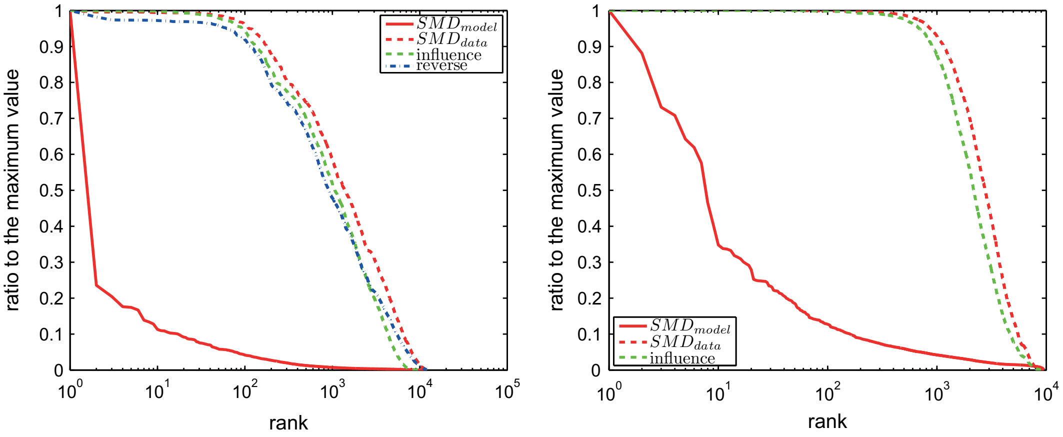
We observed the same tendencies for the Blog and Wikipedia networks. Here, we only show the distributions of the four measures for these networks in case of $r = 1.0$ in Fig. 3(a) and (b), respectively. Note that since the Wikipedia network is bidirectional, the reverse-influence degree is the same as the influence degree, so it is not shown in Fig. 3(b).

Using the Enron network, we further analyzed the properties of model-driven super-mediators. First, we investigated the effect of diffusion probability by varying the value of r. Fig. 4(a) and (b), and Tables 3 and 4 show the results for the case of $r = 0.5$ and $r = 0.25$, respectively. Here, each table shows the ranks and the values of “ratio to the maximum value” of the top 3 model-based super mediators and their corresponding ranks and the values in terms of the other three measures (data-driven super-mediators, influence, reverse-influence). It is obvious that the distribution curves shown in Fig. 4(a) and (b) share the same tendency as is shown in Fig. 2. The notable difference is that the flat region shrinks for the data-driven super-mediator degree, influence degree and reverse-influence degree as the diffusion probability becomes smaller. This is because both $\varphi(v; G)$ and $\varphi^{-}(v; G)$ 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 $r = 1.0$. The data-driven super-mediator measure remains to be the second best, but it gets closer to the reverse-influence degree measure as r becomes smaller. Further we notice that all the values for the 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 third ranks for $r = 0.25$, but overall both the influence degree and reverse-influence degree are high for the high ranked model-driven super-mediators. Indeed, these nodes are within the top 3% nodes at $r = 0.5$, and the top 1% nodes at $r = 0.25$ in both the influence degree ranking and reverse-influence degree ranking.

We can conclude that the model-driven super-mediator degree can characterize each individual node by far clearly and is the most discriminative among these measures in that different ranks have different scores. This is the most important feature of the model-driven super-mediator measure.

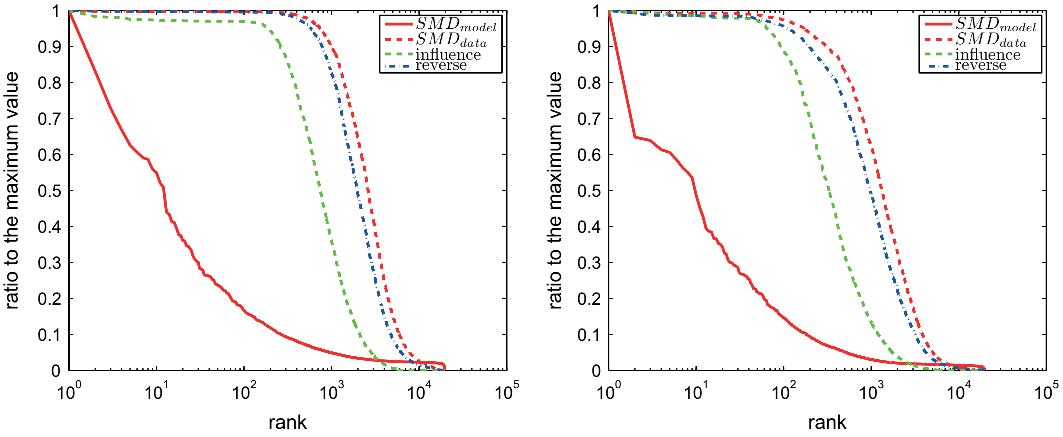
Table 2
The rank of the top 3 model-driven super-mediators for the Enron network in the data-driven super-mediator, influence, and reverse-influence degree ranking ($r = 1.0$).

ID	Rank (ratio to the maximum degree value)			
	SMD_{model}	SMD_{data}	Influence	Reverse-influence
642	1 (1.000)	352 (0.999)	441 (0.947)	217 (0.999)
471	2 (0.870)	38 (0.999)	122 (0.970)	374 (0.999)
535	3 (0.843)	41 (0.999)	126 (0.970)	377 (0.999)



(a) Blog network (b) Wikipedia network

Fig. 3. Distribution of the model-driven super-mediator/data-driven 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. 4. Distribution of the super-mediator/influence/reverse-influence degree for the Enron network in cases of $r = 0.5$ and $r = 0.25$.

5.3. Properties of super-mediators

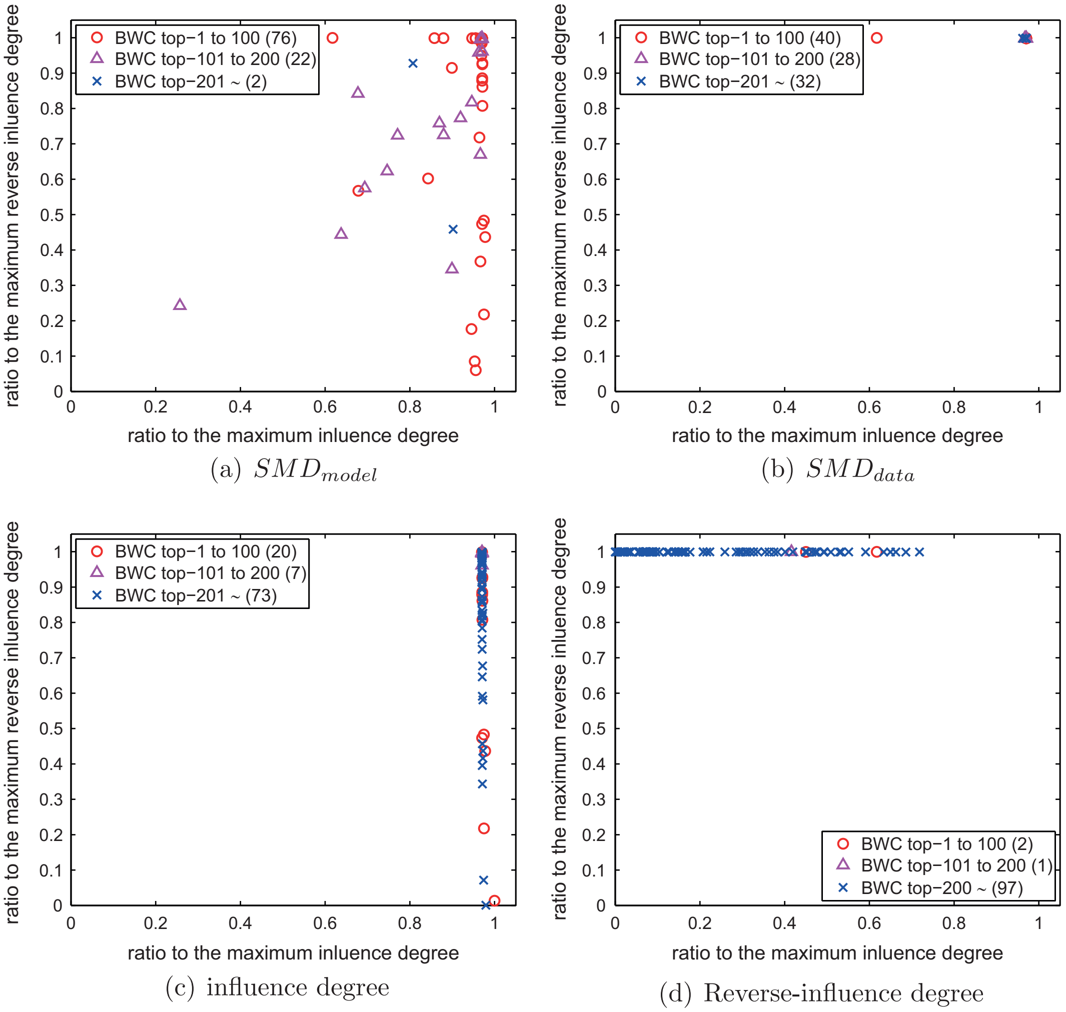
Because the curves are flat for the data-driven super-mediator degree, influence degree and reverse-influence degree, there are many nodes that have similar high values in these three measures. Thus, to further investigate the relationships among these measures, we plot the top 100 nodes in each degree using their influence and reverse-influence degrees in Fig. 5. Here, we focused on the top 100 nodes because the values of "ratio to the maximum value" of nodes after the top 100 is less than 0.2 for the model-driven super-mediator degree. We further indicate how these nodes are ranked in the betweenness centrality (BWC): ○ for those ranked in the top 1 to 100, △ in the top 101 to 200, and × below the top 200,

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

ID	Rank (ratio to the maximum degree value)			
	SMD_{model}	SMD_{data}	Influence	Reverse-influence
535	1 (1.000)	29 (0.997)	66 (0.970)	94 (0.996)
471	2 (0.831)	103 (0.997)	114 (0.969)	128 (0.995)
642	3 (0.728)	309 (0.992)	426 (0.742)	341 (0.985)

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

ID	Rank (ratio to the maximum degree value)			
	SMD_{model}	SMD_{data}	Influence	Reverse-influence
535	1 (1.000)	45 (0.989)	52 (0.970)	46 (0.979)
6	2 (0.648)	1 (1.000)	185 (0.734)	1 (1.000)
471	3 (0.639)	148 (0.959)	154 (0.799)	144 (0.931)



(a) SMD_{model}

(b) SMD_{data}

(c) influence degree

(d) Reverse-influence degree

Fig. 5. Relation between the influence and reverse-influence degrees of the top 100 nodes in each of model-driven super-mediator (SMD_{model}), data-driven super-mediator (SMD_{data}), influence degree, and reverse-influence degree for the Enron network.

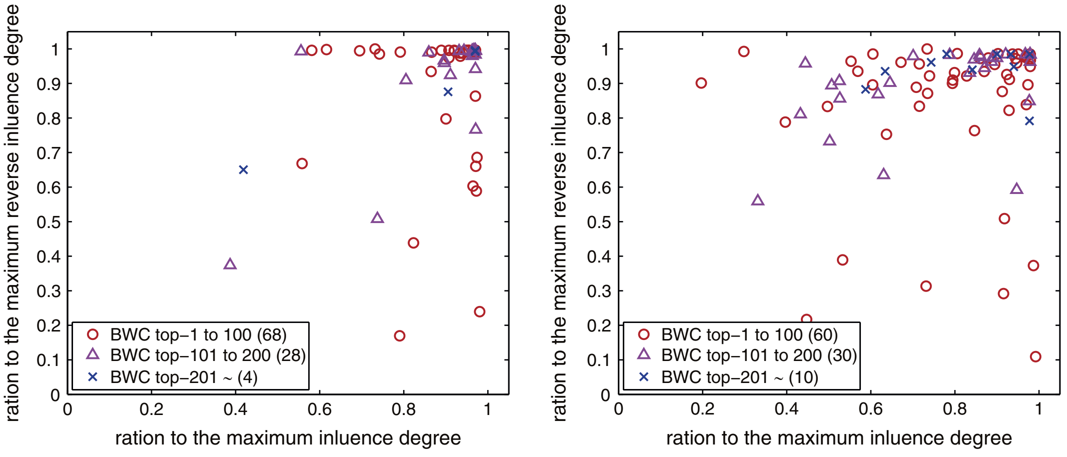
respectively, in Fig. 5. This is because it is natural to think that the betweenness centrality is closely related to the super-mediator degree (both model-driven and data-driven) as the node with high betweenness value can mediate many other nodes in a network, and thus is thought to play an important role for information diffusion. The value in parentheses in the legend of Fig. 5 is the number of nodes that fall into the corresponding range.

From Fig. 5, we can see each centrality degree has quite different tendency. The top 100 model-driven super-mediators shown in Fig. 5(a) have high values for either the influence degree or the reverse-influence degree, but not necessarily for both. 76% of nodes is ranked in the top 100 and 98% of nodes in the top 200 in the betweenness centrality. Only two nodes are below the top 200. Betweenenss centrality is high when either one of the influence degree or reverse-influence degree is high, and it cannot be small when both are small for the model-driven super-mediator degree to be high. Thus, the model-driven super-mediators are closely related to the betweenness centrality.

Next, from Fig. 5(b), we find that almost all of the top 100 data-driven super-mediators have high scores both in the influence degree and in the reverse-influence degree, which is what we expected in the beginning. However, we also see from Fig. 2 that the flat part of the curve of data-driven super-mediators is wider than the other two measures (influence and reverse-influence) and these two fall down earlier. This means that there are other factors that make the scores high. If we consider the top 1,000 to 3,000, the plots in Fig. 5(b) are more scattered. We further note that 68% of nodes is ranked in the top 200 nodes in the betweenness centrality (they overlap), which implies that the data-driven super-mediators are also related to the betweenness centrality, but not so closely compared to the 98% for the model-driven super-mediator degree. From Fig. 5(c) and (d), we find that the top 100 nodes in the influence degree do not necessarily have high scores in the reverse-influence degree, and vice versa. This also means that many of the nodes that are within the top 100 in terms of the data-driven super-mediator ranking (Fig. 5(b)) are below the top 100 in terms of both the influence degree ranking and the reverse-influence degree ranking. It is noted that in both figures, many nodes are outside of the top 200 nodes in the betweenness centrality, which means there is no clear relation between these two degree values and the betweenness centrality. From these results, we can say that the model-driven super-mediator degree has different characteristics compared to the data-driven super-mediator degree, influence degree, and reverse-influence degree, and what is for sure is that it is closely related to the betweenness centrality, and is most discriminative. The last point is the most important. As stated earlier, there are so many nodes that have practically the same high score in the other three measures, nodes returned mechanically by ranking for these measures do not make much sense.

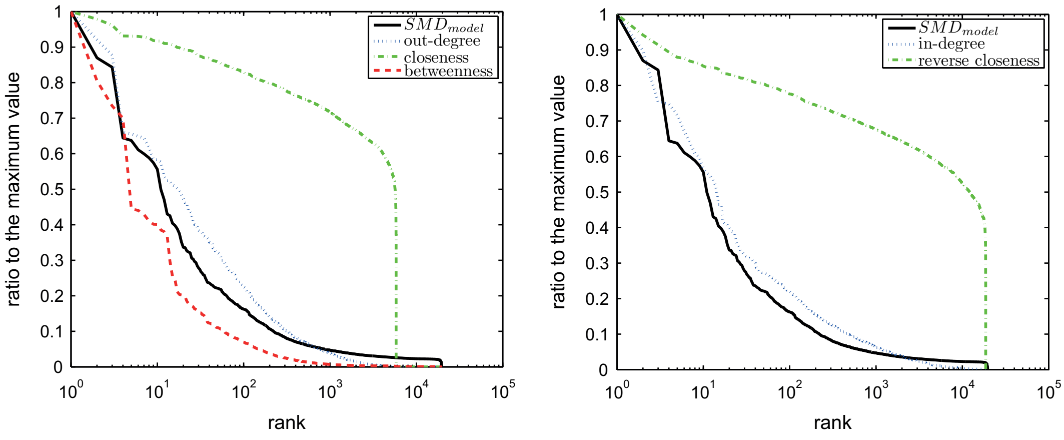
Here, again, we investigated the relationship among the four measures in Fig. 6 , in which, similarly to Fig. 5(a), the top 100 model-driven super-mediators are plotted using their influence degree and reverse-influence degree, by varying the value of r. From the results for the case of $r = 0.5$ and $r = 0.25$, respectively, it is found that the variance of both the influence degree and reverse-influence degree becomes larger as the diffusion probability becomes smaller, while the close relationship between the model-driven super-mediator degree and the betweenness centrality is kept maintained for a wide range of the diffusion probability.

As mentioned earlier, we originally anticipated that super-mediators defined either way (data-driven or model-driven) have always high scores for both influence degree and reverse-influence degree. However, there is a difference between data-driven super-mediators and model-driven super-mediators: model-driven super-mediators have high scores for either one of influence degree or reverse-influence degree, while data-driven super-mediators have always high scores for both of them. Thus, data-driven super-mediator is useful to identify nodes with both high influence degree and high reverse influence degree without knowing network topology, provided that the information diffusion data are available.



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

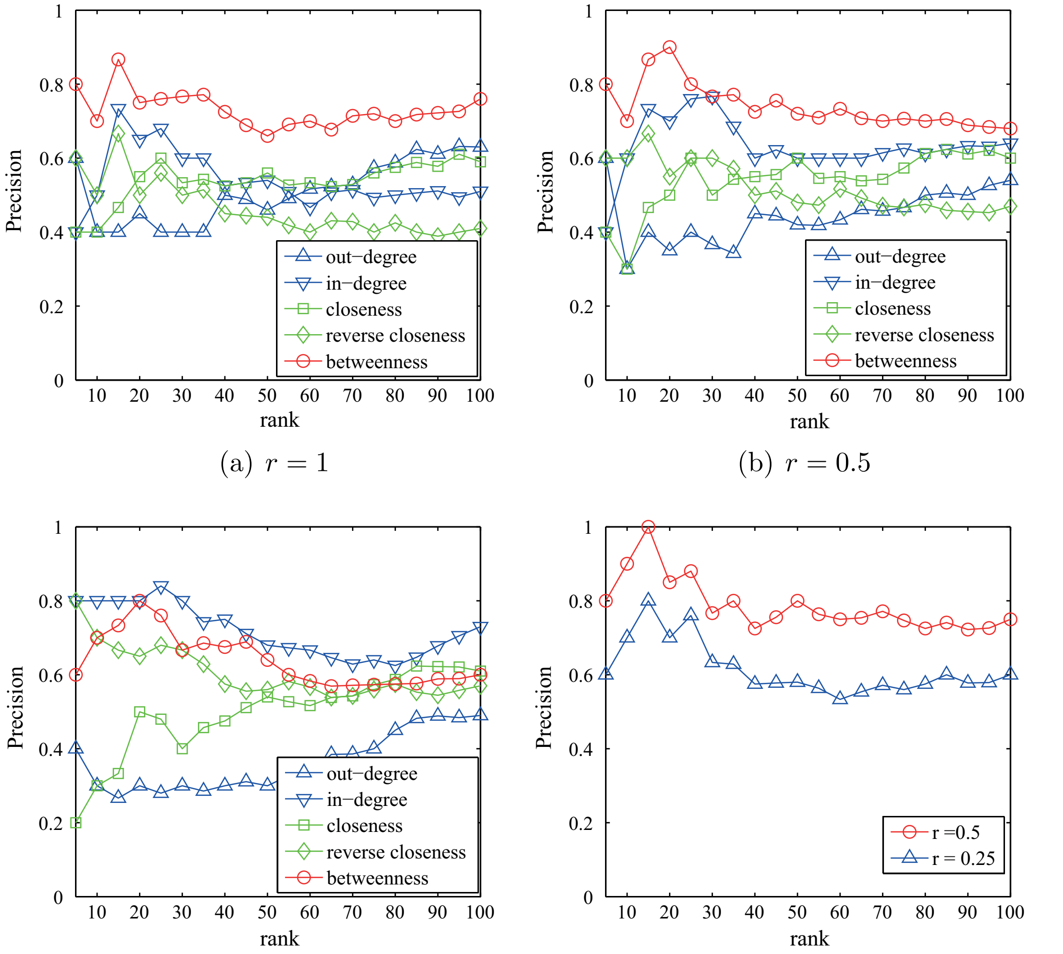
Fig. 6. Relation between the influence and reverse-influence degrees of the top 100 model-driven super-mediators for the Enron network in cases of $r = 0.5$ and 0.25.



(a) Out-degree, closeness, and betweenness

(b) In-degree and reverse closeness

Fig. 7. Distributions of conventional centrality measures for the Enron network.



(a) $r = 1$

(b) $r = 0.5$

(c) $r = 0.25$

(d) Comparison of super-mediators with r

Fig. 8. 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	10,779	203	10,779	203	642
3	535	535	451	684	471

5.4. *Comparison with other centralities*

In this subsection we compare the model-driven super-mediator with the other centralities discussed in 4.2. Fig. 7 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 model-driven super-mediator degree and, thus, these three are as discriminative as the model-driven super mediator degree, while the distributions of closeness $close(v)$ and reverse closeness $close^{-}(v)$ are similar to the distributions of the data-driven super-mediator degree, the influence degree and the reverse-influence degree, but their discriminability is slightly better.

Next, we focused on the top 100 nodes again, and examined the similarity between the model-driven 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 precision $P(k)$ defined by

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

Fig. 8(a), (b) and (c) shows the results for the cases of $r = 1.0, r = 0.5$ and $r = 0.25$, respectively. Fig. 8(d) displays the similarities between the model-driven super-mediator ranking for the case of $r = 1.0$ and that of $r = 0.5$ and $r = 0.25$. We notice that the model-driven super-mediators depend on the value of the diffusion probability from Fig. 8(d). We further notice that the betweenness centrality is the best when the diffusion probability is large (Fig. 8(a) and (b)) and the in-degree centrality becomes better when the diffusion probability gets smaller (Fig. 8(c)). 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, while when the diffusion probability is small, many of the diffusion sequences are short, in which case node degree plays a key role. 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. Finally, we added an extra experiment in which we examined the correlation among the top 100 nodes by PCA (principal components analysis) with these six measures, *i.e.*, $SMD_{model}(w), deg^{+}(v), deg^{-}(v), close(v), close^{-}(v)$, and $betw(v)$. Fig. 9 plots the first two principal components constructed from the 6-dimensional vectors for these nodes. We observe that the nodes from about the top 20 exhibit high correlations, but as the rank goes down nodes loose correlations. Nodes below rank 60 are scattered widely. In summary model-driven

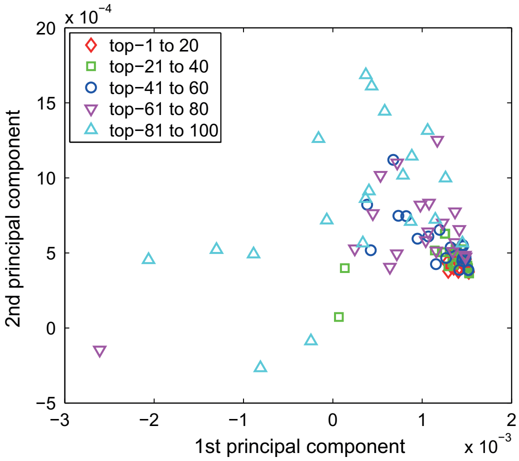


Fig. 9. Correlation analysis of conventional centrality and super-mediator degree for the Enron network.

super-mediators are closely correlated with betweenness centrality, but the strength of the correlation depends on the value of diffusion probability. When the value is small, model-driven super-mediators are closely correlated with in-degree centrality. No single conventional 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.

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 would reduce the information spread if they are removed from the network. 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 out to many other nodes. We quantified the degree of importance as super-mediator degree and formulated this as the difference of the average influence degree with respect to node removal, *i.e.*, 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 definition requires use of information diffusion model to estimate influence degree of each node, and thus it is “model-driven”. Estimating influence degree is computationally hard because it is defined to be the expected number of active nodes at the end of information diffusion process. 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. The other definition of super-mediators, which we defined earlier and is more empirical, is that super-mediators are those nodes that appear frequently in long diffusion sequences but much less frequently in short diffusion sequences. This definition does not require any model but does require abundant information diffusion data and thus it is “data-driven”.

We attempted to characterize the property of super-mediators from various perspectives: how the resulting super-mediators are different between these two definitions and which is better and more reasonable, how super-mediators are compared with nodes identified by other centralities, *e.g.*, betweenness, degree, closeness, etc., how super mediators are different from the solution of well-studied influence maximization problem, and the solution of reverse-influence maximization problem, which is a new measure of importance born from this study, *i.e.*, the expected number of initial source nodes from which the information reaches a node at the end of information diffusion. In fact reverse-influence degree of a node in a graph is the same as the influence degree of the same node of the reverse graph in which the edge direction is reversed for all edges. Thus it can be efficiently computed by the same bond percolation technique.

We have performed extensive experiments using the IC model as a model for information diffusion and applying it to three real world networks (Enron, Blog and Wikipedia). We conclude that among the four definitions of expressing node importance that explicitly take account of stochastic information diffusion process, the definition of model-driven super-mediator is the most natural and there is a clear difference in discrimination ability compared to other measures (data-driven super-mediator degree, influence degree and reverse-influence degree). Data-driven super-mediator is the second best in light of the returned results in the Enron network. It is a useful measure if network topology is unknown and no model can be applicable, but it assumes that the information diffusion data are available. Rank–score curves for the data-driven super-mediator degree, influence degree and reverse-influence degree are all flat for the top ranked nodes and they are much less discriminative than the model-driven super-mediator degree, but the latter two are slightly better than the first. We originally anticipated that super-mediators defined either way (data-driven or model-driven) have always high scores for both influence degree and reverse-influence degree. This holds only for the data-driven super-mediators and does not necessarily hold for the model-driven super-mediators. Either influence degree or reverse-influence degree has high score for the latter. We also found that those super mediators for which both influence degree and reverse-influence degree are small have always large betweenness centrality values. Model-based super-mediators are closely correlated with betweenness centrality. However, the strength of the correlation depends on the value of diffusion probability. As the value goes smaller, in-degree centrality becomes a better measure. In short, no single conventional centrality measure which are defined solely by network topology works equally well for a wide range of the diffusion probability. The model-driven super-mediators we proposed in this paper explicitly consider information diffusion process, but is different from influence degree and reverse-influence degree, and serve as a new useful centrality measure that can be added to the existing pool.

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. We also plan to test out whether what we found out holds for other type of information diffusion models, *e.g.*, LT model.

Acknowledgments

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research - United States, under Grant No. AOARD-13-4042, and JSPS Grant-in-Aid for Scientific Research (C) (No. 26330352), Japan.

References

[1] E. Bakshy, B. Karrer, L.A. Adamic, Social influence and the diffusion of user-created content, in: Proceedings of the 10th ACM conference on Electronic Commerce (EC'09), 2009, pp. 325–334.

[2] E. Bakshy, J. Hofman, W. Mason, D. Watts, Everyone's an influencer: quantifying influences on twitter, in: Proceedings of the 4th International Conference on Web Search and Data Mining (WSDM'11), 2011, pp. 65–74.

[3] M.H. Chehreghani, An efficient algorithm for approximate betweenness centrality computation, Comput. J. 57 (2014) 1371–1382.

[4] P. Dow, L. Adamic, A. Friggeri, The anatomy of large facebook cascades, in: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM'13), 2013.

[5] J. Goldenberg, B. Libai, E. Muller, Talk of the network: a complex systems look at the underlying process of word-of-mouth, Market. Lett. 12 (2001) 211–223.

[6] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, Information diffusion through blogspace, SIGKDD Explor. 6 (2004) 43–52.

[7] D. Kempe, J. Kleinberg, E. Tardos, 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'03), 2003, pp. 137–146.

[8] M. Kimura, K. Saito, R. Nakano, Extracting influential nodes for information diffusion on a social network, in: Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI'07), 2007, pp. 1371–1376.

[9] M. Kimura, K. Saito, H. Motoda, Efficient estimation of influence functions for SIS model on social networks, in: Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09), 2009.

[10] M. Kimura, K. Saito, H. Motoda, Blocking links to minimize contamination spread in a social network, ACM Trans. Knowl. Discov. Data 3 (2009) 9:1–9:23.

[11] M. Kimura, K. Saito, R. Nakano, H. Motoda, 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 (SBP'09), 2009, pp. 138–145.

[12] M. Kimura, K. Saito, R. Nakano, H. Motoda, Extracting influential nodes on a social network for information diffusion, Data Min. Knowl. Discov. 20 (2010) 70–97.

[13] M. Kimura, K. Saito, K. Ohara, H. Motoda, Efficient analysis of node influence based on SIR model over huge complex networks, in: Proceedings of the 2014 International Conference on Data Science and Advanced Analytics (DSAA'14), 2014, pp. 216–222.

[14] J. Kleinberg, The convergence of social and technological networks, Commun. ACM 51 (11) (2008) 66–72.

[15] B. Klimt, Y. Yang, The enron corpus: a new dataset for email classification research, in: Proceedings of the 2004 European Conference on Machine Learning (ECML'04), 2004, pp. 217–226.

[16] J. Leskovec, L.A. Adamic, B.A. Huberman, The dynamics of viral marketing, in: Proceedings of the 7th ACM Conference on Electronic Commerce (EC'06), 2006, pp. 228–237.

[17] M.E.J. Newman, S. Forrest, J. Balthrop, Email networks and the spread of computer viruses, Phys. Rev. E 66 (2002) 035101.

[18] M.E.J. Newman, The structure and function of complex networks, SIAM Rev. 45 (2003) 167–256.

[19] M. Richardson, P. Domingos, Mining knowledge-sharing sites for viral marketing, in: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02), 2002, pp. 61–70.

[20] D. Romero, B. Meeder, J. Kleinberg, 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 (WWW'11), 2011, pp. 695–704.

[21] K. Saito, M. Kimura, H. Motoda, Discovering influential nodes for sis models in social networks, in: Proceedings of the Twelfth International Conference of Discovery Science (DS'09), LNAI 5808, Springer, 2009, pp. 302–316.

[22] K. Saito, M. Kimura, K. Ohara, H. Motoda, Learning continuous-time information diffusion model for social behavioral data analysis, in: Proceedings of the 1st Asian Conference on Machine Learning (ACML'09), LNAI 5828, 2009, pp. 322–337.

[23] K. Saito, M. Kimura, K. Ohara, H. Motoda, 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 (SBP'10), LNCS 6007, 2010, pp. 149–158.

[24] K. Saito, M. Kimura, K. Ohara, H. Motoda, Discovery of super-mediators of information diffusion in social networks, in: Proceedings of the Thirteenth International Conference of Discovery Science (DS'10), LNAI 6332, Springer, 2010, pp. 144–158.

[25] K. Saito, M. Kimura, K. Ohara, H. Motoda, 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'13), LNCS 7812, 2013, pp. 359–367.

[26] K. Saito, M. Kimura, K. Ohara, H. Motoda, Identifying super-mediators of information diffusion in social networks, in: Proceedings of the Sixteenth International Conference of Discovery Science (DS'13), LNAI 8140, Springer, 2013, pp. 170–184.

[27] D. Sheldon, B. Dilkina, A. Elmachtoub, R. Finseth, A. Sabharwal, J. Conrad, C. Gomes, D. Shmoys, W. Allen, O. Amundsen, W. Vaughan, Maximizing the spread of cascades using network design, in: Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI'10), 2010, pp. 517–526.

[28] G.V. Steeg, R. Ghosh, K. Lerman, What stops social epidemics? in: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11), 2011, pp. 377–384.

[29] J. Tang, X. Hu, H. Liu, Social recommendation: a review, Soc. Netw. Anal. Min. 3 (2013) 1113–1133.

[30] C. Wang, W. Chen, Y. Wang, Scalable influence maximization for independent cascade model in large-scale social networks, Data Min. Knowl. Discov. 25 (2012) 545–576.

[31] D.J. Watts, A simple model of global cascades on random networks, Proc. Nat. Acad. Sci. USA 99 (2002) 5766–5771.

[32] D.J. Watts, P.S. Dodds, Influence, networks, and public opinion formation, J. Consum. Res. 34 (2007) 441–458.