Discovery of 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
52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan
k-saito@u-shizuoka-ken.ac.jp
[2] Department of Electronics and Informatics, Ryukoku University
Otsu 520-2194, Japan
kimura@rins.ryukoku.ac.jp
[3] Department of Integrated Information Technology, Aoyama Gakuin University
Kanagawa 229-8558, Japan
ohara@it.aoyama.ac.jp
[4] Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
motoda@ar.sanken.osaka-u.ac.jp

Abstract. We address the problem of discovering a different kind of influential nodes, which we call "super-mediator", i.e. those nodes which play an important role to pass the information to other nodes, and propose a method for discovering super-mediators from information diffusion samples without using a network structure. We divide the diffusion sequences 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 sequences by the F-measure. We apply this measure to the information diffusion samples generated by two real networks, identify and rank the super-mediator nodes. We show that the high ranked super-mediators are also the high ranked influential nodes when the diffusion probability is large, i.e. the influential nodes also play a role of super-mediator for the other source nodes, and interestingly enough that when the high ranked super-mediators are different from the top ranked influential nodes, which is the case when the diffusion probability is small, those super-mediators become the high ranked influential nodes when the diffusion probability becomes larger. This finding will be useful to predict the influential nodes for the unexperienced spread of new information, e.g. spread of new acute contagion.

1 Introduction

There have been tremendous interests in the phenomenon of influence that members of social network can exert on other members and how the information propagates through the network. Social networks (both real and virtual) are now recognized as an important medium for the spread of information. A variety of information that includes news, innovation, hot topics, ideas, opinions and even malicious rumors, propagates in the form of so-called "word-of-mouth" communications. Accordingly, a considerable amount of studies has been made for the last decade [1–20].

Among them, widely used information diffusion models are the *independent cascade (IC)* [1, 8, 13] and the *linear threshold (LT)* [4, 5] and their variants [14, 15, 6, 16–18]. These two models focus on different information diffusion aspects. The IC model is sender-centered and each active node *independently* influences its inactive neighbors with given diffusion probabilities. The LT model is receiver-centered and a node is influenced by its active neighbors if their total weight exceeds the threshold for the node. Which model is more appropriate depends on the situation and selecting the appropriate one is not easy [18].

The major interests in the above studies are finding influential nodes, i.e. finding nodes that play an important role of spreading information as much as possible. This problem is called *influence maximization problem* [8, 10]. The node influence can only be defined as the expected number of active nodes (nodes that have become influenced due to information diffusion) because the diffusion phenomenon is stochastic, and estimating the node influence efficiently is still an open problem. Under this situation, solving an optimal solution, i.e. finding a subset of nodes of size K that maximizes the expected influence degree with K as a parameter, faces with combinatorial explosion problem and, thus, much of the efforts has been directed to finding algorithms to efficiently estimate the expected influence and solve this optimization problem. For the latter, a natural solution is to use a greedy algorithm at the expense of optimality. Fortunately, the expected influence degree is submodular, i.e. its marginal gain diminishes as the size K becomes larger, and the greedy solution has a lower bound which is 63% of the true optimal solution [8]. Various techniques to reduce the computational cost have been attempted including bond percolation [10] and pruning [14] for the former, and lazy evaluation [21], burnout [15] and heuristics [22] for the latter.

Expected influence degree is approximated by the empirical mean of the influence degree of many independent information diffusion simulations, and by default it has been assumed that the degree distribution is Gaussian. However, we noticed that this assumption is not necessarily true, which motivated to initiate this work. In this paper, we address the problem of discovering a different kind of influential nodes, which we call ”super-mediator”, i.e. those nodes which play an important role in passing the information to other nodes, try to characterize such nodes, and propose a method for discovering super-mediator nodes from information diffusion sequences (samples) without using a network structure. We 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 sequences by the F-measure (more in subsection 3.2).

We tested our assumption of existence of super-mediators using two real networks[1] and investigated the utility of the F-measure. As before, we assume that information diffusion follows either the independent cascade (IC) model or the linear threshold (LT) model. We first analyze the distribution of influence degree averaged over all the initial nodes[2] based on the above diffusion models, and empirically show that it becomes a

[1] Note that we use these networks only to generate the diffusion sample data, and thus are not using the network structure for the analyses.

[2] Each node generates one distribution, which is approximated by running diffusion simulation many times and counting the number of active nodes at the end of simulation.

power-law like distribution for the LT model, but it becomes a mixture of two distributions (power-law like distribution and lognormal like distributions) for the IC model. Based on this observation, we evaluated our super-mediator discovery method by focusing on the IC model. It is reasonable to think that the super mediators themselves are the influential nodes, and we show empirically that the high ranked super-mediators are indeed the high ranked influential nodes, i.e. the influential nodes also play a role of super-mediator for the other source nodes, but this is true only when the diffusion probability is large. What we found more interesting is that when the high ranked super-mediators are different from the top ranked influential nodes, which is the case when the diffusion probability is small, those super-mediators become the high ranked influential nodes when the diffusion probability becomes larger. We think that this finding is useful to predict the influential nodes for the unexperienced spread of new information from the known experience, e.g. spread of new acute contagion from the spread of known moderate contagion for which there are abundant data.

The paper is organized as follows. We start with the brief explanation of the two information diffusion models (IC and LT) and the definition of influence degree in section 2, and then describe the discovery method based on the likelihood maximization and F-measure in section 3. Experimental results are detailed in section 4 together with some discussion. We end this paper by summarizing the conclusion in section 5.

2 Information Diffusion Models

We mathematically model the spread of information through a directed network $G = (V, E)$ without self-links, where V and E ($\subset V \times V$) stand for the sets of all the nodes and links, respectively. For each node v in the network G, we denote $F(v)$ as a set of child nodes of v, i.e. $F(v) = \{w; (v, w) \in E\}$. Similarly, we denote $B(v)$ as a set of parent nodes of v, i.e. $B(v) = \{u; (u, v) \in E\}$. We call nodes *active* if they have been influenced with the information. In the following models, we assume that nodes can switch their states only from inactive to active, but not the other way around, and that, given an initial active node set H, only the nodes in H are active at an initial time.

2.1 Independent Cascade Model

We recall the definition of the IC model according to [8]. In the IC model, we specify a real value $p_{u,v}$ with $0 < p_{u,v} < 1$ for each link (u, v) in advance. Here $p_{u,v}$ is referred to as the *diffusion probability* through link (u, v). The diffusion process unfolds in discrete time-steps $t \geq 0$, and proceeds from a given initial active set H in the following way. When a node u 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 become active at time-step t, then their activation attempts are sequenced in an arbitrary order, but all performed at time-step t. Whether or not u succeeds, it cannot make any further attempts to activate v in subsequent rounds. The process terminates if no more activations are possible.

2.2 Linear Threshold Model

In the LT model, for every node $v \in V$, we specify a *weight* ($\omega_{u,v} > 0$) from its parent node u in advance such that $\sum_{u \in B(v)} \omega_{u,v} \leq 1$. The diffusion process from a given initial active set H proceeds according to the following randomized rule. First, for any node $v \in V$, a *threshold* θ_v is chosen uniformly at random from the interval $[0, 1]$. At time-step t, an inactive node v is influenced by each of its active parent nodes, u, according to weight $\omega_{u,v}$. If the total weight from active parent nodes of v is no less than θ_v, that is, $\sum_{u \in B_t(v)} \omega_{u,v} \geq \theta_v$, then v will become active at time-step $t + 1$. Here, $B_t(v)$ stands for the set of all the parent nodes of v that are active at time-step t. The process terminates if no more activations are possible.

2.3 Influence Degree

For both models on G, we consider information diffusion from an initially activated node v, i.e. $H = \{v\}$. Let $\varphi(v; G)$ denote the number of active nodes at the end of the random process for either the IC or the LT model on G. Note that $\varphi(v; G)$ is a random variable. We refer to $\varphi(v; G)$ as the *influence degree* of node v on G. Let $\mathcal{E}(v; G)$ denote the expected number of $\varphi(v; G)$. We call $\mathcal{E}(v; G)$ the *expected influence degree* of node v on G. In theory we can simply estimate $\mathcal{E}$ by the simulations based on either the IC or the LT model in the following way. First, a sufficiently large positive integer M is specified. Then, the diffusion process of either the IC or the LT model is simulated from the initially activated node v, and the number of active nodes at the end of the random process, $\varphi(v; G)$, is calculated. Last, $\mathcal{E}(v; G)$ for the model is estimated as the empirical mean of influence degrees $\varphi(v; G)$ that are obtained from M such simulations.

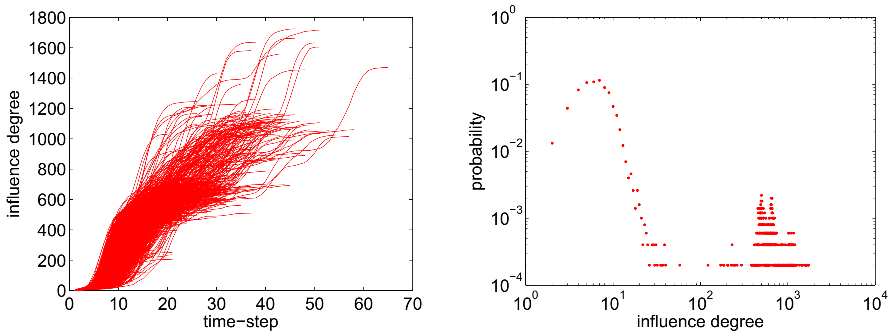
From now on, we use $\varphi(v)$ and $\mathcal{E}(v)$ instead of $\varphi(v; G)$ and $\mathcal{E}(v; G)$, respectively if G is obvious from the context.

3 Discovery Method

3.1 Super-mediator

As mentioned in section 1, we address the problem of discovering a different kind of influential nodes, which we call ”super-mediator”. These are the nodes which appear frequently in long diffusion sequences with many active nodes and less frequently in short diffusion sequences, i.e. those nodes which play an important role to pass the information to other nodes. Figure 1 (a) shows an example of information diffusion samples. In this figure, by independently performing simulations $5,000$ times based on the IC model, we plotted $5,000$ curves for influence degree of a selected information source node with respect to time steps[3]. From this figure, we can observe that 1) due to its stochastic nature, each diffusion sample varies in a quite wide range for each simulation; and 2) some curves clearly exhibit sigmoidal behavior in part, in each of which the influence degree suddenly becomes relatively high during a certain time interval.

[3] The network used to generate these data is the blog network (see subsection 4.1).



(a) Diffusion samples (b) Influence degree distribution

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

In Figure 1 (b), we plotted the distribution of the final influence degree for the above $5,000$ simulations. From this figure, we can observe that there exist a number of 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 several lognormal like distributions. Together with the observation from Figure 1 (a), we conjecture that super-mediators appear as a limited number of active nodes in some lognormal components with relatively 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.

3.2 Clustering of Diffusion samples

Let $\mathcal{S}(v) = \{1, 2, \cdots, M(v)\}$ denote a set of indices with respect to information diffusion samples for an information source node v, i.e. $\{d_1(v), d_2(v), \cdots, d_{M(v)}(v)\}$. Here note that $d_m(v)$ stands for a set of active nodes in the m-th diffusion sample. As described earlier, in order to discover super-mediator nodes, we consider dividing $\mathcal{S}(v)$ into two groups, $\mathcal{S}_1(v)$ and $\mathcal{S}_2(v)$, which are the upper group of samples with relatively high influence degree and the lower group, respectively. Namely, $\mathcal{S}_1(v) \cup \mathcal{S}_2(v) = \mathcal{S}(v)$ and $\min_{m\in\mathcal{S}_1(v)} |d_m(v)| > \max_{m\in\mathcal{S}_2(v)} |d_m(v)|$. Although we can straightforwardly extend our approach in case of k-groups division, we focus ourselves on the simplest case ($k = 2$) because of ease of both evaluation of basic performance and the following derivation. By assuming the independence of each sample drawn from either the upper or the lower group, we can consider the following likelihood function.

$$\mathcal{L}(\mathcal{S}(v); \mathcal{S}_1(v), \Theta) = \prod_{k\in\{1,2\}} \prod_{m\in\mathcal{S}_k(v)} p(m; \theta_k), \qquad (1)$$

where $p(m; \theta_k)$ denotes some probability distribution with the parameter set θ_k for the m-th diffusion sample, and $\Theta = \{\theta_1, \theta_2\}$. If it is assumed that the influence degree distribution consists of lognormal components, we can express $p(m;\theta_k)$ by

$$p(m;\theta_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}|d_m(v)|}\exp\left(-\frac{(\log|d_m(v)| - \mu_k)^2}{2\sigma_k^2}\right), \tag{2}$$

where $\theta_k = \{\mu_k, \sigma_k^2\}$. Then, based on the maximum likelihood estimation, we can identify the optimal upper group $\hat{\mathcal{S}}_1(v)$ by the following equation.

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

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

Below we describe our method for efficiently obtaining $\hat{\mathcal{S}}_1(v)$ by focusing on the case that $p(m;\theta_k)$ is the lognormal distribution defined in Equation (2), although the applicability of the method is not limited to this case. For a candidate upper group $\mathcal{S}_1(v)$, by noting the following equations of the maximum likelihood estimation,

$$\hat{\mu}_k = \frac{1}{|\mathcal{S}_k(v)|}\sum_{m\in\mathcal{S}_k(v)}\log|d_m(v)|, \quad \hat{\sigma_k^2} = \frac{1}{|\mathcal{S}_k(v)|}\sum_{m\in\mathcal{S}_k(v)}(\log|d_m(v)| - \hat{\mu}_k)^2, \tag{4}$$

we can transform Equation (3) as follows.

$$\hat{\mathcal{S}}_1(v) = \arg\max_{\mathcal{S}_1(v)}\left\{2\log\mathcal{L}(\mathcal{S}(v);\mathcal{S}_1(v),\hat{\Theta})\right\} = \arg\max_{\mathcal{S}_1(v)}\left\{-\sum_{k\in\{1,2\}}|\mathcal{S}_k|\log\left(\hat{\sigma_k^2}\right)\right\}. \tag{5}$$

Therefore, when a candidate upper group $\mathcal{S}_1(v)$ is successively changed by shifting its boundary between $\mathcal{S}_1(v)$ and $\mathcal{S}_2(v)$, we can efficiently obtain $\hat{\mathcal{S}}_1(v)$ by simply updating the sufficient statistics for calculating the maximum likelihood estimators. Here, we define the following operation to obtain the set of elements with the maximum influence degree,

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

because there might exist more than one diffusion sample with the same influence degree. Then, we can summarize our algorithm as follows.

1. Initialize $\mathcal{S}_1(v) \leftarrow \eta(\mathcal{S}(v))$, $\mathcal{S}_2(v) \leftarrow \mathcal{S}(v) \setminus \eta(\mathcal{S}(v))$, and $\hat{L} \leftarrow -\infty$.
2. Iterate the following procedure:
2-1. Set $\mathcal{S}_1(v) \leftarrow \mathcal{S}_1(v) \cup \eta(\mathcal{S}_2(v))$, and $\mathcal{S}_2(v) \leftarrow \mathcal{S}_2(v) \setminus \eta(\mathcal{S}_2(v))$.
2-2. If $\mathcal{S}_2(v) = \eta(\mathcal{S}_2(v))$, then terminate the iteration.
2-3. Calculate $L = -\sum_{k\in\{1,2\}}|\mathcal{S}_k(v)|\log(\hat{\sigma_k^2})$.
2-4. If $\hat{L} < L$ then set $\hat{L} \leftarrow L$ and $\hat{\mathcal{S}}_1(v) \leftarrow \mathcal{S}_1(v)$
3. Output $\hat{\mathcal{S}}_1(v)$, and terminate the algorithm.

We describe the computational complexity of the above algorithm. Clearly, the number of iterations performed in step 2 is at most $(M(v) - 2)$. On the other hand, when applying the operator $\eta(\cdot)$ in steps 1 and 2.1 (or 2.2), by classifying each diffusion sample according to its influence degree in advance, we can perform these operations with computational complexity of $O(1)$. Here note that since the influence degree is a positive integer less than or equal to $|V|$, we can perform the classification with computational complexity of $O(M(v))$. As for step 2.3, by adding (or removing) statistics calculated from $\eta(\mathcal{S}_2(v))$, we can update the maximum likelihood estimators $\hat{\Theta}$ defined in Equation (4) with computational complexity of $O(1)$. Therefore, the total computational complexity of our clustering algorithm is $O(M(v))$. Note that the above discussion can be applicable to a more general case for which the sufficient statistics of $p(m; \theta_k)$ is available to its parameter estimation.

A standard approach to the above clustering problem might be applying the EM algorithm by assuming a mixture of lognormal components. However, this approach is likely to confront the following drawbacks: 1) due to the local optimal problem, a number of parameter estimation trials are generally required by changing the initial parameter values, and we cannot guarantee the global optimality for the final result; 2) since many iterations are required for each parameter estimation trial, we need a substantially large computational load for obtaining the solution, which results in a prohibitively large processing time especially for a large data set; and 3) in case that a data set contains malicious outlier samples, we need a special care to avoid some unexpected problems such as degradation of $\hat{\sigma_k^2}$ to 0. Actually, our preliminary experiments based on this approach suffered from these drawbacks. In contrast, our proposed method always produces the optimal result with computational complexity of $O(M(v))$.

3.3 Super-mediator Discovery

Next, we describe our method for discovering super-mediator nodes. Let $D = \{d_m(v); v \in V, m = 1, \cdots, M(v)\}$ denote a set of observed diffusion samples. By using the above clustering method, we can estimate the upper group $\hat{\mathcal{S}}_1$ for each node $v \in V$. For $\hat{\mathcal{S}}_1(v)$, we employ, as a natural super-mediator score for a node $w \in V$, the following F-measure $F(w; v)$, a widely used measure in information retrieval, which is the harmonic average of recall and precision of a node w for the node v. Here 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.

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

Note that instead of the F-measure, we can employ the other measures such as the Jaccard coefficients, but for our objective that discovers characteristic nodes appearing in $\hat{\mathcal{S}}_1(v)$, we believe that the F-measure is most basic and natural. Then, we can consider the following expected F-measure for D.

$$\mathcal{F}(w) = \sum_{v \in V} F(w; v) r(v), \tag{8}$$

where $r(v)$ stands for the probability that the node v becomes an information source node, which can be empirically estimated by $r(v) = M(v) / \sum_{v \in V} M(v)$. Therefore, we can discover candidates for the super-mediator nodes by ranking the nodes according to the above expected F-measure.

In order to confirm the validity of the F-measure and characterize its usefulness, we compare the ranking by the F-measure with the rankings by two other measures, and investigate how these rankings are different from or correlated to each other considering several situations. The first one is the expected influence degree defined in Section 2.3. From observed diffusion samples D, we can estimate it as follows.

$$\mathcal{E}(w) = \frac{1}{M(w)} \sum_{m=1}^{M(w)} |d_m(w)|. \tag{9}$$

The second one is the following measure:

$$\mathcal{N}(w) = \sum_{v \in V} |\{m; w \in d_m(v)\}| r(v). \tag{10}$$

This measure ranks high those nodes that are easily influenced by many other nodes.

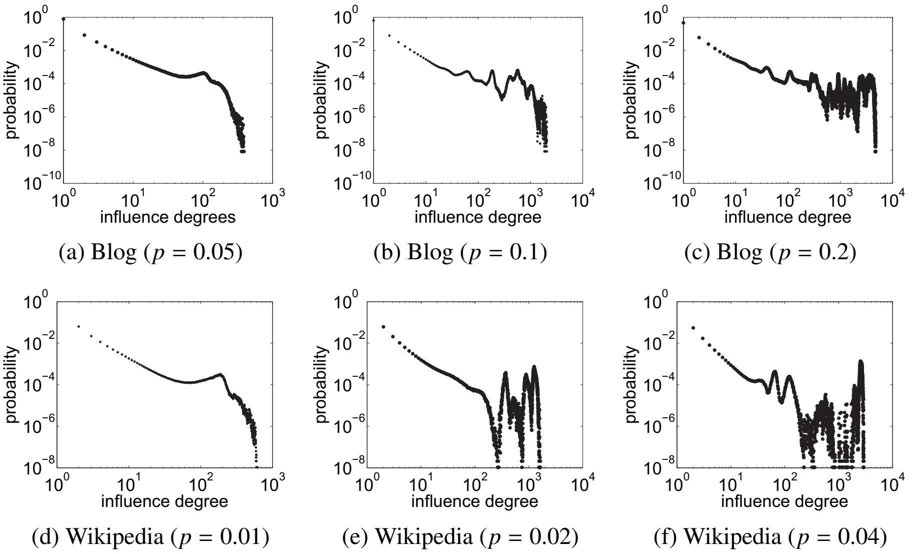
4 Experimental Evaluation

4.1 Data Sets

We employed two datasets of large real networks, which are both bidirectionally connected networks. The first one is a trackback network of Japanese blogs used in [13] and has $12,047$ nodes and $79,920$ directed links (the blog network). The other one is a network of people derived from the "list of people" within Japanese Wikipedia, also used in [13], and has $9,481$ nodes and $245,044$ directed links (the Wikipedia network).

Here, according to [17], we assumed the simplest case where the parameter values are uniform across all links and nodes, i.e. $p_{u,v} = p$ for the IC model. As for the LT model, we assumed $\omega_{u,v} = q|B(v)|^{-1}$, and adopted q $(0 \leq q \leq 1)$ as the unique parameter for a network instead of $\omega_{u,v}$ as in [18]. According to [8], we set p to a value smaller than $1/\bar{d}$, where $\bar{d}$ is the mean out-degree of a network. Thus, the value of p was set to 0.1 for the blog network and 0.02 for the Wikipedia network. These are the base values, but in addition to them, we used two other values, one two times larger and the other two times smaller for our analyses, i.e. 0.02 and 0.05 for the blog network, and 0.04 and 0.01 for the Wikipedia network. We set the base value for q to be 0.9 for the both networks to achieve reasonably long diffusion results. Same as p, we also adopted two other values, one two times larger and the other two times smaller. Since the double of 0.9 exceeds the upper-bound of q, i.e. 1.0, we used 1.0 for the larger value, and we used 0.45 for the smaller one.

For each combination of these values, information diffusion samples were generated for the corresponding model on each network using each node in the network as the initial active node. In our experiments, we set $M = 10,000$, which means $10,000$ information diffusion samples were generated for each initial active node. Then, we analyzed them to discover super-mediators. To efficiently generate those information diffusion samples and estimate the expected influence degree $\mathcal{E}$ of an initial active node,



(a) Blog (p = 0.05) (b) Blog (p = 0.1) (c) Blog (p = 0.2)

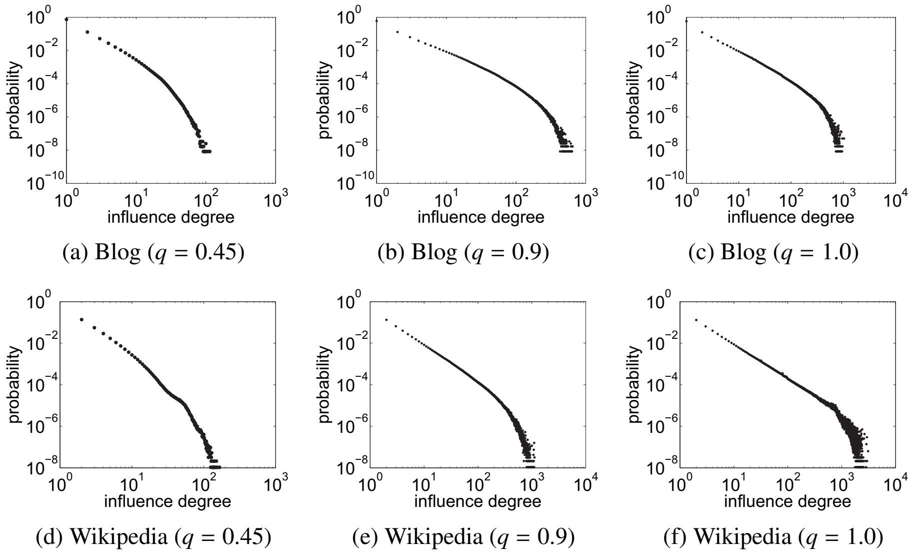
(d) Wikipedia (p = 0.01) (e) Wikipedia (p = 0.02) (f) Wikipedia (p = 0.04)

Fig. 2: The average influence degree distribution of the IC model

we adopted the method based on the bond percolation proposed in [14]. Note that we only use these two networks to generate the diffusion sample data which we assume we observed. Once the data are obtained, we no more use the network structure.

4.2 Influence Degree Distribution

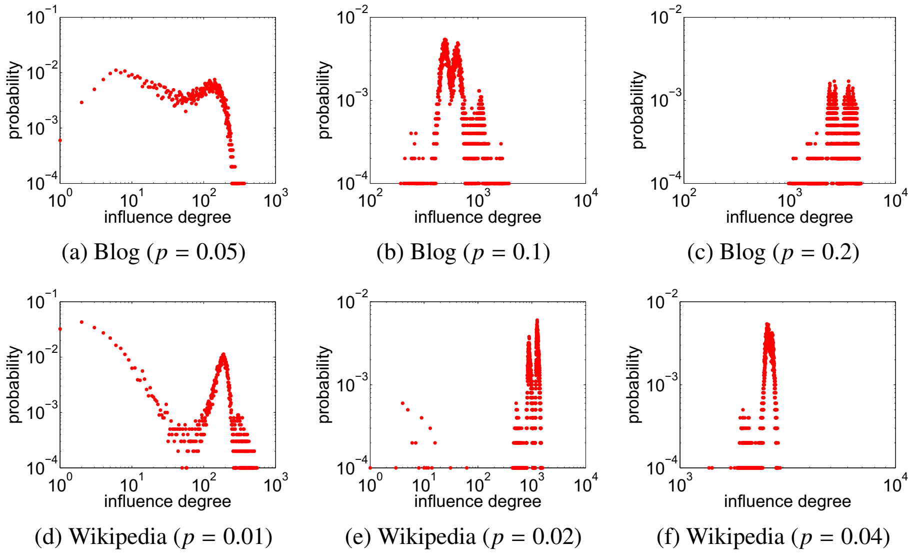
First, we show the influence degree distribution for all nodes. Figure 2 is the results of the IC model and Fig. 3 is the results of the LT model. $M(= 10,000)$ simulations were performed for each initial node $v \in V$ and this is repeated for all the nodes in the network. Since the number of the nodes $|V|$ is about 10,000 for both the blog and the Wikipedia networks, these results are computed from about one hundred million diffusion samples and exhibits global characteristics of the distribution. We see that the distribution of the IC model consists of lognormal like distributions for a wide range of diffusion probability p with clearer indication for a larger p. Here it is known that if the variance of the lognormal distribution is large, it can be reasonably approximated by a power-law distribution [23]. On the contrary, we note that the distribution of the LT model is different and is a monotonically decreasing power-law like distribution. This observation is almost true of the distribution for an individual node v except that the distribution has one peak for the LT model. One example is already shown in Fig 1 (b) for the IC model. Figures 4 and 5 show some other results for the both models. In each of these figures the most influential node for the parameter used was chosen as the initial activated source node v. From this observation, the discovery model we derived in subsections 3.2 and 3.3 can be straightforwardly applied to the IC model by assuming that the probability distribution consists of lognormal components and the succeeding experiments were performed for the IC model. However, this does not necessarily



(a) Blog (q = 0.45) (b) Blog (q = 0.9) (c) Blog (q = 1.0)

(d) Wikipedia (q = 0.45) (e) Wikipedia (q = 0.9) (f) Wikipedia (q = 1.0)

Fig. 3: The average influence degree distribution of the LT model

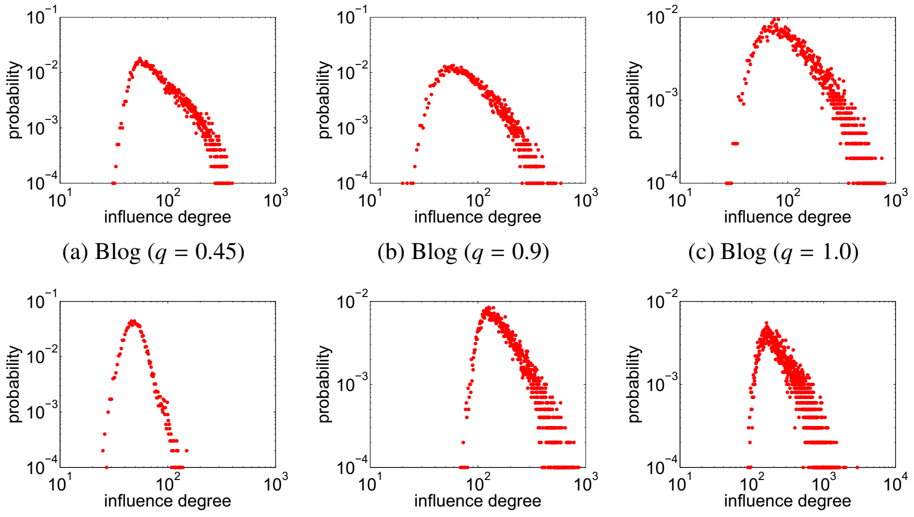


(a) Blog (p = 0.05) (b) Blog (p = 0.1) (c) Blog (p = 0.2)

(d) Wikipedia (p = 0.01) (e) Wikipedia (p = 0.02) (f) Wikipedia (p = 0.04)

Fig. 4: The influence degree distribution for a specific node v of the IC model

mean that the notion of super-mediator is only applicable to the IC model. Finding a reasonable and efficient way to discover super-mediator nodes for the LT model is our on-going research topic. Further, the assumption of dividing the groups into only two need be justified. This is also left to our future work.



(a) Blog (q = 0.45) (b) Blog (q = 0.9) (c) Blog (q = 1.0)

(d) Wikipedia (q = 0.45) (e) Wikipedia (q = 0.9) (f) Wikipedia (q = 1.0)

Fig. 5: The influence degree distribution for a specific node v of the LT model

4.3 Super-mediator Ranking

Tables 1, 2 and 3 summarize the ranking results. Ranking is evaluated for two different values of diffusion probability ($p = 0.1$ and $p = 0.05$ for the blog data, and $p = 0.02$ and 0.01 for the Wikipedia data) and for the three measures mentioned in subsection 3.3. Rank by all the measures is based on the value rounded off to three decimal places. So the same rank appears more than once. The first two (Tables 1 and 2) rank the nodes by $\mathcal{F}$ for $p = 0.1$ and 0.05 (blog data) and $p = 0.02$ and 0.01 (Wikipedia data), respectively,

Table 1: Comparison of the ranking by $\mathcal{F}$ with rankings by $\mathcal{E}$ and $\mathcal{N}$ for a large diffusion probability.

(a) Blog network (p = 0.1)

Ranking by $\mathcal{F}$		Ranking by $\mathcal{E}$, $\mathcal{N}$	
Ranking	Node ID	$\mathcal{E}$	$\mathcal{N}$
1	146	2	2
1	155	1	1
3	140	3	3
3	150	4	4
5	238	5	5
5	278	6	6
5	240	7	7
5	618	10	8
9	136	8	9
9	103	9	10

(b) Wikipedia network (p = 0.02)

Ranking by $\mathcal{F}$		Ranking by $\mathcal{E}$, $\mathcal{N}$	
Ranking	Node ID	$\mathcal{E}$	$\mathcal{N}$
1	790	1	1
1	8340	2	2
3	323	3	3
3	279	4	4
5	326	5	5
6	772	6	6
6	325	7	7
8	1407	8	8
9	4924	9	9
10	3149	11	10

Table 2: Comparison of the ranking by $\mathcal{F}$ with rankings by $\mathcal{E}$ and $\mathcal{N}$ for a small diffusion probability.

(a) Blog network (p = 0.05)

Ranking by $\mathcal{F}$		Ranking by $\mathcal{E}, \mathcal{N}$	
Ranking	Node ID	$\mathcal{E}$	$\mathcal{N}$
1	155	26	28
2	146	29	29
3	140	41	44
4	150	63	66
5	238	92	93
6	618	79	81
6	240	113	112
8	103	84	86
8	490	95	96
8	173	88	89

(b) Wikipedia network (p = 0.01)

Ranking by $\mathcal{F}$		Ranking by $\mathcal{E}, \mathcal{N}$	
Ranking	Node ID	$\mathcal{E}$	$\mathcal{N}$
1	790	167	168
2	279	199	198
2	4019	1	1
4	3729	2	2
4	7919	3	3
4	1720	7	4
4	4465	5	6
4	1712	6	7
9	4380	4	5
9	3670	9	8

Table 3: Comparison of the ranking by $\mathcal{E}$ for a high diffusion probability with rankings by $\mathcal{E}$, $\mathcal{F}$, and $\mathcal{N}$ for a low diffusion probability.

(a) Blog network

Ranking by $\mathcal{E}$ for p = 0.1		Ranking by $\mathcal{E}, \mathcal{F}, \mathcal{N}$ for p = 0.05		
Ranking	Node ID	$\mathcal{E}$	$\mathcal{F}$	$\mathcal{N}$
1	155	26	1	28
2	146	29	2	29
3	140	41	3	44
4	150	63	4	66
5	238	92	5	93
6	278	161	18	154
7	240	113	6	112
8	136	83	8	85
9	103	84	8	86
10	618	79	6	81

(b) Wikipedia network

Ranking by $\mathcal{E}$ for p = 0.02		Ranking by $\mathcal{E}, \mathcal{F}, \mathcal{N}$ for p = 0.01		
Ranking	Node ID	$\mathcal{E}$	$\mathcal{F}$	$\mathcal{N}$
1	790	167	1	168
2	8340	200	9	201
3	323	196	14	200
4	279	199	2	198
5	326	212	24	206
6	325	231	51	236
7	772	242	41	235
8	1407	257	80	264
9	4924	305	111	298
10	2441	279	103	287

and compare each ranking with those by $\mathcal{E}$ and $\mathcal{N}$. From these results we observe that when the diffusion probability is large all the three measures ranks the nodes in a similar way. This means that the influential nodes also play a role of super-mediator for the other source nodes. When the diffusion probability is small, the Wikipedia data still shows the similar tendency but the blog data does not. We further note that $\mathcal{E}$ and $\mathcal{N}$ rank the nodes in a similar way regardless of the value of diffusion probability. This is understandable because the both networks are bidirectional. In summary, when the diffusion probability is large, all the three measures are similar and the influential nodes also play a role of super-mediator for the other source nodes.

The third one (Table 3) ranks the nodes by $\mathcal{E}$ for $p = 0.01$ (blog data) and $p = 0.02$ (Wikipedia data) and compares them with the rankings by the three measures for $p = 0.05$ (blog data) and $p = 0.01$ (Wikipedia data). The results say that the influential nodes are different between the two different diffusion probabilities, but what is strikingly interesting to note is that the nodes that are identified to be influential (up to 10th) at a large diffusion probability are almost the same as the nodes that rank high by $\mathcal{F}$ at a small diffusion probability for the blog data. This correspondence is not that clear for the Wikipedia data but the correlation of the rankings by $\mathcal{E}$ (at a large diffusion probability) and $\mathcal{F}$ (at a small diffusion probability) is much larger than the corresponding correlation by the other two measures ($\mathcal{E}$ and $\mathcal{N}$). This implies that the super-mediators at a small diffusion probability become influential at a large diffusion probability. Since the F-measure can be evaluated by the observed information sample data alone and there is no need to know the network structure, this fact can be used to predict which nodes become influential when the diffusion probability switches from a small value for which we have enough data to a large value for which we do not have any data yet.

4.4 Characterization of Super-mediator and Discussions

If we observe that some measure evaluated for a particular value of diffusion probability gives an indication of the influential nodes when the value of diffusion probability is changed, it would be a useful measure for finding influential nodes for a new situation. It is particularly useful when we have abundant observed set of information diffusion samples with normal diffusion probability and we want to discover high ranked influential nodes in a case where the diffusion probability is larger. For example, this problem setting corresponds to predicting the influential nodes for the unexperienced rapid spread of new information, e.g. spread of new acute contagion, because it is natural to think that we have abundant data for the spread of normal moderate contagion.

The measure based on $\mathcal{E}$ ranks high those nodes that are also influential where the diffusion probability is different from the current value if nodes are not sensitive to the diffusion probability, i.e. a measure useful to estimate influential nodes from the known results when the diffusion probability changes under such a condition. The measure based on $\mathcal{N}$ ranks high those nodes that are easily influenced by many other nodes. It is a measure useful to estimate influential nodes from the known results if they are the nodes easily influenced by other nodes. In our experiments, the influential nodes by $\mathcal{E}$ for the much larger diffusion probability, i.e. $p = 0.2$ (blog data) and $p = 0.04$ (Wikipedia data) were almost the same as the high ranked ones by any one of the three measures $\mathcal{E}$, $\mathcal{N}$ and $\mathcal{F}$ for $p = 0.1$ (blog data) and $p = 0.02$ (Wikipedia data), although we have to omit the details due to the space limitation.

In the previous subsection we showed that the super-mediators at a small diffusion probability become influential at a large diffusion probability. In a situation where there are relatively large number of active nodes, the probability that more than one parent try to activate their same child increases, which mirrors the situation where the diffusion probability is effectively large. It is the super-mediators that play the central role in these active node group under such a situation. This would explain why the super-mediators at a small diffusion probability become influential nodes at a large diffusion probability.

5 Conclusion

We found that the influence degree for the IC model exhibits a distribution which is a mixture of two distributions (power-law like distribution and lognormal like distribution). This implied that there are nodes that may play different roles in information diffusion process. We made a hypothesis that there should be nodes that play an important role to pass the information to other nodes, and called these nodes "super-mediators". These nodes are different from what is usually called "influential nodes" (nodes that spread information as much as possible). We devised an algorithm based on maximum likelihood and linear search which can efficiently identify the super-mediator node group from the observed diffusion sample data, and proposed a measure based on recall and precision to rank the super-mediators. We tested our hypothesis by applying it to the information diffusion sample data generated by two real networks. We found that the high ranked super-mediators are also the high ranked influential nodes when the diffusion probability is large, i.e. the influential nodes also play a role of super-mediator for the other source nodes, but not necessarily so when the diffusion probability is small, and further, to our surprise, that when the high ranked super-mediators are different from the top ranked influential nodes, which is the case when the diffusion probability is small, those super-mediators become the high ranked influential nodes when the diffusion probability becomes larger. This finding will be useful to predict the influential nodes for the unexperienced spread of new information from the known experience, e.g. prediction of influential nodes for the spread of new acute contagion for which we have no available data yet from the abundant data we already have for the spread of moderate contagion.

Acknowledgments

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research under Grant No. AOARD-10-4053, and JSPS Grant-in-Aid for Scientific Research (C) (No. 20500147).

References

1. 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** (2001) 211–223
2. Newman, M.E.J., Forrest, S., Balthrop, J.: Email networks and the spread of computer viruses. Physical Review E **66** (2002) 035101
3. Newman, M.E.J.: The structure and function of complex networks. SIAM Review **45** (2003) 167–256
4. Watts, D.J.: A simple model of global cascades on random networks. Proceedings of National Academy of Science, USA **99** (2002) 5766–5771
5. Watts, D.J., Dodds, P.S.: Influence, networks, and public opinion formation. Journal of Consumer Research **34** (2007) 441–458
6. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. SIGKDD Explorations **6** (2004) 43–52

7. Domingos, P.: Mining social networks for viral marketing. IEEE Intelligent Systems **20** (2005) 80–82
8. 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). (2003) 137–146
9. 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). (2006) 228–237
10. 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). (2007) 1371–1376
11. Kimura, M., Saito, K., Nakano, R., Motoda, H.: Extracting influential nodes on a social network for information diffusion. Data Mining and Knowledge Discovery, Springer **20** (2010) 70–97
12. Kimura, M., Saito, K., Motoda, H.: Minimizing the spread of contamination by blocking links in a network. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI-08). (2008) 1175–1180
13. Kimura, M., Saito, K., Motoda, H.: Blocking links to minimize contamination spread in a social network. ACM Transactions on Knowledge Discovery from Data **3** (2009) 9:1–9:23
14. 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)
15. 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), Springer, LNAI 5808 (2009) 302–316
16. Kimura, M., Saito, K., Nakano, R., Motoda, H.: Finding influential nodes in a social network from information diffusion data. In: Proceedings of the International Workshop on Social Computing and Behavioral Modeling (SBP09). (2009) 138–145
17. 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). (2009) 322–337
18. Saito, K., Kimura, M., Ohara, K., Motoda, H.: Behavioral analyses of information diffusion models by observed data of social network. In: Proceedings of the International Workshop on Social Computing and Behavioral Modeling (SBP10). (2010) 149–158
19. Goyal, A., Bonchi, F., Lakshhmanan, L.V.S.: Learning influence probabilities in social networks. In: Proceedings of the third ACM international conference on Web Search and Data Mining. (2010) 241–250
20. Bakshy, E., Karrer, B., Adamic, L.A.: Social influence and the diffusion of user-created content. In: Proceedings of the tenth ACM conference on Electronic Commerce. (2009) 325–334
21. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007). (2007) 420–429
22. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2009). (2009) 199–208
23. Mitzenmacher, M.: A brief history of generative models for power law and lognormal distributions. Internet Mathematics **1** (2004) 226–251