What Does an Information Diffusion Model Tell about Social Network Structure?

Takayasu Fushimi[1], Takashi Kawazoe[1], Kazumi Saito[1], Masahiro Kimura[2], and Hiroshi Motoda[3]

[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, Shiga 520-2194, Japan
kimura@rins.ryukoku.ac.jp
[3] Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
motoda@ar.sanken.osaka-u.ac.jp

Abstract. In this paper, we attempt to answer a question "What does an information diffusion model tell about social network structure?" To this end, we propose a new scheme for empirical study to explore the behavioral characteristics of representative information diffusion models such as the IC (Independent Cascade) model and the LT (Linear Threshold) model on large networks with different community structure. To change community structure, we first construct a GR (Generalized Random) network from an originally observed network. Here GR networks are constructed just by randomly rewiring links of the original network without changing the degree of each node. Then we plot the expected number of influenced nodes based on an information diffusion model with respect to the degree of each information source node. Using large real networks, we empirically found that our proposal scheme uncovered a number of new insights. Most importantly, we show that community structure more strongly affects information diffusion processes of the IC model than those of the LT model. Moreover, by visualizing these networks, we give some evidence that our claims are reasonable.

1 Introduction

We can now obtain digital traces of human social interaction with some relating topics in a wide variety of on-line settings, like Blog (Weblog) communications, email exchanges and so on. Such social interaction can be naturally represented as a large-scale social network, where nodes (vertices) correspond to people or some social entities, and links (edges) correspond to social interaction between them. Clearly these social networks reflect complex social structures and distributed social trends. Thus, it seems worth putting some effort in attempting to find empirical regularities and develop explanatory accounts of basic functions in the social networks. Such attempts would be valuable for understanding social structures and trends, and inspiring us to lead to the discovery of new knowledge and insights underlying social interaction.

A social network can also play an important role as a medium for the spread of various information [7]. For example, innovation, hot topics and even malicious rumors can propagate through social networks among individuals, and computer viruses can diffuse through email networks. Previous work addressed the problem of tracking the propagation patterns of topics through network spaces [3, 1], and studied effective "vaccination" strategies for preventing the spread of computer viruses through networks [8, 2]. Widely-used fundamental probabilistic models of information diffusion through networks are the *independent cascade (IC) model* and the *linear threshold (LT) model* [4, 3]. Researchers have recently investigated the problem of finding a limited number of influential nodes that are effective for the spread of information through a network under these models [4, 5]. Moreover, the influence maximization problem has recently been extended to general influence control problems such as a contamination minimization problem [6].

To deepen our understanding of social networks and accelerating study on information diffusion models, we attempt to answer a question "What does an information diffusion model tell about social network structure?" We except that such attempts derive some improved methods for solving a number of problems based on information diffusion models such as the influence maximization problem [5]. In this paper, we propose a new scheme for emperical study to explore the behavioral characteristics of representative information diffusion models such as the IC model and the LT model on large networks with different community structure. We perform extensive numerical experiments on two large real networks, one generated from a large connected trackback network of blog data, resulting in a directed graph of $12,047$ nodes and $79,920$ links, and the other, a network of people, generated from a list of people within a Japanese Wikipedia, resulting in an undirected graph of $9,481$ nodes and $245,044$ links. Through these experiments, we show that our proposed scheme could uncover a number of new insights on information diffusion processes of the IC model and the LT model.

2 Information Diffusion Models

We mathematically model the spread of information through a directed network $G = (V, E)$ under the IC or LT model, where V and E $(\subset V \times V)$ stands for the sets of all the nodes and links, respectively. We call nodes *active* if they have been influenced with the information. In these models, the diffusion process unfolds in discrete time-steps $t \geq 0$, and it is assumed that nodes can switch their states only from inactive to active, but not from active to inactive. Given an initial set S of active nodes, we assume that the nodes in S have first become active at time-step 0, and all the other nodes are inactive at time-step 0.

2.1 Independent Cascade Model

We define the IC model. In this model, for each directed link (u, v), we specify a real value $\beta_{u,v}$ with $0 < \beta_{u,v} < 1$ in advance. Here $\beta_{u,v}$ is referred to as the *propagation probability* through link (u, v). The diffusion process proceeds from a given initial active set S in the following way. When 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 $\beta_{u,v}$. 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 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.

For an initial active set S, let $\varphi(S)$ denote the number of active nodes at the end of the random process for the IC model. Note that $\varphi(S)$ is a random variable. Let $\sigma(S)$ denote the expected value of $\varphi(S)$. We call $\sigma(S)$ the *influence degree* of S.

2.2 Linear Threshold Model

We define the LT model. In this model, for every node $v \in V$, we specify, in advance, a *weight* $\omega_{u,v}$ (> 0) from its parent node u such that

$$\sum_{u \in \Gamma(v)} \omega_{u,v} \leq 1,$$

where $\Gamma(v) = \{u \in V; (u,v) \in E\}$. The diffusion process from a given initial active set S 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 at least threshold θ_v, that is,

$$\sum_{u \in \Gamma_t(v)} \omega_{u,v} \geq \theta_v,$$

then v will become active at time-step $t+1$. Here, $\Gamma_t(v)$ 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.

The LT model is also a probabilistic model associated with the uniform distribution on $[0,1]^{|V|}$. Similarly to the IC model, we define a random variable $\varphi(S)$ and its expected value $\sigma(S)$ for the LT model.

2.3 Bond Percolation Method

First, we revisit the bond percolation method [5]. Here, we consider estimating the influence degrees $\{\sigma(v;G);\ v \in V\}$ for the IC model with propagation probability p in graph $G = (V, E)$. For simplicity we assigned a uniform value p for $\beta_{u,v}$.

It is known that the IC model is equivalent to the bond percolation process that independently declares every link of G to be "occupied" with probability p [7].

It is known that the LT model is equivalent to the following bond percolation process [4]: For any $v \in V$, we pick at most one of the incoming links to v by selecting link (u,v) with probability $\omega_{u,v}$ and selecting no link with probability $1 - \sum_{u \in \Gamma(v)} \omega_{u,v}$. Then, we declare the picked links to be "occupied" and the other links to be "unoccupied". Note here that the equivalent bond percolation process for the LT model is considerably different from that of IC model.

Let M be a sufficiently large positive integer. We perform the bond percolation process M times, and sample a set of M graphs constructed by the occupied links,

$$\{G^m = (V, E^m);\ m = 1, \cdots, M\}.$$

Then, we can approximate the influence degree $\sigma(v; G)$ by

$$\sigma(v; G) \simeq \frac{1}{M} \sum_{m=1}^{M} |\mathcal{F}(v; G^m)|.$$

Here, for any directed graph $\tilde{G} = (V, \tilde{E})$, $\mathcal{F}(v; \tilde{G})$ denotes the set of all the nodes that are *reachable* from node v in the graph. We say that node u is reachable from node v if there is a path from u to v along the links in the graph. Let

$$V = \bigcup_{u \in \mathcal{U}(G^m)} \mathcal{S}(u; G^m)$$

be the strongly connected component (SCC) decomposition of graph G^m, where $\mathcal{S}(u; G^m)$ denotes the SCC of G^m that contains node u, and $\mathcal{U}(G^m)$ stands for a set of all the representative nodes for the SCCs of G^m. The bond percolation method performs the SCC decomposition of each G^m, and estimates all the influence degrees $\{\sigma(v; G);\ v \in V\}$ in G as follows:

$$\sigma(v; G) = \frac{1}{M} \sum_{m=1}^{M} |\mathcal{F}(u; G^m)|, \quad (v \in \mathcal{S}(u; G^m)), \tag{1}$$

where $u \in \mathcal{U}(G^m)$.

3 Proposed Scheme for Experimental Study

We technically describe our proposed scheme for empirical study to explore the behavioral characteristics of representative information diffusion models on large networks different community structure. In addition, we present a method for visualizing such networks in terms of community structure. Hereafter, the degree of a node v, denoted by $deg(v)$, means the number of links connecting from or to the node v.

3.1 Affection of Community Structure

As mentioned earlier, our scheme consists of two parts. Namely, to change community structure, we first construct a GR (generalized random) network from an originally observed network. Here GR networks are constructed just by randomly rewiring links of the original network without changing the degree of each node [7]. Then we plot the influence degree based on an information diffusion model with respect to the degree of each information source node.

First we describe the method for constructing a GR network. By arbitrary ordering all links in a given original network, we can prepare a link list $L_E = (e_1, \cdots, e_{|E|})$. Recall that each directed link consists of an ordered pair of *from*-part and *to*-part nodes, i.e., $e = (u, v)$. Thus, we can produce two node lists from the list L_E, that is, the *from*-part node list L_F and the *to*-part node list L_T. Clearly the frequency of each node v appearing in L_F (or L_T) is equivalent to the out (or in) degree of the node v. Therefore, by randomly reordering the node list L_T, then concatenating it with the other node list L_F, we can produce a link list for a GR network. More specifically, let L'_T be a shuffled node list, and we denote the i-th order element of a list L by $L(i)$, then the link list of the GR network is $L'_E = ((L_F(1), L'_T(1)), \cdots, (L_F(|E|), L'_T(|E|)))$. Here note that to fairly compare the GR network with original one in terms of influence degree, we excluded some types of shuffled node lists, each of which produces a GR network with self-links of some node or multiple-links between any two nodes.

By using the bond percolation method described in the previous section, we can efficiently obtain the influence degree $\sigma(v)$ for each node v. Thus we can straightforwardly plot each pair of $deg(v)$ and $\sigma(v)$. Moreover, to examine their tendency of nodes with the same degree δ, we also plot the average influence degree $\mu(\delta)$ calculated by

$$\mu(\delta) = \frac{1}{|\{v : deg(v) = \delta\}|} \sum_{\{v:deg(v)=\delta\}} \sigma(v). \tag{2}$$

Clearly we can guess that nodes with larger degrees influence many other nodes in any information diffusion models, but we consider that it is worth examining its curves in more details.

3.2 Visualization of Community Structure

In order to intuitively grasp the original and GR networks in terms of community structure, we present a visualization method that is based on the cross-entropy algorithm [11] for network embedding, and the k-core notion [10] for label assignment.

First we describe the network embedding problem. Let $\{\mathbf{x}_v : v \in V\}$ be the embedding positions of the corresponding $|V|$ nodes in an R dimensional Euclidean space. As usual, we define the Euclidean distance between $\mathbf{x}_u$ and $\mathbf{x}_w$ as follows:

$$d_{u,w} = \|\mathbf{x}_u - \mathbf{x}_w\|^2 = \sum_{r=1}^{R} (x_{u,r} - x_{w,l})^2.$$

Here we introduce a monotonic decreasing function $\rho(s) \in [0, 1]$ with respect to $s \geq 0$, where $\rho(0) = 1$ and $\rho(\infty) = 0$. Let $a_{u,w} \in \{0, 1\}$ be an adjacency information between two nodes u and w, indicating whether their exist a link between them ($a_{u,w} = 1$) or not ($a_{u,w} = 0$). Then we can introduce a cross-entropy (cost) function between $a_{u,w}$ and $\rho(d_{u,w})$ as follows:

$$\mathcal{E}_{u,w} = -a_{u,w} \ln \rho(d_{u,w}) - (1 - a_{u,w}) \ln(1 - \rho(d_{v,w})).$$

Since $\mathcal{E}_{u,w}$ is minimized when $\rho(d_{u,w}) = a_{u,w}$, this minimization with respect to $\mathbf{x}_u$ and $\mathbf{x}_w$ basically coincides with our problem setting. In this paper, we employ $\rho(s) = \exp(-s/2)$ as the monotonic decreasing function. Then the total cost function (objective function) can be defined as follows:

$$\mathcal{E} = \frac{1}{2} \sum_{u \in V} \sum_{w \in V} a_{u,w} d_{u,w} - \sum_{u \in V} \sum_{w \in V} (1 - a_{u,w}) \ln(1 - \rho(d_{u,w})). \tag{3}$$

Namely the cross-entropy algorithm minimizes the objective function defined in (3) with respect to $\{\mathbf{x}_v : v \in V\}$.

Next we explain the k-core notion. For a given node v in the network $G = (V_G, E_G)$, we denote $A_G(v)$ as a set of *adjacent nodes* of v as follows:

$$A_G(v) = \{w : \{v, w\} \in E_G\} \cup \{u : \{u, v\} \in E_G\}.$$

A subnetwork $C(k)$ of G is called *k-core* if each node in $C(k)$ has more than or equal to k adjacent nodes in $C(k)$. More specifically, we can define k-core subnetwork as follows. For a given order k, the k-core is a subnetwork $C(k) = (V_{C(k)}, E_{C(k)})$ consisting of the following node set $V_{C(k)} \subset V_G$ and link set $V_{C(k)} \subset V_G$:

$$V_{C(k)} = \{v : |A_{C(k)}(v)| \geq k\}, \quad E_{C(k)} = \{e : e \subset V_{C(k)}\}.$$

Here according to our purpose, we focus on the subnetwork of maximum size with this property as a k-core subnetwork $C(k)$.

Finally we describe the label assignment strategy. As a rough necessary condition, we assume that each community over a network includes a higher order k-core as its part. Here we consider that a candidate for such higher core order is greater than the average degree calculated by $\overline{d} = |E|/|V|$. Then we can summarize our visualization method as follows: after embedding a given network into an R (typically $R = 2$) dimensional Euclidean space by use of the cross-entropy algorithm, our visualization method plots each node position by changing the appearance of nodes belonging to its $([\overline{d}] + 1)$-core subnetwork. Here note that $[\overline{d}]$ denotes the greatest integer smaller than $\overline{d}$. By this visualization method, we can expect to roughly grasp community structure of a given network.

4 Experimental Evaluation

4.1 Network Data

In our experiments, we employed two sets of real networks used in [5], which exhibit many of the key features of social networks as shown later. We describe the details of these network data.

The first one is a trackback network of blogs. Blogs are personal on-line diaries managed by easy-to-use software packages, and have rapidly spread through the World Wide Web [3]. Bloggers (*i.e.*, blog authors) discuss various topics by using trackbacks. Thus, a piece of information can propagate from one blogger to another blogger through a trackback. We exploited the blog “Theme salon of blogs” in the site “goo” [2], where a blogger can recruit trackbacks of other bloggers by registering an interesting theme. By tracing up to ten steps back in the trackbacks from the blog of the theme “JR Fukuchiyama Line Derailment Collision”, we collected a large connected trackback network in May, 2005. The resulting network had $12,047$ nodes and $79,920$ directed links, which features the so-called “power-law” distributions for the out-degree and in-degree that most real large networks exhibit. We refer to this network data as the blog network.

[2] `http://blog.goo.ne.jp/usertheme/`

The second one is a network of people that was derived from the "list of people" within Japanese Wikipedia. Specifically, we extracted the maximal connected component of the undirected graph obtained by linking two people in the "list of people" if they co-occur in six or more Wikipedia pages. The undirected graph is represented by an equivalent directed graph by regarding undirected links as bidirectional ones[3]. The resulting network had $9,481$ nodes and $245,044$ directed links. We refer to this network data as the Wikipedia network.

4.2 Characteristics of Network Data

Newman and Park [9] observed that social networks represented as undirected graphs generally have the following two statistical properties that are different from non-social networks. First, they show positive correlations between the degrees of adjacent nodes. Second, they have much higher values of the *clustering coefficient* C than the corresponding *configuration model* defined as the ensemble of GR networks. Here, the clustering coefficient C for an undirected network is defined by

$$C = \frac{1}{|V|} \sum_{u \in V} \frac{|\{(v \in V, w \in V) : v \neq w, w \in A_G(v)\}|}{|A_G(u)|(|A_G(u)| - 1)}.$$

Another widely-used statistical measure of network is the average length of shortest paths between any two nodes defined by

$$L = \frac{1}{|V|(|V| - 1)} \sum_{u \neq v} l(u, v).$$

where $l(u, v)$ denotes the shortest path length between nodes u and v. In terms of information diffusion processes, when L becomes smaller the probability that any information source nodes can activate the other nodes, becomes larger in general.

Table 1 shows the basic statistics of the blog and Wikipedia networks, together with their GR networks. We can see that the measured value of C for the original blog network is substantially larger than that of the GR blog network, and the measured value of L for the original blog network is somehow larger than that of the GR blog network indicating that there exisit communities. We can observe a similar tendency for the Wikipedia networks. Note that we have already confirmed for the original Wikipedia network that the degrees of adjacent nodes were positively correlated, although we derived the network from Japanese Wikipedia. Therefore, we can say that the Wikipedia network has the key features of social networks.

4.3 Experimental Settings

We describe our experimental settings of the IC and LT models. In the IC model, we assigned a uniform probability β to the propagation probability $\beta_{u,v}$ for any directed link (u, v) of a network, that is, $\beta_{u,v} = \beta$. As our β setting, we employed a reciprocal

[3] For simplicity, we call a graph with bi-directional links an undirected graph

Table 1: Basic statistics of networks.

network	$\|V\|$	$\|E\|$	C	L
original blog	12,047	79,920	0.26197	8.17456
GR blog	12,047	79,920	0.00523	4.24140
original Wikipedia	9,481	245,044	0.55182	4.69761
GR Wikipedea	9,481	245,044	0.04061	3.12848

of the average degree, i.e., $\beta = |V|/|E|$. The resulting propagation probability for the original and GR blog networks was $\beta = 0.1507$, and $\beta = 0.0387$ for the original and GR Wikipedia networks. Incidentally, these values were reasonably close to those used in former study, i.e., $\beta = 0.2$ for the blog networks and $\beta = 0.03$ for the Wikipedia networks were used in the former experiments [6].

In the LT model, we uniformly set weights as follows. For any node v of a network, the weight $\omega_{u,v}$ from a parent node $u \in \Gamma(v)$ is given by $\omega_{u,v} = 1/|\Gamma(v)|$. This experimental setting is exactly the same as the one performed in [5].

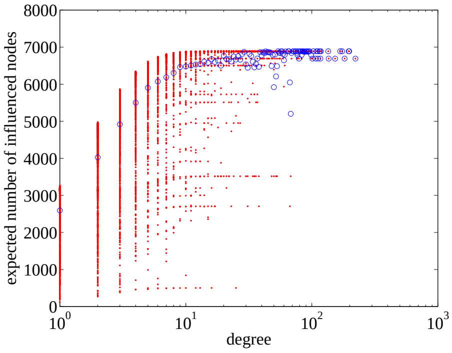
For the proposed method, we need to specify the number M of performing the bond percolation process. In the experiments, we used $M = 10,000$ [5]. Recall that the parameter M represents the number of bond percolation processes for estimating the influence degree $\sigma(v)$ of a given initial active node v. In our preliminary experiments, we have already confirmed that the influence degree of each node for these networks with $M = 10,000$ are comparable to those with $M = 300,000$.

4.4 Experimental Results Using Blog Network

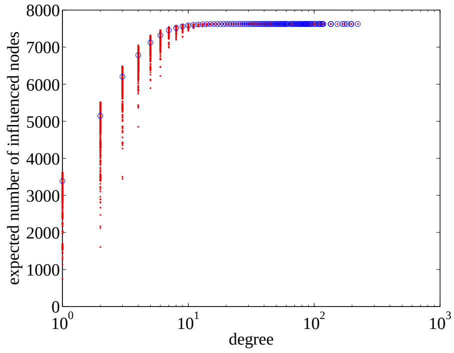
Figure 1a shows the influence degree based on the IC model with respect to the degree of each information source node over the original blog network, Figure 1b shows those of the IC model over the GR blog network, Figure 1c shows those of the LT model over the original Wikipedia network, and Figure 1d shows those of the LT model over the GR Wikipedia network. Here the red dots and blue circles respectively stand for the levels of the influence degree of individual nodes and their averages for the nodes with the same degree.

In view of the difference between the information diffusion models, we can clearly see that although nodes with larger degrees influenced many other nodes in both of the IC and LT models, their average curves exhibit opposite curvatures as shown in these results. In addition, we can observe that the influence degree of the individual nodes based on the IC model have quite large variances compared with those of the LT model.

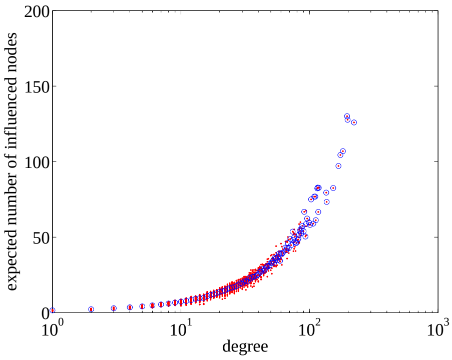
In view of the difference between the original and GR networks, we can see that compared with the original networks, the levels of the influence degree were somewhat larger in the GR networks. We consider that this is because the averages of shortest path lengths became substantially larger than those of the GR networks, especially for the IC model. In the case of the LT model over the GR network (Figure 1d), we can observe that the influence degree was almost uniquely determined by the degree of each node. As the most remarkable point, in the case of the IC model, we can observe a number of lateral lines composed of the individual influence degree over the original networks (Figure 1a), but these lines disappeared over the GR networks (Figure 1b).



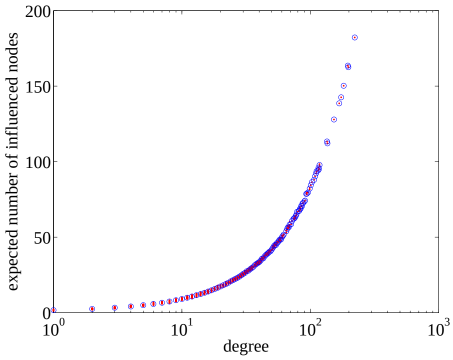
(a) Original network by IC model



(b) GR network by IC model



(c) Original network by LT model



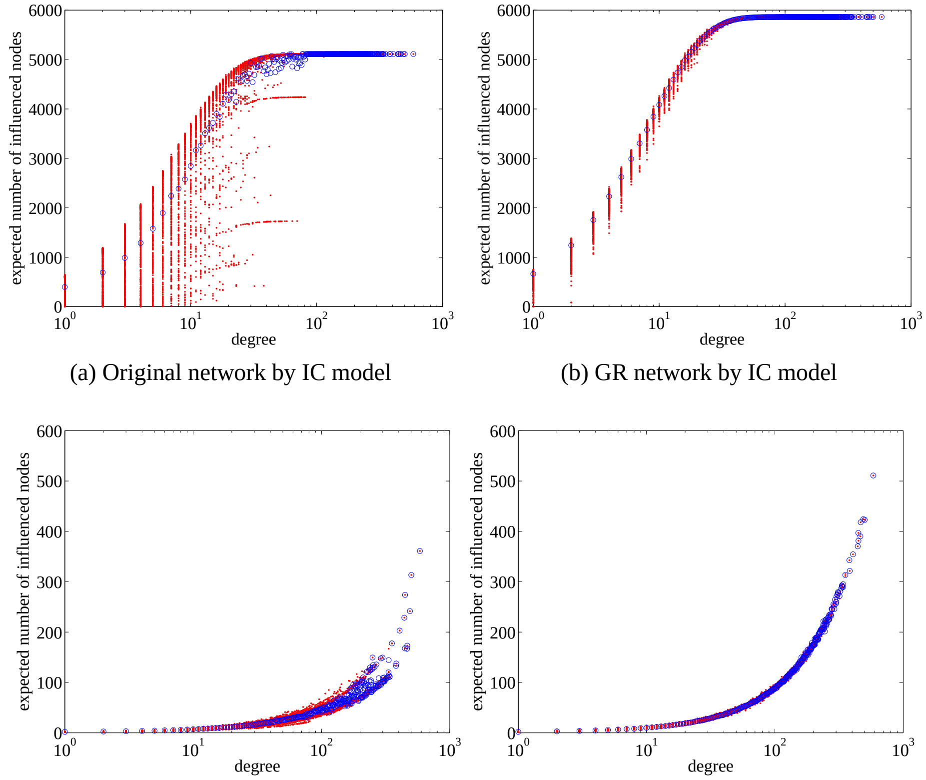
(d) GR network by LT model

Fig. 1: Comparison of information diffusion processes using blog network

4.5 Experimental Results Using Wikipedia Network

Figure 2 shows the same experimental results using the Wikipedia networks. From these results, we can derive arguments similar to those of the blog networks. Thus we consider that our arguments were substantially strengthen by these experiments.

We summarize the main points below. 1) Nodes with larger degrees influenced many other nodes, but their average curves of the IC and LT models exhibited opposite curvatures; 2) The levels of the influence degree over the GR networks were somewhat larger than those of the original networks in both of the IC and LT models; 3) The influence degree was almost uniquely determined by the degree of each node in the case of LT model using the GR network (Figure 2d); and 4) A number of lateral lines composed of the individual influence degree appeared in the case of IC model using the original network (Figure 2a).



(a) Original network by IC model

(b) GR network by IC model

(c) Original network by LT model

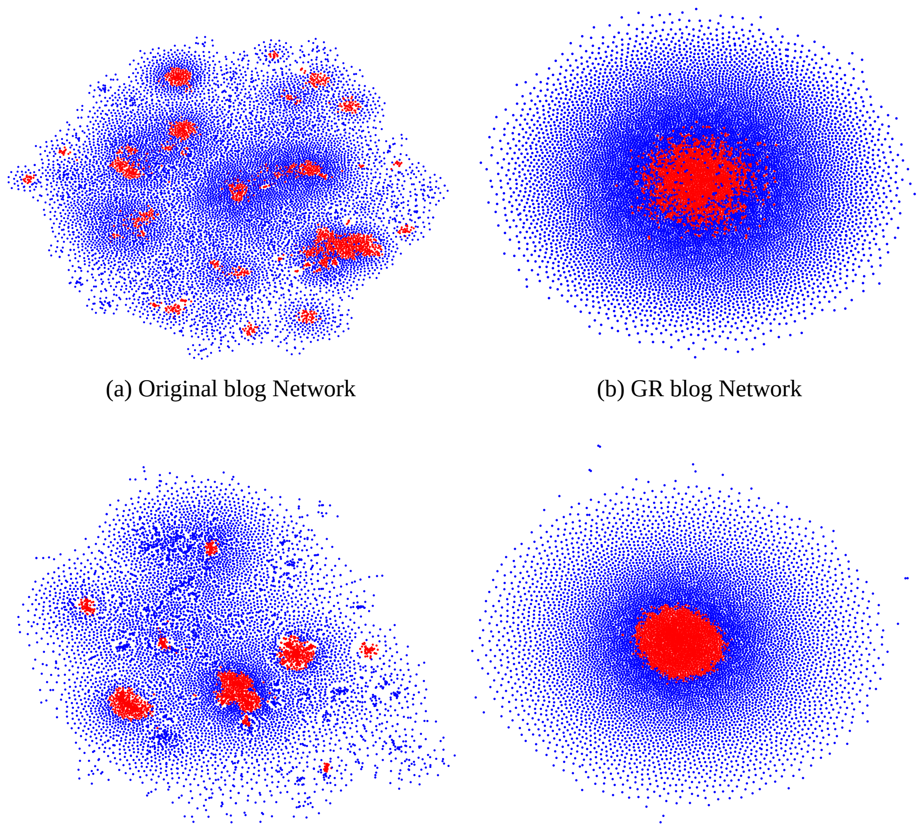
(d) GR network by LT model

Fig. 2: Comparison of information diffusion processes using wikipedia network

4.6 Community Structure Analysis

Figure 3 shows our visualization results. Here, in the case of the blog networks, since the average degree was $\overline{d} = 6.6340$, we represented the nodes belonging to the 7-core subnetwork by red points, and others by blue points. Similarly, in the case of the Wikipedia networks, since the average degree was $\overline{d} = 25.8458$, we represented the nodes belonging to the 26-core subnetwork by red points, and others by blue points. These visualization results show that the nodes of higher core order are scattered here and there in the original networks (Figures 3a and 3c), while those nodes are concentrated near the center in the GR network (Figures 3b and 3d). This clearly indicates that the transformation to GR networks changes community structure from distributed to lumped ones.

Since the main difference between the original and GR networks are their community structure, we consider that a number of lateral lines appeared in the original networks using the IC model (Figures 1a and 2a), are closely related to distributed community structure of social networks. On the other hand, we cannot observe such remarkable characteristics for the LT model (Figures 1b and 2b). In consequence, we can say that community structure more strongly affects information diffusion processes of the IC model than those of the LT model.



(a) Original blog Network

(b) GR blog Network

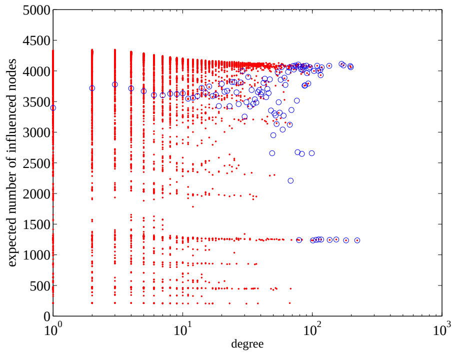
(c) Original wikipedia Network

(d) GR wikipedia Network

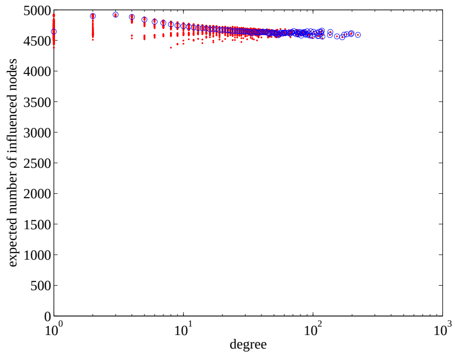
Fig. 3: Visualization of Networks

4.7 Experimental Results by Changing the Parameter Settings

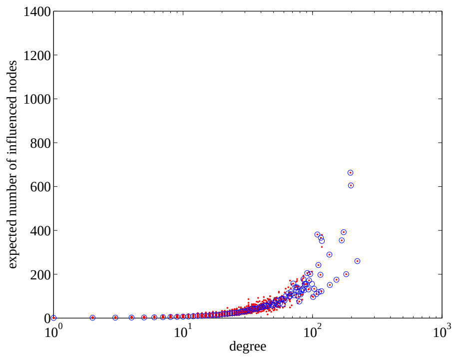
We consider to change the parameter settings to explore the intrinsic characteristics of the information diffusion models. First, we modify the IC model so as to roughly equalize the expected number of influence nodes obtained from any information source



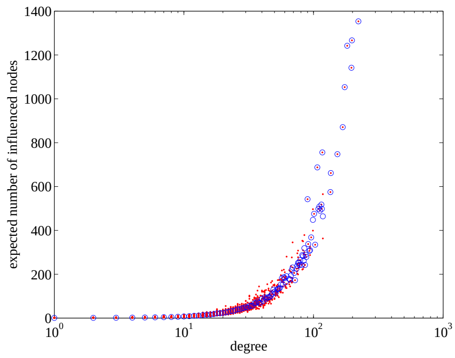
(a) Original network by IC model2



(b) GR network by IC model2



(c) Original network by LT model2



(d) GR network by LT model2

Fig. 4: Comparison of information diffusion processes using blog network by changed probability setting

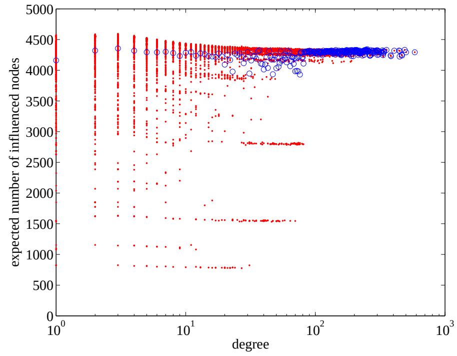
node. In the previous experiments, since we assigned the same diffusion probability to all links, the nodes with larger degrees have more advantage in diffusing information than those with smaller degrees. In order to suppress such a variation, we specified the diffusion probability for each directed link (u, v) as follows:

$$\beta_{u,v} = \frac{1}{|A_G(u)|}$$

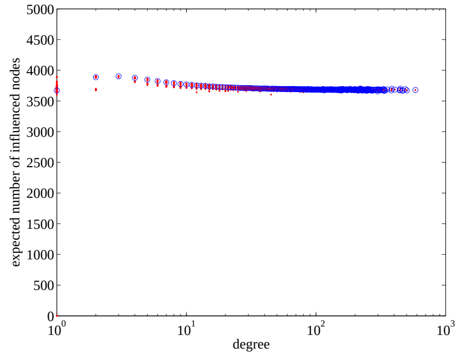
Second, we modify the LT model so as to raise the expected numbers of influenced nodes obtained form information source nodes with larger degrees, just like the IC model using the common diffusion probability. To this end, we specified the weight for each directed link (u, v) as follows:

$$\omega_{u,v} = \frac{A_G(u)}{\sum_{x \in \Gamma(v)} |A_G(x)|}$$

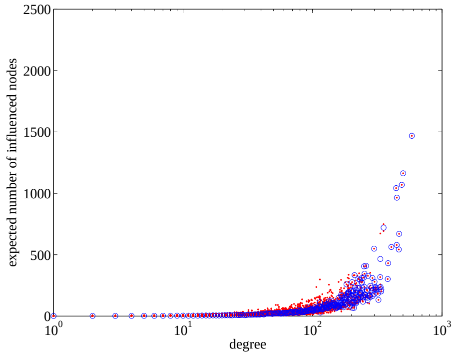
Figures 4 and 5 show the experimental results, where we employed the same experimental settings used in the previous experiments, except for the above parameter set-



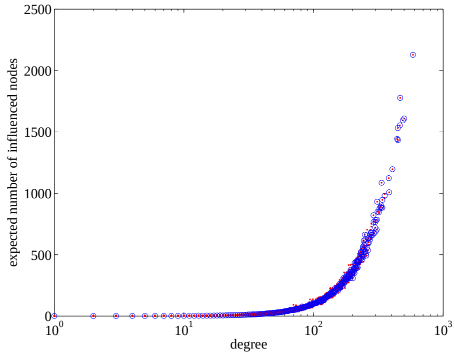
(a) Original network by IC model2



(b) GR network by IC model2



(c) Original network by LT model2



(d) GR network by LT model2

Fig. 5: Comparison of information diffusion processes using wikipedia network by changed weight setting

tings. From these results, we can similarly confirm the fact that the community structure more strongly affects information diffusion process of the IC model than those of the LT model. We believe that this observation strengthens our claim. Here, as we expected, the average curves of the IC model show almost flat lines, and the number of influence nodes in the LT model becomes larger than previous experimental results.

5 Conclusion

In this paper, we proposed a new scheme for empirical study to explore the behavioral characteristics of representative information diffusion models such as the Independent Cascade model and the Linear Threshold model on large networks with different community structure. The proposed scheme consists of two parts, i.e., GR (generalized random) network construction from an originally observed network, and plotting of the influence degree of each node based on an information diffusion model. Using large real networks, we empirically found that our proposal scheme uncovers a number of new insights. Most importantly, we showed that community structure more strongly affects information diffusion processes of the IC model than those of the LT model. Our future work includes the analysis of relationships between community structure and information diffusion models by using a wide variety of social networks.

Acknowledgments

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research, U.S. Air Force Research Laboratory under Grant No. AOARD-08-4027, and JSPS Grant-in-Aid for Scientific Research (C) (No. 20500147).

References

1. Adar, E., & Adamic, L. (2005). Tracking information epidemics in blogspace. *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 207–214).
2. Balthrop, J., Forrest, S., Newman, M. E. J., & Williampson, M. W. (2004). Technological networks and the spread of computer viruses. *Science*, *304*, 527–529.
3. Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. *Proceedings of the 13th International World Wide Web Conference* (pp. 107–117).
4. Maximizing the spread of influence through a social network. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 137–146).
5. Kimura, M., Saito, K., & Nakano, R. (2007). Extracting influential nodes for information diffusion on a social network. *Proceedings of the 22nd AAAI Conference on Artificial Intelligence* (pp. 1371–1376).
6. Kimura, M., Saito, K., & Motoda, H. (2008). Minimizing the spread of contamination by blocking links in a network. *Proceedings of the 23nd AAAI Conference on Artificial Intelligence* (pp. 1175–1180).
7. Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167–256.
8. Newman, M. E. J., Forrest, S., & Balthrop, J. (2002). Email networks and the spread of computer viruses. *Physical Review E*, *66*, 035101.
9. Newman, M. E. J. & Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E*, *68*, 036122.
10. S.B. Seidman, S. B. (1983). Network Structure and Minimum Degree, *Social Networks*, *5*, 269–287.
11. Yamada, T., Saito, K., & Ueda, N. (2003). Cross-entropy directed embedding of network data. *Proceedings of the 20th International Conference on Machine Learning* (pp. 832–839).