Community Analysis of Influential Nodes for Information Diffusion on a Social Network

Masahiro Kimura, Kazumasa Yamakawa, Kazumi Saito, and Hiroshi Motoda

***Abstract*— We consider the problem of finding influential nodes for information diffusion on a social network under the independent cascade model. It is known that the greedy algorithm can give a good approximate solution for the problem. Aiming to obtain efficient methods for finding better approximate solutions, we explore what structual feature of the underlying network is relevant to the *greedy solution* that is the approximate solution by the greedy algorithm. We focus on the SR-community strucutre, and analyze the greedy solution in terms of the SR-community structure. Using real large social networks, we experimentally demonstrate that the SR-community structure can be more strongly correlated with the greedy solution than the community structure introduced by Newman and Leicht.**

I. Introduction

Recently, considerable attention has been devoted to social network analysis [9], [14], [1], [2], [8], [13], [7], since the rise of the Internet and the World Wide Web has enabled us to collect real large social networks. Here, a social network is the network of relationships and interactions among social entities such as individuals, organizations and groups. Examples include blog networks, collaboration networks, and email networks.

A social network plays an important role for the spread of information since a piece of information can propagate from one node to another node through a link on the network in the form of "word-of-mouth" communication [3]. Thus, it is an important research issue to find influential nodes for information diffusion on a social network in terms of sociology and "viral marketing". In fact, researchers [5], [6] have recently studied a combinatorial optimization problem called the *influence maximization problem* on a network under the *independent cascade (IC) model* that is a widely-used fundamental probabilistic model of information diffusion. Here, the influence maximization problem of size k is the problem of extracting a set of k nodes to target for initial activation such that it yields the largest expected spread of information, where k is a given positive integer. Kempe *et al.* [5] experimentally showed on large collaboration networks that the greedy algorithm can give a good approximate solution for the influence maximization problem under the IC model. We refer to the approximate solution obtained by the greedy algorithm as the *greedy solution*. Using an analysis framework based on submodular functions, Kempe *et al.* [5] mathematically proved a performance guarantee of the greedy solution. Moreover, Kimura *et al.* [6] presented a method of efficiently estimating the greedy solution on the basis of bond percolation and graph theory. However, it is desirable to construct efficient methods of obtaining better approximate solutions for the influence maximization problem on a network under the IC model. Towards this aim, it is important to understand what structural feature of the underlying network is correlated with the greedy solution.

As a structual feature of a given network, we focus on the *SR-community structure* $\mathcal{U} = \langle U_m;\, m = 1, 2, 3, \cdots \rangle$ [15] that is a sequence of densely connected sets of nodes in the network. Here, the mth SR-community U_m is defined as the set of nodes in the network that maximizes the average number of links within it after removing all the links within U_j, $(j = 0, \cdots, m-1)$, where U_0 is the empty set $\emptyset$. In this paper, we analyze the greedy solution for the influence maximization problem under the IC model in terms of the SR-community structure $\mathcal{U}$. For the influence maximization problem of size k, we extract the minimal sequence of SR-communities in $\mathcal{U}$, $\mathcal{U}_k = \langle U_m; m = 1, \cdots, M_k \rangle$, such that it covers the greedy solution, and investigate the similarity between the set of nodes influenced by each node v_i in the greedy solution and the SR-community in $\mathcal{U}_k$ that corresponds to the node v_i. On the basis of this manner, we quantify the strength of the correlation between the greedy solution and the SR-community struture. Using real large social networks, we experimentally demonstrate that unlike the community structure introduced by Newman and Leicht [12], the SR-community structure can be strongly correlated wirh the greedy solution.

II. Influential Nodes for Information Diffusion

Throughout this paper, we consider a social network represented by an undirected graph, and discuss the spread of a certain information through the network under the IC model by regarding those undirected links as bidirectional ones. We call nodes *active* if they have accepted the information.

A. Independent Cascade Model

We define the IC model. In this model, the diffusion process unfolds in discrete time-steps $t \geq 0$, and it is

Masahiro Kimura is with the Department of Electronics and Informatics, Faculty of Science and Technology, Ryukoku University, Otsu 520-2194, Japan (phone: +81 77 543 7406; fax: +81 77 543 7749; email: kimura@rins.ryukoku.ac.jp).

Kazumasa Yamakawa is with the Division of Electronics and Informatics, Graduate School of Science and Technology, Ryukoku University, Otsu 520-2194, Japan (email: t07m025@mail.ryukoku.ac.jp).

Kazumi Saito is with the School of Administration and Informatics, University of Shizuoka, Shizuoka 422-8526, Japan (email: k-saito@u-shizuoka-ken.ac.jp).

Hiroshi Motoda is with the Institute of Scientific and Industrial Research, Osaka University, Osaka 567-0047, Japan (email: motoda@ar.sanken.osaka-u.ac.jp).

assumed that nodes can switch from being inactive to being active, but cannot switch from being active to being inactive. Given an initial set X of active nodes, we assume that the nodes in X have first become active at step 0, and all the other nodes are inactive at step 0. We specify a real value $\beta_{u,v} \in [0,1]$ for each directed link (u,v) in advance. Here, $\beta_{u,v}$ is referred to as the *propagation probability* through link (u,v).

When an initial set X of active nodes is given, the diffusion process proceeds in the following way. When node u first becomes active at step t, it is given a single chance to activate each currently inactive neighbor v, and succeeds with probability $\beta_{u,v}$. If u succeeds, then v will become active at step $t+1$. If multiple parents of v first become active at step t, then their activation attempts are sequenced in an arbitrary order, but performed at 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 X, let $\sigma(X)$ denote the expected number of active nodes at the end of the random process in the IC model. We call $\sigma(X)$ the *influence degree* of initial active set X.

B. Influence Maximization Problem

We consider the influence maximization problem of size k under the IC mode. Let S be the set of all the nodes in the network. The problem is defined as follows: Given a positive integer k, find a set X_k^* of k nodes to target for initial activation such that $\sigma(X_k^*) \geq \sigma(Y)$ for any set Y of k nodes. To approximately solve this optimization problem, we consider the following greedy algorithm:

1) Set $X \leftarrow \emptyset$.
2) **for** $i = 1$ to k **do**
3) Choose a node $v_i \in V$ maximizing $\sigma(X \cup \{v\})$, $(v \in S \setminus X)$.
4) Set $X \leftarrow X \cup \{v_i\}$.
5) **end for**

Let S_k denote the set of k nodes obtained by this algorithm. We call S_k the *greedy solution* of the influence maximization problem of size k.

Using large collaboration networks, Kempe *et al.* [5] experimetally demonstrated that the greedy solution S_k outperforms the approximate solutios obtained by the high-degree and centrality heuristics that are commonly used in the sociology literature. It is also known that

$$\sigma(S_k) \geq \left(1 - \frac{1}{e}\right) \sigma(X_k^*),$$

that is, a performance guarantee of the greedy solution S_k is obtained [5]. For any initial active set X, a good estimate of $\sigma(X)$ was conventionally obtained by simulating the random process of the IC model many times. Thus, any straightforward method to estimate the greedy solution S_k needed a large amount of computation on a large network. However, Kimura *et al.* [6] gave an efficient method for estimating S_k on the basis of bond percolation and graph theory. In this paper, using their method, we estimate the greedy solution S_k.

III. SR-community Structure

In this section, we define the SR-community structure, and describe a method for efficiently estimating it according to the work of Saito *et al.* [15].

A. Definition

Let $\mathbf{A}$ be the adjacency matrix of a network, and let

$$S = \{1, \cdots, N\}$$

be the set of all the nodes in the network. Namely, each (i,j)-element of the adjacency matrix, denoted by $A(i,j)$, is set to 1 if there exists a link (edge) between nodes i and j; otherwise 0. In this paper, we focus on undirected graphs without self-connections, i.e., $A(i,j) = A(j,i)$, $A(i,i) = 0$, $(i,j = 1, \cdots, N)$. For any subset of nodes, $T \subset S$, we can define *the average number of links within* T as follows:

$$G(T) = \frac{1}{2} \sum_{i \in T} \sum_{j \in T} \frac{A(i,j)}{|T|}, \tag{1}$$

where $|T|$ stands for the number of elements in T. First, let U_1 denote the subset of S that maximizes the average number of links within it (see, (1)). Next, for the network constructed through removing all the links within U_1 from the original network, let U_2 denote the subset of S that maximizes the average number of links within it (see, (1)). Next, for the network constructed thorugh removing all the links within U_1 and U_2 from the original network, let U_3 denote the subset of S that maximizes the average number of links within it (see, (1)). By repeatedly performing these procedures, we define the sequence of subsets of S,

$$\mathcal{U} = \langle U_m;\, m = 1, 2, 3, \cdots \rangle.$$

Here, $\mathcal{U}$ is called the *SR-community structure* of the original network, and each U_m is referred to as the *mth SR-community*. Note that the SR-community structure $\mathcal{U}$ represents a structual feature of the network.

In the case of a large network, any straightforward method for detecting the SR-community structure is likely to suffer from combinatorial explosion. To cope with such a large network, we employ the method presented by Saito *et al.* [15].

B. Relaxation problem

For a subset T of S , we define an N dimensional indicator vector $\mathbf{q}$ by setting $q(i) = 1$ if $i \in T$; otherwise $q(i) = 0$. Then we can rewrite (1) as follows:

$$G(\mathbf{q}) = \frac{1}{2} \frac{\mathbf{q}^T \mathbf{A} \mathbf{q}}{\mathbf{q}^T \mathbf{q}}, \tag{2}$$

where $\mathbf{q}^T$ stands for a transposed vector of $\mathbf{q}$. Now we consider a relaxation problem by letting $\mathbf{q}$ take continuous values. Then, according to the Rayleigh-Ritz theorem [4],

the solution of maximizing $G(\mathbf{q})$ is given by the principal eigenvector $\mathbf{q}^*$ of the adjacency matrix $\mathbf{A}$.

In order to obtain the eigenvector $\mathbf{q}*$, we employ the following procedure based on the power iteration [4].

E1. Initialize $\mathbf{q}^{(0)} = (1, \cdots, 1)^T$, and set $\tau \leftarrow 1$;
E2. Calculate $\widetilde{\mathbf{q}} = \mathbf{A}\mathbf{q}^{(\tau-1)}$ and $\mathbf{q}^{(t)} = \widetilde{\mathbf{q}}/\max_i \widetilde{q}_i$;
E3. Terminate if $\max_i |q^{(\tau)}(i) - q^{(\tau-1)}(i)| < \varepsilon$;
E4. Set $\tau \leftarrow \tau + 1$, and return to **E2.**.

Here a small positive parameter ε controls the termination condition, and we can obtain the final solution as $\mathbf{q}^* = \mathbf{q}^{(\tau)}$ after its termination. Since all the elements of $\mathbf{A}$ and $\mathbf{q}^{(0)}$ have non-negative values, we can guarantee that all the elements of $\widetilde{\mathbf{q}}$ also have non-negative values after any number of iterations. Moreover, due to the scaling operation in **E2**, we can guarantee that $0 \leq q^{(\tau)}(i) \leq 1$ for any τ and i. Thus we consider that the above formulation gives one of desirable relaxation solutions to the original problem.

C. Quantization problem

By ranking nodes according to the values of eigenvector elements, we can obtain a list of nodes, $R = [r(1), \cdots, r(N)]$, where $r(i)$ stands for a mapping from ranks to nodes. Note that $q^*(r(i)) \geq q^*(r(i+1))$ for any i. By considering a set of the top h nodes,

$$T(h) = \{r(i) : i = 1, \cdots, h\}, \tag{3}$$

we can calculate the average number of links within $T(h)$ as follows:

$$G(h) = \sum_{i=1}^{h-1} \sum_{j=i+1}^{h} \frac{A(r(i), r(j))}{h}. \tag{4}$$

In our method, instead of directly solving (1), we compute a node set $T(h^*)$, where h^* maximizes (4).

In order to efficiently calculate h^*, we utilize the following update formula:

$$G(h+1) = G(h) + \frac{\Delta(h+1) - G(h)}{h+1}, \tag{5}$$

where $\Delta(h+1)$ stands for the increment by adding node $r(h+1)$, calculated by

$$\Delta(h+1) = \sum_{j=1}^{h} A(r(j), r(h+1)). \tag{6}$$

Note that $G(1) = 0$. The above procedure can be summarized as follows.

F1. Compute $r(i)$ by sorting elements of $\mathbf{q}*$;
F2. Calculate $G(2), \cdots, G(N)$ by using (5) and (6);
F3. Output $T(h^*)$ such that $h^* = \arg\max_h G(h)$;

D. Detection algorithm

By repeatedly performing the above procedures, M times, we can detect M densely connected portions for a given network as follows.

G1. Repeat the following steps for $m = 1$ to M;
G2. Calculate $\mathbf{q}_m^*$ using **E1** to **E4**;
G3. Calculate T_m^* using **F1** to **F3**;
G4. Set $A(i,j) = 0$ if $i, j \in T_m^*$.

Here, the number M of communities is determined by a user. We estimate the mth SR-community U_m as T_m^* for any integer m with $1 \leq m \leq M$.

IV. Community Analysis of Influential Nodes

For a given network, we consider the influence maximization problem of size k under the IC model. Let $S_k = \{v_i;\ i = 1, \cdots, k\}$ be the greedy solution, and let $\mathcal{U} = \langle U_m;\ m = 1, 2, 3, \cdots \rangle$ be the SR-community structure of the network. We analyze the greedy solution S_k in terms of the SR-community structure $\mathcal{U}$.

First, we extract the minimal sequence of SR-communities in $\mathcal{U}$ such that it covers the greedy solution S_k,

$$\mathcal{U}_k = \langle U_m;\ m = 1, \cdots, M_k \rangle,$$

that is, M_k is the minimal integer M satisfying

$$\bigcup_{m=1}^{M} U_m \supset S_k.$$

Note that $\mathcal{U}_k$ can be regarded as a rough approximation to the greedy solution S_k. We call M_k the *SR-covering number* of the greedy solution S_k. For any $v_i \in S_k$, let $\alpha(v_i)$ denote the minimal integer m satisfying $U_m \ni v_i$. $U_{\alpha(v_i)}$ is reffered to as the SR-community that corresponds to the node v_i.

Next, for any $v_i \in S_k$ and a real value $p \in [0, 1]$, we consider the *influence set* $H(v_i, p)$ of v_i with probability p. Here, $H(v_i, p)$ is the set of nodes v in the network such that when $\{v_i\}$ is the initial active set, the probability that v is active at the end of the diffusion process under the IC model is more than p. Note that $v_i \in H(v_i, p) \subset H(v_i, p')$ if $0 \leq p' \leq p \leq 1$.

We investigate the correlation between the greedy solution S_k and the SR-community structure $\mathcal{U}$. In terms of F-measure, we quantify the similarity between an influence set $H(v_i, p)$ of each node v_i in the greedy solution S_k and the SR-community $U_{\alpha(v_i)}$ that correspond to v_i, that is, we measure how close the sets $H(v_i, p)$ and $U_{\alpha(v_i)}$ are by

$$F_0(p; v_i) = 200\, \frac{\left|H(v_i, p) \cap U_{\alpha(v_i)}\right|}{|H(v_i, p)| + \left|U_{\alpha(v_i)}\right|}. \tag{7}$$

Moreover, we quantify the strength of the correlation between the greedy solution S_k and the SR-community structure $\mathcal{U}$ as follows:

$$F(k) = \frac{1}{k} \sum_{i=1}^{k} F_1(v_i), \tag{8}$$

where

$$F_1(v_i) = \max_{0 \leq p \leq 1} F_0(p; v_i),\ \ (i = 1, \cdots, k).$$

V. Experimental Evaluation

Using real large networks, we experimentally evaluate the strength of the correlation between the greedy solution of the influence maximization problem under the IC model and the SR-community structure. Let $S_k = \{v_1, \cdots, v_k\}$ be the greedy solution for the influence maximization problem of size k.

A. Network Datasets

In the evaluation experiments, we should desirably use large networks that exhibit many of the key features of real social networks. Here, we report on the experimental results for two different datasets of such real networks.

First, we employed a trackback network of blogs, since a piece of information can propagate from one blog author to another blog author through a trackback. Since bloggers discuss various topics and establish mutual communications by putting trackbacks on each other's blogs, we regarded a link created by a trackback as a biderectional link for simplicity. By tracing ten steps ahead the trackbacks from the blog of the theme "JR Fukuchiyama Line Derailment Collision" in the site "goo"[1], we collected a large connected trackback network in May, 2005. This network was an undirected graph of $12,047$ nodes and $39,960$ links. This network showed the so-called "power-law" degree distribution that most real large networks exhibit. Here, the degree distribution is the distribution of the number of links for every node. We refer to this network data as *the blog network dataset*.

Next, we employed 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. We refer to this network data as *the Wikipedia network dataset*. Here, the total numbers of nodes and links were $9,481$ and $122,522$, respectively.

Newman and Park [11] observed that social networks represented as undirected graphs generally have the following two statistical properties unlike non-social networks. First, they show positive correlations between the degrees of adjacent nodes. Second, they have much higer values of the *clustering coefficient* than the corresponding *configuration models* (i.e., random network models). Here, the clustering coefficient C for an undirected graph is defined by

$$C = \frac{3 \times \text{number of triangles on the graph}}{\text{number of connected triples of nodes}},$$

where a "triangle" means a set of three nodes each of which is connected to each of the others, and a "connected triple" means a node connected directly to an unordered pair of others. Note that in terms of sociology, C measures the probability that two of your friends will also be friends of one another. Given a degree distribution, the corresponding configuration model of random network is defined as the ensemble of all possible graphs that possess the degree distribution, with each having equal weight. The value of C for the configuration model can be exactly calculated [10]. For the Wikipedia network, the value of C of the corresponding configuration model was 0.046, while the actual measured value of C was 0.39. Moreover, the degrees of adjacent nodes were positively correlated for the Wikipedia network dataset. Therefore, we consider that the Wikipedia network dataset can be used as an example of social network.

[1] http://blog.goo.ne.jp/usertheme/

B. A Comparison Method

In order to quantitatively evaluate the strength of the correlation between the greedy solution for the influence maximization problem under the IC model and the SR-community structure, we employ the community structure obtined by the method of Newman and Leicht [12] as a baseline.

Given an integer k, the method of Newman and Leicht [12] divides the set $S = \{1, \cdots, N\}$ of nodes in the network into k communities, that is, k disjoint subsets of S, according to some probabilistic mixture model that is a probabilistic mixture of multinomial distributions. More specifically, their method is as follows: First, a probabilistic generative model for network is given. Namely, the probability that a network with adjacency matirix $\mathbf{A}$ appears is defiend by

$$P(\mathbf{A} \,|\, \lambda, \theta) = \prod_{n=1}^{N} P(\mathbf{A}(n,:) \,|\, \lambda, \theta),$$

where $\mathbf{A}(n,:)$ denotes the nth row vector of $\mathbf{A}$,

$$\begin{aligned} \lambda &= \{\lambda_\ell;\ \ell = 1, \cdots, k\}, \\ \theta &= \{\theta_{\ell,j};\ \ell = 1, \cdots, k,\ j = 1, \cdots, N\} \end{aligned}$$

are sets of parameters, and

$$P(\mathbf{A}(n,:) \,|\, \lambda, \theta) = \sum_{\ell=1}^{k} \lambda_\ell P(\mathbf{A}(n,:) \,|\, \ell, \theta),$$

$$P(\mathbf{A}(n,:) \,|\, \ell, \theta) \propto \prod_{j=1}^{N} (\theta_{\ell,j})^{A(n,j)},$$

for $\ell = 1, \cdots, k$ and $n, j = 1, \cdots, N$. Here, each λ_ℓ is the mixture weight (prior probability) of the ℓth community, and

$$\lambda_\ell > 0,\ (\ell = 1, \cdots, k),\ \ \sum_{\ell=1}^{k} \lambda_\ell = 1.$$

Also, each $\theta_{\ell,j}$ is the probability that the jth node connects with a node beloging to the ℓth community, and

$$\theta_{\ell,j} > 0,\ \ \sum_{j=1}^{N} \theta_{\ell,j} = 1,$$

for $\ell = 1, \cdots, k$ and $j = 1, \cdots, N$. By performing the maximal likelihood estimation using the EM algorithm, we estimate the values of λ and θ. Then, applying Bayes' rule, we define the community label $\ell^*(n)$ for each node n as

$$\ell^*(n) = \arg \max_{1 \le \ell \le k} P(\ell \,|\, \mathbf{A}(n,:), \lambda, \theta).$$

For the greedy solution $S_k = \{v_1, \cdots, v_k\}$, we detect the set of k communities,

$$\mathcal{Z}_k = \{Z_1, \cdots, Z_k\},$$

by using the method of Newman and Leicht. For every v_i, we define $\gamma(v_i)$ by the condition $Z_{\gamma(v_i)} \ni v_i$. In the same way as the SR-community structure, we quantify the strength of the correlation between S_k and $\mathcal{Z}_k$ by $F(k)$ (see, (8)). Here, we modify the definition of $F(k)$ through changing each $F_0(p; v_i)$ (see, (7)) to

$$F_0(p; v_i) = 200\,\frac{\left|H(v_i, p) \cap Z_{\gamma(v_i)}\right|}{|H(v_i, p)| + \left|Z_{\gamma(v_i)}\right|}.$$

C. Experimental Settings

In our experiments, we assigned a uniform probability β to the propagation probability $\beta_{u,v}$ for any directed link (u, v) of the network. As investigate by Leskovec *et al.* [7], it seems that large cascades of information diffusion happen rarely. Thus, we examined the IC model with relatively small β according to Kempe *at al.* [5].

We estimated the greedy solution $S_k = \{v_1, \cdots, v_k\}$ using the method of Kimura *et al.* [6] with the parameter value $10,000$. Here, the parameter represents the number of bond percolation processes for estimating the influence degree $\sigma(X)$ of a given initial active set X. Also, we estimated the influence set $H(v_i, p)$ of node v_i with probability p through $300,000$ simulations of the IC model.

D. Experimental Results

We describe the results for the experiments using the blog network dataset and the Wikipedia network dataset.

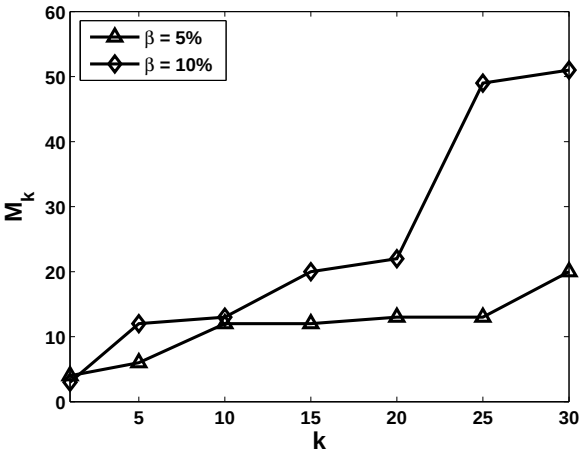

Fig. 1. SR-covering number M_k of greedy solution S_k on the blog network dataset.

Figures 1 and 2 plot the SR-covering number M_k of the greedy solution S_k with respect to k on the blog network dataset and the Wikipedia network dataset, respectively. For almost all k, we observe that the larger the value of propagation probability β is, the larger the SR-covering number M_k of S_k is.

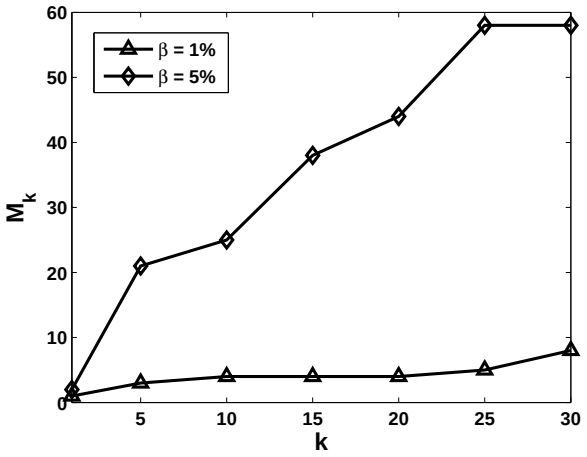

Fig. 2. SR-covering number M_k of greedy solution S_k on the Wikipedia network dataset.

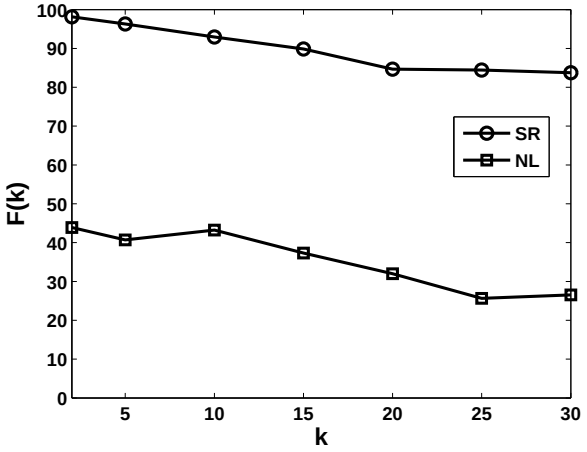

Fig. 3. Strength $F(k)$ of correlation with greedy solution S_k on the blog network dataset ($\beta = 5\%$).

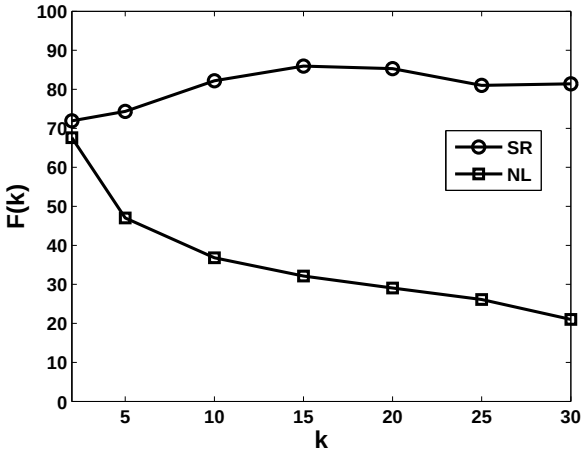

Fig. 4. Strength $F(k)$ of correlation with greedy solution S_k on the blog network dataset ($\beta = 10\%$).

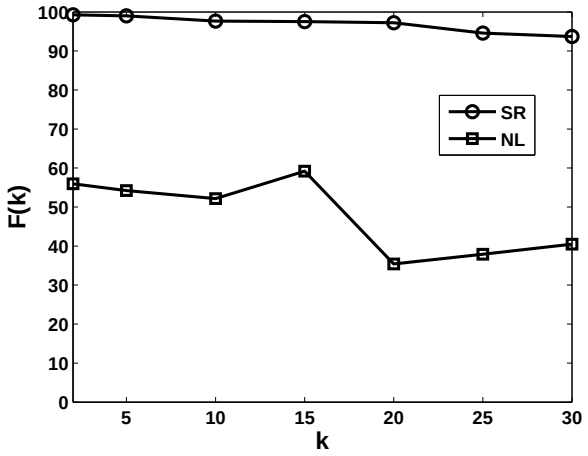

Fig. 5. Strength $F(k)$ of correlation with greedy solution S_k on the Wikipedia network dataset ($\beta = 1\%$).

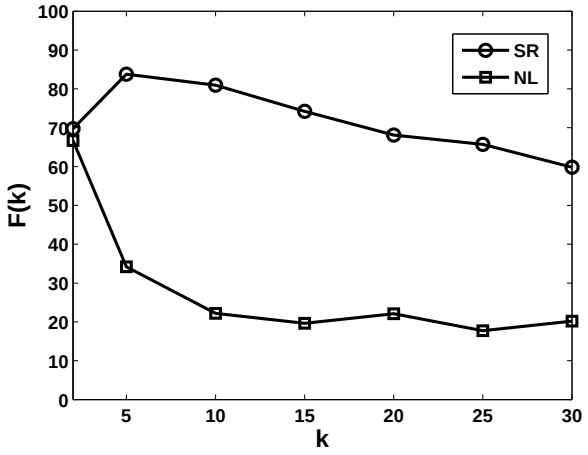

Fig. 6. Strength $F(k)$ of correlation with greedy solution S_k on the Wikipedia network dataset ($\beta = 5\%$).

Figures 3, 4, 5, and 6 plot the strength $F(k)$ of correlation with the greedy solution S_k with respect to k, $(2 \leq k \leq 30)$. In Figures 3, 4, 5, and 6, the circles indicate the strength of the correlation between the greedy solution and the SR-community structure (SR), and the squares indicate the strength of the correlation between the greedy solution and the community structure obtained by the method of Newman and Leicht (NL). Figures 3 and 4 show the results for the blog network dataset, and Figures 5 and 6 show the results for the Wikipedia network dataset. These results imply that for the IC model with relatively small propagation probability β, the SR-community structure can be more strongly correlated with the greedy solution than the community structure introduced by Newman and Leicht.

VI. Concluding Remarks

Aiming to obtain efficient methods for finding better approximate solutions for the influence maximization problem on a social network under the IC model, we have explored what structual feature of the undelying network is correlated with the greedy solution. We have focused on the SR-community structure of the network, and analyzed the greedy solution in terms of the SR-community structure. Using real large social networks including a blog network, we have experimentally demonstrated that in comparison with the community structure introduced by Newman and Leicht, the SR-community structure can be strongly correlated with the greedy solution.

On the other hand, extensive verification of this proposition with various real social networks remains an important task. However, we have already made substantial progress, and we are encouraged by our initial results.

Acknowledgment

This work was partly supported by Asian Office of Aerospace Research and Development, The U.S. Air Force Research Laboratory under Grant No. AOARD-08-4027.

References

[1] E. Adar and L. A. Adamic, "Tracking information epidemics in blogspace," *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, Compiègne, France, 2005, pp. 207–214.

[2] P. Domingos, "Mining social networks for viral marketing," *IEEE Intelligent Systems*, vol. 20, pp. 80–82, 2005.

[3] K. J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Letters*, vol. 12, pp. 211–223, 2001.

[4] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, Maryland, USA, 1989.

[5] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, 2003, pp. 137–146.

[6] M. Kimura, K. Saito, and R. Nakano, "Extracting influential nodes for information diffusion on a social network," *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, Vancouver, British Columbia, Canada, 2007, pp. 1371-1376.

[7] J. Leskovec, A. Singh, and J. Kleinberg, "Patterns of influence in a recommendation network," *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Singapore, 2006, pp. 380–389.

[8] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and role discovery in social networks," *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland, 2005, pp. 786–791.

[9] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 404–409, 2001.

[10] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.

[11] M. E. J. Newman and J. Park, "Why social networks are different from other types of networks", *Physical Review E*, vol. 68, 036122, 2003.

[12] M. E. J. Newman and E. A. Leicht, "Mixture models and exploratory analysis in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 9564–9569, 2007.

[13] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814–818, 2005.

[14] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing", *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, 2002, pp. 61–70.

[15] K. Saito, N. Ueda, M. Kimura, K. Kazama, and S. Sato, "Filtering search engine spam based on anomaly detection approach," *Proceedings of the KDD2005 Workshop on Data Mining Methods for Anomaly Detection*, Chicago, Illinois, USA, 2005, pp. 62–66.