Efficient Analysis of Node Influence Based on SIR Model over Huge Complex Networks

Masahiro Kimura*, Kazumi Saito†, Kouzou Ohara‡ and Hiroshi Motoda§¶
*Department of Electronics and Informatics, Ryukoku University, Japan
†School of Administration and Informatics, University of Shizuoka, Japan
‡Department of Integrated Information Technology, Aoyama Gakuin University, Japan
§Institute of Scientific and Industrial Research, Osaka University, Japan
¶ School of Computing and Information Systems, University of Tasmania, Australia

***Abstract*—Node influence is yet another useful concept to quantify how important each node is over a network and can share the same role that other centrality measures have. It can provide new insight into the information diffusion phenomena such as existence of epidemic threshold which the other topology-based centralities cannot do. We focus on information diffusion process based on the SIR model, and address the problem of efficiently estimating the influence degree for all the nodes in the network. The proposed approach is a further improvement over the existing work of the bond percolation process [1], [2] which was demonstrated to be very effective, *i.e.*, three orders of magnitude faster than direct Monte Carlo simulation, in approximately solving the influence maximization problem under a greedy search strategy. We introduce two pruning techniques which improve computational efficiency by an order of magnitude. This is a generic approach for the SIR model setting and can be instantiated to any specific diffusion model. It does not require any approximations or assumptions to the model, *e.g.*, small diffusion probability, shortest path, maximum influence path, etc., that were needed in the existing approaches. We demonstrate its effectiveness by extensive experiments on two large real social networks. Main finding includes that different network structures have different epidemic thresholds and the node influence can identify influential nodes that the other centrality measures cannot.**

I. Introduction

Studies of the structure and functions of large complex networks have attracted a great deal of attention in many different fields such as sociology, biology, physics and computer science [3]. Pursuing fundamental network analysis, it has been recognized that developing methods/tools for quantifying the importance of each individual node in a network is crucially important. Networks mediate the spread of information, and it sometimes happens that small initial shocks cascade to affect large portions of networks [4]. Such information cascade phenomena are observed in many situations: for example, cascading failures can occur in power grids (*e.g.*, the August 10, 1996 accident in the western US power grid), diseases can spread over networks of contacts between individuals, innovations and rumors can propagate through social networks, and large grass-roots social movements can begin in the absence of centralized control (*e.g.*, the Arab Spring). Understanding these phenomena involves dynamic analysis of diffusion process. Thus, the node influence with respect to information cascade is a useful measure of node importance. It shares the same role that existing centrality measures have. The well-known centrality measures include, but not limited to, degree centrality [5], eigenvector centrality [6], Katz centrality [7], PageRank [8], closeness centrality [5], betweenness centrality [5], and topological centrality [9]. Notable feature of these existing measures is that they all are defined by only network topology. Node influence is different from them in that it is defined through dynamical processes of a network. Thus, it can provide new insight into the information diffusion phenomena such as existence of epidemic threshold which the topology-based centrality measures can never do.

Basic models of information diffusion over a network often assume that each node has three states, *susceptible*, *infective*, and *recovered* from the analogy of epidemiology. A node in the susceptible state means that it has not yet been influenced with the information. A node in the infective state means that it is influenced with the information, and can propagate the information to its neighbor nodes. A node in the recovered state means that it can no longer propagate the information to its neighbor nodes once it has been influenced with the information, *i.e.*, immune. The *SIR model* is typical among such basic models and well exploited in many fields [3]. Here, the SIR model is a discrete-time stochastic process model, and assumes that a susceptible node becomes infective with a certain probability when its neighbor nodes get infective, and becomes subsequently recovered. In particular, it is known that the SIR model on a network can be exactly mapped onto a *bond percolation process* on the same network [3], [10]. The dynamical behaviors of the SIR model have been widely studied in physics literature. One such important analysis is to examine the *epidemic threshold* p_G^* of a network G, where most nodes of the network remain uninfected (*i.e.*, a small outbreak) if the probability that a susceptible node receives information from its infective neighbor is smaller than p_G^*, and the number of infected (recovered) nodes rapidly increase (*i.e.*, a large outbreak) as the probability becomes greater than p_G^* [3]. We must be able to estimate node influence very efficiently to make this kind of analysis feasible. In this paper, we focus on the node influence based on the SIR model, and regard it as one of the centrality measures and refer to it as the *influence degree centrality* for convenience sake.

Let $G = (V, E)$ be a directed network, where V and E ($\subset V \times V$) stand for the sets of all nodes and links, respectively. For the SIR model over G, the *influence degree* $\sigma_G(v)$ of a node $v \in V$ is defined as the expected number of recovered nodes at the end of the information diffusion process (i.e., when there are no nodes in the infective state), assuming that at the initial time $t = 0$, only v is in infective state and all other nodes are in susceptible state. In order to examine the influence degree centrality in G, it is necessary to estimate the influence degree $\sigma_G(v)$ for every single node $v \in V$. We refer to $\sum_{v \in V} \sigma_G(v)/|V|$ as the *average influence degree* of G. In order to examine the epidemic threshold of G, we must further calculate the average influence degree of G for various values of diffusion probability of the SIR model. Note that it is difficult to calculate the influence degree exactly since the SIR model is defined by a stochastic process [1], [13], [2]. In general, the influence degree is approximately estimated through a number of simulations while the existing centrality measures described above are exactly calculated. Thus, it is an important research issue to estimate the influence degrees $\{\sigma_G(v) \,|\, v \in V\}$ efficiently.

In this paper, we propose an improved method of efficiently estimating the influence degrees of all the nodes in network G, $\{\sigma_G(v) \,|\, v \in V\}$ under the SIR model setting. Estimating influence degree is a sub problem in the *influence maximization problem*, which has recently attracted tremendous interest in the field of social network mining [11]. The task of the influence maximization problem is to identify a limited number of seed nodes that maximize the expected spread of influence over G. Kempe et al [10] first formalized this problem and presented a good solution by using a greedy search strategy. Since then, many researchers have proposed various techniques for improving the efficiency in finding high-quality approximate solutions [1], [12], [13], [14], [15], [16], [17], [18]. These techniques include both of those that aim at improving the efficiency of estimating the expected spread for a given seed node set and those that aim at improving the efficiency of the search for the seed node set. The proposed method belongs to the former. Thus, it can naturally be applied to the influence maximization problem through the greedy search. It can also be utilized for identifying super-mediators of information diffusion in social networks [19].

Many of the techniques cited above are designed for a specific diffusion model, *e.g.*, independent cascade or linear threshold models, and introduce approximations and/or assumptions to the model chosen such as assuming that the diffusion probability is small enough to allow for linear approximation, considering only the shortest diffusion path or the maximum influence path between a pair of nodes is enough, approximating the diffusion path in the original network to be a DAG for information spread, etc. To the best of our knowledge, two groups of work, one [1], [2] (called bond percolation) and the other [13] (called new greedy algorithm) are the only ones that do not introduce any approximations and/or assumptions to the model. Both use the same idea, and in this paper we call it *BP method* for short.

The BP method was shown to be very efficient, three orders of magnitude faster than direct Monte Carlo simulation in computing the node influence degree [1], [2]. Our proposed method for estimating the influence degree centrality $\{\sigma_G(v) \,|\, v \in V\}$ in network G makes it even faster by an order of magnitude by introducing two new pruning techniques: the *redundant-edge pruning (REP) technique* and the *marginal-component pruning (MCP) technique*. The REP technique prunes redundant edges for reachability analysis among three vertices and the MCP technique recursively prunes vertices of in-degree 1 or out-degree 1 from the quotient graph obtained by decomposing the graph generated by the corresponding bond percolation process into the strongly connected components (SCCs). We extensively evaluate the proposed method using two large real social networks, compare the computation time,[1] and show that the proposed method significantly outperforms the existing BP method. The MCP technique is found to be more effective than the REP technique. Use of both techniques is always better than the single use of either technique. The proposed method inherits the good feature of the BP method. It is a generic framework to estimate the influence degree centrality under the SIR model setting without need for any approximations and assumptions. With this improved efficiency it is now possible to estimate the node influence of every single node of a network with one million nodes and 157 millions links and analyze the existence of epidemic threshold. We further confirmed that the node influence identifies nodes that are deemed indeed influential which are not identifiable by the existing centrality measures.

II. BP Method

We briefly revisit the BP method (see [2] for more detail). A bond percolation process on a given network $G = (V, E)$ is the process in which each link of G is stochastically designated either "occupied " or "unoccupied" according to some probability distribution. The occupation probability distribution is determined according to the assumed information diffusion model and its associated parameter values. Now, we consider M times of bond percolation processes. Let E_m ($\subset E$) denote the set of occupied links at the m-th bond percolation process and let G_m denote the network (V, E_m). For any node $v \in V$, we define $\bar{\sigma}_G(v)$ by

$$\bar{\sigma}_G(v) = \frac{1}{M} \sum_{m=1}^{M} |R_{G_m}(v)|, \tag{1}$$

where $R_{G_m}(v)$ stands for the set of *reachable* nodes from v on G_m, and $|R_{G_m}(v)|$ is the number of nodes in $R_{G_m}(v)$. Here, we say that a node $w \in V$ is *reachable* from node v on G_m if there exists a path from v to w in the network G_m. It is known [3] that the influence degree $\sigma_G(v)$ can be estimated by $\bar{\sigma}_G(v)$ with a reasonable accuracy if M is sufficiently large. [2] Here note that the bond percolation technique decomposes each network G_m into its SCCs, where an SCC (strongly connected component) is a maximal subset C of V such that for all v, $w \in C$ there is a path from v to w on G_m. Note that $R_{G_m}(v) = R_{G_m}(w)$ $(v, w \in C)$. Thus, we can obtain $R_{G_m}(v)$ for any node $v \in V$ by calculating $R_{G_m}(v)$ for only one node v in each component C. Let $Q_m = (C_m, \mathcal{E}_m)$ be the quotient graph obtained by the SCC decomposition of $G_m = (V, E_m)$, where C_m is the set of all the SCCs of G_m, and $\mathcal{E}_m$ ($\subset C_m \times C_m$) is the set of edges in Q_m, *i.e.*, $(C, D) \in \mathcal{E}_m$ if there exist some pair of nodes $v \in C$ and $w \in D$ which satisfies $(v, w) \in E_m$. Note that the quotient graph Q_m is a DAG (directed acyclic graph). For each component $C \in C_m$, we can also consider the set of *reachable* components from C on Q_m, which is denoted by $R_{Q_m}(C)$. Here, a component $D \in C_m$ is an element of $R_{Q_m}(C)$ when there exists a path from

[1] The estimation accuracy of $\{\sigma_G(v) \,|\, v \in V\}$ is the same because of no new approximations and assumptions introduced.

[2] It is shown that setting M to a few thousands usually gives good accuracy in experiments using real social networks (see [2]).

vertex C to vertex D on the graph Q_m. Then, for any node $v \in C$, we can calculate the number of reachable nodes from v on the network G_m by

$$|R_{G_m}(v)| = |C| + \sum_{D \in R_{Q_m}(C)} |D|. \tag{2}$$

In case of the MCP technique as described later, this equation is replaced as follows:

$$|R_{G_m}(v)| = h_m(C) + \sum_{D \in R_{Q_m}(C)} h_m(D), \tag{3}$$

where $h_m(D)$ is initially set to $h_m(D) = |D|$ for any component $D \in \mathcal{C}_m$, and it is to be updated iteratively. Note that in general $|R_{G_m}(v)| \neq |C| + \sum_{D \in \mathcal{F}_m(C)} |R_{G_m}(w_D)|$ for any node $v \in C$, unless Q_m is a tree. Here, $\mathcal{F}_m(C)$ denotes the set of *child components* of component C in G_m, defined by

$$\mathcal{F}_m(C) = \{D \in \mathcal{C}_m \mid (C, D) \in \mathcal{E}_m\},$$

and w_D stands for a representative node of a component $D \in \mathcal{C}_m$.

In summary, the existing BP method first computes the subset $R_{Q_m}(C)$ of $\mathcal{C}_m$ for each component $C \in \mathcal{C}_m$ by following the edges on the quotient graph Q_m, then calculates $|R_{G_m}(v_C)|$ for only one node $v_C \in C$ by using Eq. (2), and finally sets as follows:

$$|R_{G_m}(v)| \leftarrow |R_{G_m}(v_C)|, \quad (\forall v \in C \setminus \{v_C\}).$$

III. Proposed Method

We enhance the existing BP method by introducing two techniques: redundant-edge pruning (REP) and marginal-component pruning (MCP). Again, we focus on the quotient graph $Q_m = (\mathcal{C}_m, \mathcal{E}_m)$ of the network $G_m = (V, E_m)$ constructed through the m-th bond percolation process.

The REP technique performs pruning redundant edges for reachability analysis among three components in G_m, *i.e.*, three vertices on Q_m. For each component $C \in \mathcal{C}_m$ in G_m, an edge $(C, D) \in \mathcal{E}_m$ is called a *redundant edge* with respect to C if component D is reachable from C via another component $X \in \mathcal{C}_m$. Let $\mathcal{EP}_{Q_m}(C)$ denote the set of all redundant edges with respect to $C \in \mathcal{C}_m$. Then, we have

$$\mathcal{EP}_{Q_m}(C) = \left\{ (C, D) \in \mathcal{E}_m \;\middle|\; D \in \bigcup_{X \in \mathcal{F}_m(C)} \mathcal{F}_m(X) \right\}. \tag{4}$$

Note that if an edge $(C, D) \in \mathcal{E}_m$ is a redundant edge with respect to component C, *i.e.*, $(C, D) \in \mathcal{EP}_{Q_m}(C)$, then it is possible to correctly compute $R_{Q_m}(C)$ without using the edge (C, D). Thus, the REP technique prunes the set of redundant edges $\mathcal{EP}_{Q_m}(C)$ when computing $R_{Q_m}(C)$ for any component $C \in \mathcal{C}_m$. If interpreted as a network motifs [20], the REP technique detects such 3-vertices $\{C, X, D\}$ on graph Q_m that form a feedforward motif pattern $\{(C, X), (X, D), (C, D)\}$, and prunes its short-cut edge (C, D) from them. Let $\mathcal{EP}_{Q_m}$ denote the set of all the redundant edges, *i.e.*,

$$\mathcal{EP}_{Q_m} = \bigcup_{C \in \mathcal{C}_m} \mathcal{EP}_{Q_m}(C).$$

In summary, the REP technique computes the set of all the redundant edges $\mathcal{EP}_{Q_m}$, and replaces the set of edges on Q_m as follows:

$$\mathcal{E}_m \leftarrow \mathcal{E}_m \setminus \mathcal{EP}_{Q_m}.$$

The MCP technique recursively performs pruning components of in-degree 1 or out-degree 1 in the network G_m. Here, we define the sets of components of in-degree 1 and out-degree 1 by Eqs. (5) and (6), respectively:

$$\mathcal{CPI}_{Q_m} = \{C \in \mathcal{C}_m \mid |\mathcal{B}_m(C)| = 1, |\mathcal{F}_m(C)| = 0\}, \tag{5}$$
$$\mathcal{CPO}_{Q_m} = \{C \in \mathcal{C}_m \mid |\mathcal{F}_m(C)| = 1, |\mathcal{B}_m(C)| = 0\}. \tag{6}$$

Here, $\mathcal{B}_m(C)$ denotes the set of all parent components of C,

$$\mathcal{B}_m(C) = \{D \in \mathcal{C}_m \mid (D, C) \in \mathcal{E}_m\}.$$

We define the set $\mathcal{CP}_{Q_m}$ of components of in-degree 1 or out-degree 1 in G_m by $\mathcal{CP}_{Q_m} = \mathcal{CPI}_{Q_m} \cup \mathcal{CPO}_{Q_m}$. Below we explain two basic ideas of the MCP technique. First, for any component $C \in \mathcal{CPI}_{Q_m}$ of in-degree 1, we can easily prove the following properties:

1) $|R_{G_m}(v)| = |C|$ for any $v \in C$.
2) Setting $h_m(D) \leftarrow h_m(D) + |C|$ for the unique parent component $D \in \mathcal{B}_m(C)$, $|R_{G_m}(v_X)|$ is obtained by

$$|R_{G_m}(v_X)| = h_m(X) + \sum_{Y \in R_{Q_m}(X) \setminus \{C\}} h_m(Y)$$

(see Eq. (3)) for any component $X \in \mathcal{C}_m \setminus \{C\}$, where v_X stands for a representative node of X.

Second, for any component $C \in \mathcal{CPO}_{Q_m}$ of out-degree 1, we can easily prove that if $|R_{G_m}(v_D)|$, $(v_D \in D)$ is given for the unique child component $D \in \mathcal{F}_m(C)$, then $|R_{G_m}(v_C)|$, $(v_C \in C)$ is obtained by

$$|R_{G_m}(v_C)| = |C| + |R_{G_m}(v_D)|$$

without computing $R_{Q_m}(C)$ by following the edges on Q_m. Therefore, it is possible to prune the components of in-degree 1 or out-degree 1 in G_m from $\mathcal{C}_m$ when computing $R_{Q_m}(C)$ for any component $C \in \mathcal{C}_m$.

For a component $X \in \mathcal{C}_m$, let $\mathcal{IE}_{Q_m}(X)$ be the set of all edges attached to X in Q_m. We define the operation of pruning a component $C \in \mathcal{C}_m$ in graph Q_m by

$$Q_m \ominus C = (\mathcal{C}_m \setminus \{C\}, \mathcal{E}_m \setminus \mathcal{IE}_{Q_m}(C)).$$

Evidently, after pruning a component C, there might exist some component $D \in \mathcal{C}_m$ such that $D \notin \mathcal{CP}_{Q_m}$ and $D \in \mathcal{CP}_{Q_m \ominus C}$. Thus, the MCP technique need to recursively perform pruning components. In summary, unless $|\mathcal{CP}_{Q_m}| = 0$, the MCP technique recursively selects a component $C \in \mathcal{CP}_{Q_m}$, and prunes C by

$$Q_m \leftarrow Q_m \ominus C$$

after 1) setting

$$|R_{G_m}(v_C)| \leftarrow |C|, \quad (v_C \in C)$$
$$h_m(D) \leftarrow h_m(D) + |C|$$

for the unique parent component $D \in \mathcal{B}_m(C)$ if $C \in \mathcal{CPI}_{Q_m}$, and 2) setting

$$|R_{G_m}(v_C)| \leftarrow |C| + |R_{G_m}(v_D)|$$

when $|R_{G_m}(v_D)|$, $(v_D \in D)$ has been computed for the unique child component $D \in \mathcal{F}_m(C)$ if $C \in \mathcal{CPO}_{Q_m}$.

In our proposed method, the REP technique is applied before the MCP techniques, because it is naturally conceivable that the REP technique increases the number of components of in-degree 1 or out-degree 1. Clearly we can individually incorporate these techniques into the existing BP method. Hereafter, we refer to the proposed method without the MCP technique as the REP method, and the proposed method without the REP technique as the MCP method. Since it is difficult to analytically examine the effectiveness of these techniques, we empirically evaluate the computational efficiency of these three methods in comparison to the existing BP method.

IV. Experiments

Using large real networks, we evaluated the effectiveness of the proposed method.

A. Network Datasets

We employed two large social networks, where all the networks are represented as directed graphs. Here, we adopt the notation for a link in which the link creator is the target node in order to emphasize the direction of information flow.

The first one is a network extracted from "@cosme",[3] a Japanese word-of-mouth communication site for cosmetics, in which each user page can have *fan links*. A fan link (u, v) means that user v registers user u as her favorite user. We traced up to ten steps in the fan-link network from a randomly chosen user in December 2009, and extracted a large weakly-connected network consisting of $45,024$ nodes and $351,299$ directed links. We refer to this directed network as the Cosme network.

The second one is a network extracted from a set of message posts from "Japanese Twitter",[4] which totally consist of $201,297,161$ messages (tweets) made by $1,088,040$ active users (micro-bloggers or twitters who posted no less than 200 messages) during the period of almost three weeks (from March 5, 2011 to March 24, 2011), when the massive earthquake and consequent tsunami in eastern Japan occurred on March 11, 2011. We used the network constructed from the *follower links* between these users, which resulted in a network consisting of $1,088,040$ nodes and $157,371,628$ directed links. We refer to this huge network as the Twitter network.

B. Experimental Settings

One of the simplest models of the SIR framework is the *independent cascade (IC) model* [10], where nodes have two states (*active* and *inactive*), and can switch their states only from inactive to active. The IC model on a network $G = (V, E)$ has a *diffusion probability* $p_{u,v}$ with $0 < p_{u,v} < 1$ for each link $(u, v) \in E$ as a parameter. Suppose that a node $u \in V$ first becomes active at time-step t, it is given a single chance to activate each currently inactive child node $v \in V$ with $(u, v) \in E$, and succeeds with probability $p_{u,v}$. If u succeeds, then v will become active at time-step $t + 1$. If multiple parent nodes of v first become active at time-step t, then their activation trials are sequenced in an arbitrary order, but all performed at time-step t. Whether u succeeds or not, it cannot make any further trials to activate v in subsequent rounds. The process terminates if no more activations are possible. It is well known [10] that the IC model on G for diffusion probabilities $\{p_{u,v} \,|\, (u, v) \in E\}$ is equivalent to the bond percolation process on G for occupation probabilities $\{p_{u,v} \,|\, (u, v) \in E\}$, that is, these two models have the same probability distribution for the final active (recovered) nodes. In the experiments, we employed the IC model.

Now, we explain the setting of diffusion probabilities $\{p_{u,v} \,|\, (u, v) \in E\}$ for the IC model. We draw $\{p_{u,v} \,|\, (u, v) \in E\}$ independently assuming a generative model according to the beta distribution with a mean of μ. Note that the beta distribution is the conjugate prior probability distribution for the Bernoulli distribution corresponding to a single toss of a coin. Then, the average occupied probability of the corresponding bond percolation process over G reduces to μ. Actually, this formulation is equivalent to assigning a uniform value μ to the diffusion probability $p_{u,v}$ for any link, *i.e.*, $p_{u,v} = \mu$, $\forall (u, v) \in E$. In the experiments, we investigated the four cases of very low, low, medium, and high diffusion probabilities:

$$\mu = r/\bar{d}_G, \quad (r = 0.25, 0.5, 1.0, 2.0),$$

where $\bar{d}_G$ is the mean out-degree of network G. We refer r to the *diffusion probability factor*.

For the parameter M of the proposed method, we found $M = 1,000$ to be a reasonable value for estimating the influence degrees for the Cosme and Twitter networks through our preliminary experiments. Thus, we used $M = 1,000$ unless otherwise stated.

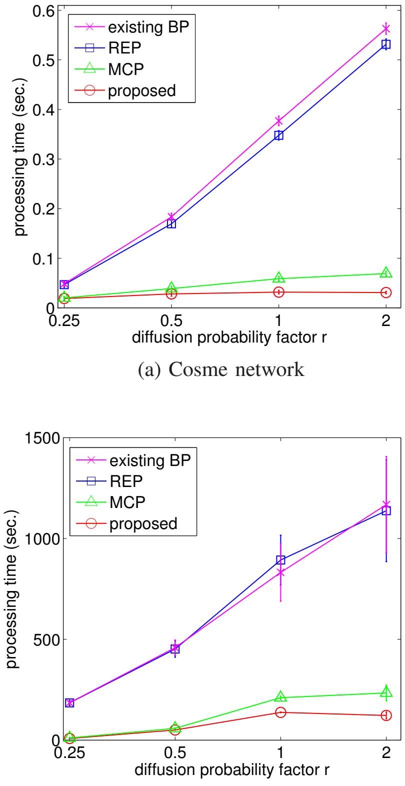
In the next subsection, we explain experimental results for computation time. All our experimentation was undertaken on a single PC with Intel(R) Xeon(R) CPU X5690 @ 3.474 GHz, with 198 GB of memory, running under Linux.

C. Efficiency Evaluation

First, we evaluated the efficiency of the proposed method. We compared the computation time of the proposed, REP, MCP, and existing BP methods. All of them are based on the bond percolation process on the same network G, and have the same accuracy for the same M (see Eq. (1)). Here, we used $M = 100$ trials, and evaluated the time for each trial (corresponding to $M = 1$), because the existing BP method needed much time for the Twitter network. Figure 1 shows the computation time of each method as a function of diffusion probability factor r, where the average values are plotted and the standard deviations are indicated by the error bars. The results show that the MCP technique can always be useful although the REP technique is not necessarily effective alone. However, the proposed method, which incorporates both techniques, always performs the best. The Twitter network requires much longer computation time than the Cosme network since the former is much larger than the latter. It is in particular important to reduce the processing time in case of large diffusion probability μ since the processing time in general increases as μ becomes larger. In case of $r = 2.0$, the proposed method is about 18 times faster than the existing BP

[3] http://www.cosme.net/

[4] http://twitter.jp

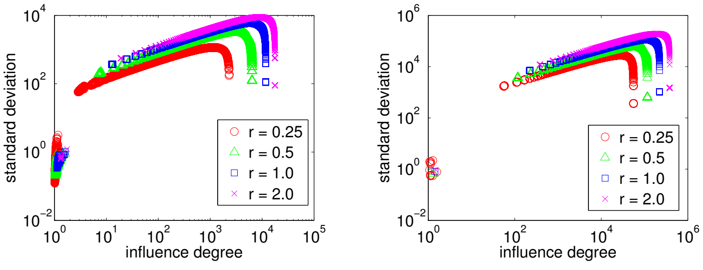


(a) Cosme network

(b) Twitter network

Fig. 1: Computation time comparison.

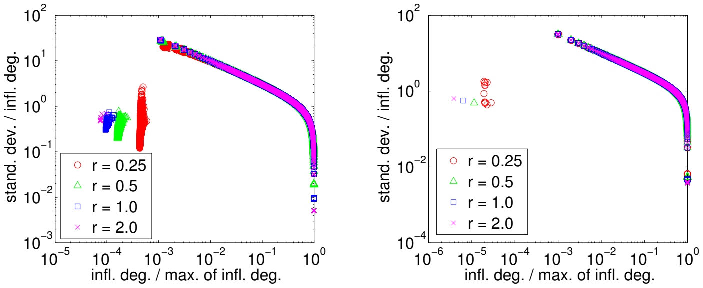
method on average for the Cosme network. Moreover, when using $M = 1$ in the Twitter network for $r = 2.0$, the proposed method requires only about 2 minutes while the existing BP method needs about 20 minutes. Thus, for $M = 1,000$, the existing BP method would have needed about two weeks while the proposed method would have required only about one day and a half. Compared to the existing BP method, the proposed method also has smaller standard deviations, especially for the diffusion probabilities with medium and high values. When the diffusion probability takes a large value, the information diffusion path length changes substantially for each trial as seen in the next experiment (see Fig. 2). This fluctuation is attributed to whether or not information diffusion paths in network G arrive at several marginal components of G, that is, we conjecture that the structure of quotient graph Q_m substantially change for each trial m. In general, it takes more time to trace down longer paths for identifying $R_{Q_m(C)}$ in the BP framework. Since the MCP technique attempts to prune such marginal components in advance, we can expect that the MCP method has smaller standard deviations than the existing BP method. Further, since the REP technique finds candidates of marginal components, we can conjecture that the proposed method combining both the REP and MCP techniques is more stable than the other three methods in terms of computation time. These results demonstrate the effectiveness of the proposed method.



(a) Cosme network

(b) Twitter network

Fig. 2: Results for “influence degree vs. standard deviation”.



(a) Cosme network

(b) Twitter network

Fig. 3: Relation between $\bar{\sigma}^1_G(v)$ and $\bar{s}^1_G(v)$.

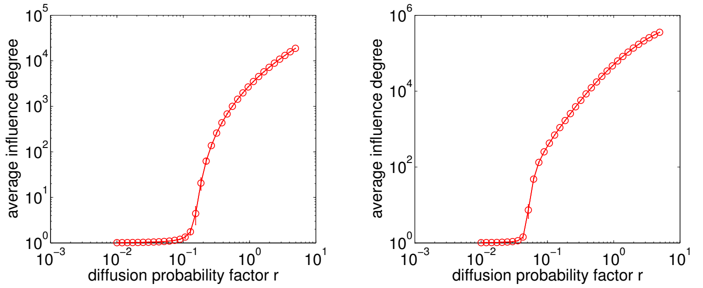
Next, we investigated a global picture of the node influence estimation of the BP method framework with $M = 1,000$ for the Cosme and Twitter networks. Using the proposed method with $M = 1,000$, we estimated the influence degree of each node v in network G by $\bar{\sigma}_G(v)$ (see Eq. (1)), and then calculated the standard deviation $\bar{s}_G(v)$ of samples $\{|R_{G_m}(v)|\}$ for each $v \in V$. Figure 2 plots the pair $(\bar{\sigma}_G(v), \bar{s}_G(v))$ for all $v \in V$. We first see that all the results are qualitatively very similar, and these plots can provide a tool of network structure analysis. In fact, there exists a critical influence degree $\bar{\sigma}_G(v_*)$ for network G such that standard deviation $\bar{s}_G(v)$ is an increasing function of influence degree $\bar{\sigma}_G(v)$ if $\bar{\sigma}_G(v) \leq \bar{\sigma}_G(v_*)$, but $\bar{s}_G(v)$ is a rapidly decreasing function of $\bar{\sigma}_G(v)$ if $\bar{\sigma}_G(v) > \bar{\sigma}_G(v_*)$. Moreover, influence degree $\bar{\sigma}_G(v)$ and its standard deviation $\bar{s}_G(v)$ increase as the diffusion probability becomes larger. We also investigated the relation between ratios $\bar{\sigma}^1_G(v)$ and $\bar{s}^1_G(v)$,

$$\bar{\sigma}^1_G(v) = \bar{\sigma}_G(v)/\max_{u\in V}\bar{\sigma}_G(u), \qquad \bar{s}^1_G(v) = \bar{s}_G(v)/\bar{\sigma}_G(v),$$

for all $v \in V$. Figure 3 plots the pair $(\bar{\sigma}^1_G(v), \bar{s}^1_G(v))$ for all $v \in V$. We observe that $\bar{s}^1_G(v)$ is essentially a decreasing function of $\bar{\sigma}^1_G(v)$, and the function form does not primarily depend on the value of diffusion probability although it does depend on network structure. Moreover, roughly speaking, $\bar{s}^1_G(v)$ becomes almost equal to or less than $10^0 = 1.0$ when the ratio $\bar{\sigma}^1_G(v)$ is larger than 10^{-1} for both the networks, which means that standard deviation $\bar{s}_G(v)$ becomes almost equal to or less than $\bar{\sigma}_G(v)$ for nodes whose influence degree $\bar{\sigma}_G(v)$ is greater than 10% of the maximum value of influence degree. These results imply that the estimation accuracy with $M = 1,000$ is acceptable from a statistical point of view.

D. Average Influence Degree

We consider finding the epidemic threshold p^*_G of the IC model for the Cosme and Twitter networks. To this end, we ex-



(a) Cosme network

(b) Twitter network

Fig. 4: Average influence degree curves.

TABLE I: Ranking results for conventional centralities in the huge Twitter network.

Rank	Degree	Betweenness	Closeness
1	**masason**	**shuzo_matsuoka**	**masason**
2	**GachapinBlog**	SNOOPYbot	**GachapinBlog**
3	higashimototiji	NHK_PR	**shuzo_matsuoka**
4	**shuzo_matsuoka**	moomin_valley	higashimototiji
5	555hamako	shuumai	takapon_jp

Rank	Hub	PageRank
1	tomo7272	**masason**
2	ktamiya	natalie_mu
3	euro_tour	JAXA_jp
4	rakko001	Hayabusa_jaxa
5	mabou77	**GachapinBlog**

TABLE II: Ranking results for the influence degree centrality in the huge Twitter network.

Rank	$r = 0.25$	$r = 0.5$	$r = 1.0$	$r = 2.0$
1	**masason**	**masason**	**masason**	**masason**
2	**GachapinBlog**	**GachapinBlog**	**GachapinBlog**	**GachapinBlog**
3	higashimototiji	itoi_shigesato	itoishigesato	**utadahikaru**
4	itoi_shigesato	higashimototiji	higashimototiji	shiro_tsubuyaki
5	555hamako	Astro_Soichi	**utadahikaru**	tenkijp

amined the relation between the diffusion probability $p_{u,v} = \mu$ and the average influence degree $\sum_{v \in V} \sigma_G(v)/|V|$. Since this is a computationally heavy task, we estimated the average influence degree using the proposed method with $M = 100$. Figure 4 shows the estimated average influence degree as a function of diffusion probability factor r, where the standard deviations (see Eq. (1)) are indicated by the error bars. Here, we investigated $r = r_1 a^{k-1}$, ($r_1 = 0.01$, $a = 1.2$, $k = 1, \ldots, 35$), that is, $1.3 \times 10^{-3} \leq \mu \leq 6.3 \times 10^{-1}$ for the Cosme network and $6.9 \times 10^{-5} \leq \mu \leq 3.4 \times 10^{-2}$ for the Twitter network. We first observe that the standard deviations are relatively small, and the accuracy with $M = 100$ is acceptable when the goal is to estimate the average influence degree. We needed about 1.1 minutes for the Cosme network and about 9.1 hours for the Twitter network to obtain the results shown in Figure 4. From Figure 4, we can find that the epidemic threshold $p^*_G = r^*_G/\bar{d}_G$ is given by $p^*_G = 1.9 \times 10^{-2}$ ($r^*_G = 0.15$) for the Cosme network and $p^*_G = 2.8 \times 10^{-4}$ ($r^*_G = 0.04$) for the Twitter network. These results imply that the epidemic threshold depends on network structure and the Twitter network spreads information more easily than the Cosme network.

E. Comparison with Conventional Centralities

Although estimating influence degree centrality for large networks is a time-consuming and difficult task, the proposed method enabled us to approximately calculate the influence degree within a reasonable time even for huge social networks. Thus, for the huge Twitter network, we evaluated whether or not the influence degree centrality can actually provide a novel concept in comparison with conventional centralities.

As conventional centralities, we examined the betweenness centrality, the closeness centrality, the hub centrality, and the PageRank centrality for network G. Here, the betweenness $betw(v)$ of a node v is defined as $betw(v) = \sum_{u \in V} \sum_{w \in V} spath^G_{u,w}(v)/spath^G_{u,w}$, where $spath^G_{u,w}$ is the total number of the shortest paths between node u and node v in G and $spath^G_{u,w}(v)$ is the number of the shortest paths between node u and node v in G that passes through node v. The closeness $close(v)$ of a node v is defined as $close(v) = (1/|V|) \sum_{u \in V} (1/dist_G(v, u))$, where $dist_G(v, u)$ stands for the graph distance from v to u in G. Also, the hub centrality score of a node is obtained by the HITS algorithm [21] that defines the hub and authority centrality, and the PageRank score of a node is provided by applying the PageRank algorithm with random jump factor 0.15 [8] to the reverse network $G^- = (V, E^-)$ that is constructed through reversing any link of G, that is, $E^- = \{(u, v) \in V \times V |\ (v, u) \in E\}$.

Tables I and II show the top five nodes in the degree, betweenness, closeness, hub, PageRank, and influence degree ($r = 0.25, 0.5, 1.0, 2.0$) centralities for the Twitter network. We can first observe that each centrality measure actually extracts its own proper nodes. For the influence degree centrality, while the diffusion probability setting affects the result, the top two nodes coincided. They were "masason" and "GachapinBlog", which also appeared in the top five of the degree, closeness and PageRank centralities. Here, "masason" is the Twitter account of Masayoshi Son who is a famous Japanese businessman and CEO of SoftBank (a big IT company), and "GachapinBlog" is the Twitter account of Gachapin who is a popular Japanese TV character in a children's program. These are very influential in Japanese Twitter. Unlike other centralities, the hub centrality extracted the representatives of a certain big community in Japanese Twitter, where "tomo7272" is the Twitter account of an ordinary person who often posts nice tweets. Note that "shuzo_matsuoka" is a famous bot in Japanese Twitter, and was extracted by the degree, betweenness and closeness centralities. However, it did not appear in the top ten of the influence degree ranking. The tweet of bot attracts many people but dies out very rapidly. Thus, it is not identified as influential by the proposed method. On the other hand, "utadahikaru" was extracted only by the influence degree centrality with medium and high diffusion probabilities while it did not appear in the top ten of other rankings. Here, "utadahikaru" is the Twitter account of Hikaru Utada who is a Japanese American singer known as one of the most influential artists in Japan. These results demonstrate that the influence degree centrality can serve as a novel measure that extracts influential nodes in terms of information diffusion which are not identified by existing measures.

V. Conclusion

We view the dynamic process of information diffusion as an important ingredient to evaluate the importance of a node in a social network, and consider that the node influence degree shares the same role that other existing topology-based centrality measured have. Unlike the existing centrality measures, the influence degree centrality is not easily computable because it is defined to be the expected number of information spread. We proposed a method that can estimate the influence degree of every single node in a large network simultaneously under the framework of *SIR model* setting. More specifically, we proposed two new pruning techniques on top of the existing bond percolation approach, in which the problem is reduced to counting the reachable nodes from each single node in the directed graph which is generated by bond percolation. We tested our algorithm using two real world networks, one with 40K nodes and the other with $1,000$K nodes. The experimental results confirmed that the new pruning techniques improve the computational efficiency by an order of magnitude over the existing bond percolation method which is already three orders of magnitude faster than direct Monte Carlo simulations.

We also demonstrated that the proposed method can estimate the epidemic threshold of the IC model even for a huge Twitter network with $1,000$K nodes in reasonable time by examining the relation between the diffusion probability and the average influence degree, and showed that the epidemic threshold depends on network structure and the Twitter network spreads information more easily than the Cosme network. Further, it is confirmed that the nodes identified as influential by the influence degree centrality based on the SIR model are not necessarily the same or similar to those identified by the other existing centralities, and the influence degree centrality can identify those nodes that are deemed indeed influential but are not identifiable by the existing methods. The bond percolation is a generic approach for the SIR model and can be instantiated to any specific diffusion model. It does not require any approximations or assumptions to the model to improve the computational efficiency, *e.g.*, small diffusion probability, shortest path, maximum influence path, etc., that were needed in the existing approaches. We instantiated it to the independent cascade (IC) model, but the same technique can be applied to other instantiations, *e.g.*, linear threshold (LT) model.

Our immediate future work is to extensively evaluate the proposed method for various instantiations of the SIR framework including the LT model by using large real networks in a variety of fields. Needless to say, it is also necessary to mathematically clarify the performance difference between the proposed method and the existing BP method in terms of computational efficiency. In several real-world networks, there exist phenomena in which the *SIS model* is more suitable than the SIR model [3], [22], where every node is allowed to be activated multiple times. It is known that the SIS-type independent cascade model on a network can be exactly mapped onto the IC model on a layered network built from the original network [10], [23]. Thus, note that the proposed method developed for the SIR setting can also be applied to the SIS setting. Our future work includes evaluating the proposed method in the SIS framework.

Acknowledgments

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

References

[1] M. Kimura, K. Saito, and R. Nakano, “Extracting influential nodes for information diffusion on a social network,” in *Proceedings of AAAI’07*, 2007, pp. 1371–1376.

[2] M. Kimura, K. Saito, R. Nakano, and H. Motoda, “Extracting influential nodes on a social network for information diffusion,” *Data Mining and Knowledge Discovery*, vol. 20, pp. 70–97, 2010.

[3] M. Newman, “The structure and function of complex networks,” *SIAM Review*, vol. 45, pp. 167–256, 2003.

[4] D. Watts, “A simple model of global cascades on random networks,” *Proceedings of National Academy of Science, USA*, vol. 99, pp. 5766–5771, 2002.

[5] L. Freeman, “Centrality in social networks: Conceptual clarification,” *Social Networks*, vol. 1, pp. 215–239, 1979.

[6] P. Bonacich, “Power and centrality: A family of measures,” *American Journal of Sociology*, vol. 92, pp. 1170–1182, 1987.

[7] L. Katz, “A new status index derived from sociometric analysis,” *Psychometrika*, vol. 18, pp. 39–43, 1953.

[8] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” *Computer Networks and ISDN Systems*, vol. 30, pp. 107–117, 1998.

[9] H. Zhuge and J. Zhang, “Topological centrality and its e-science applications,” *Journal of the American Society of Information Science and Technology*, vol. 61, pp. 1824–1841, 2010.

[10] D. Kempe, J. Kleinberg, and E. Tardos, “Maximizing the spread of influence through a social network,” in *Proceedings of KDD’03*, 2003, pp. 137–146.

[11] W. Chen, L. Lakshmanan, and C. Castillo, “Information and influence propagation in social networks,” *Synthesis Lectures on Data Management*, vol. 5(4), pp. 1–177, 2013.

[12] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, “Cost-effective outbreak detection in networks,” in *Proceedings of KDD’07*, 2007, pp. 420–429.

[13] W. Chen, Y. Wang, and S. Yang, “Efficient influence maximization in social networks,” in *Proceedings of KDD’09*, 2009, pp. 199–208.

[14] W. Chen, C. Wang, and Y. Wang, “Scalable influence maximization for prevalent viral marketing in large-scale social networks,” in *Proceedings of KDD’10*, 2010, pp. 1029–1038.

[15] W. Chen, Y. Yuan, and L. Zhang, “Scalable influence maximization in social networks under the linear threshold model,” in *Proceedings of ICDM’10*, 2010, pp. 88–97.

[16] A. Goyal, F. Bonchi, and L. Lakshmanan, “A data-based approach to social influence maximization,” *Proceedings of the VLDB Endowment*, vol. 5(1), pp. 73–84, 2011.

[17] H. Nguyen and R. Zheng, “Influence spread in large-scale social networks - a belief propagation approach,” in *Proceedings of ECML-PKDD’12*. LNAI 7524, 2012, pp. 515–530.

[18] Y. Yang, E. Chen, Q. Liu, B. Xiang, T. Xu, and S. Shad, “On approximation of real-world influence spread,” in *Proceedings of ECML-PKDD’12*. LNAI 7524, 2012, pp. 548–564.

[19] K. Saito, M. Kimura, K. Ohara, and H. Motoda, “Identifying supermediators of information diffusion in social networks,” in *Proceedings of DS’13*. LNAI 8140, 2013, pp. 170–184.

[20] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, “Network motifs: Simple building blocks of complex networks,” *Science*, vol. 298, pp. 824–827, 2002.

[21] S. Chakrabarti, B. Dom, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg, “Mining the web’s link structure,” *IEEE Computer*, vol. 32, pp. 60–67, 1999.

[22] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, “Patterns of cascading behavior in large blog graphs,” in *Proceedings of SDM’07*, 2007, pp. 551–556.

[23] M. Kimura, K. Saito, and H. Motoda, “Efficient estimation of influence functions fot sis model on social networks,” in *Proceedings of IJCAI’09*, 2009, pp. 2046–2051.