Finding Relation between PageRank and Voter Model

Takayasu Fushimi[1], Kazumi Saito[1], Masahiro Kimura[2], Hiroshi Motoda[3], and Kouzou Ohara[4]

[1] Graduate School of Administration and Informatics, University of Shizuoka
52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan
{j09118,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
Osaka 567-0047, Japan
motoda@ar.sanken.osaka-u.ac.jp
[4] Department of Integrated Information Technology, Aoyama Gakuin University
Kanagawa 229-8558, Japan
ohara@it.aoyama.ac.jp

Abstract. Estimating influence of a node is an important problem in social network analyses. We address this problem in a particular class of model for opinion propagation in which a node adopts its opinion based on not only its direct neighbors but also the average opinion share over the whole network, which we call an extended Voter Model with uniform adoption (VM). We found a similarity of this model with the well known PageRank (PR) and explored the relationships between the two. Since the uniform adoption implies the random opinion adoption of all nodes in the network, it corresponds to the random surfer jump of PR. For an undirected network, both VM and PR give the same ranking score vector because the adjacency matrix is symmetric, but for a directed network, the score vector is different for both because the adjacency matrix is asymmetric. We investigated the effect of the uniform adoption probability on ranking and how the ranking correlation between VM and PR changes using four real world social networks. The results indicate that there is little correlation between VM and PR when the uniform adoption probability is small but the correlation becomes larger when both the uniform adoption and the random surfer jump probabilities become larger. We identified that the recommended value for the uniform adoption probability is to be around 0.25 to obtain a stable solution.

1 Introduction

Recent technological innovation in the web such as blogosphere and knowledge/media-sharing sites is remarkable, which makes it possible to form various kinds of large social networks, through which behaviors, ideas and opinions can spread. Thus, substantial attention has been directed to investigating the spread of influence in these networks [12, 4, 17]. The representative problem is the influence maximization problem, that is, the problem of finding a limited number of influential nodes that are effective for the

spread of information through the network and new algorithmic approaches have been proposed under different model assumptions, e.g. descriptive probabilistic interaction models [5, 15], and basic diffusion models such as independent cascade (IC) model and the linear threshold (LT) model [8, 9, 19]. This problem has good applications in sociology and "viral marketing" [1].

Another line of work on the spread of influence is opinion share analyses, i.e. how people changes their opinions, how each opinion propagates and what the final opinion share is, etc. A good model for opinion diffusion would be a voter model [13, 16]. It is one of the most basic stochastic process model, and has the same key property with the linear threshold model that a node decision is influenced by its neighbors' decision, i.e. a person changes its opinion by the opinions of its neighbors. The basic voter model is defined on an undirected network with self-loop and each node initially holds one of K opinions, and adopts the opinion of a randomly chosen neighbor at each subsequent discrete time-step.

Even-Dar and Shapira [6] investigated the influence maximization problem (maximizing the spread of the opinion that supports a new technology) under the basic voter model with two ($K = 2$) opinions (one in favor of the new technology and the other against it) at a given target time T. They showed that the most natural heuristic solution, which picks the nodes in the network with the highest degree, is indeed the optimal solution, under the condition that all nodes have the same cost.

We propose a new model for the spread of opinions. Each person has a different influence on the other person and the person to person relation is directional. A person not only changes its opinion by its direct neighbors but also considers the overall opinion distributions of the whole society. The new model incorporates these factors and we call this model as an extended Voter Model with uniform adoption. Here we note that the new model has a strong similarity to the well known PageRank [2, 11] which is an algorithm to rank Web pages. Since the uniform adoption can be viewed as random opinion adoption of all nodes in the network, it is equivalent to the random surfer jump of PageRank.

We mathematically derive the ranking vector of the new Voter Model and compare it with that of PageRank, and explore how the two models are related by a series of extensive experiments using four real world social networks. Especially we investigate the effects of the uniform adoption probability on node ranking and how the ranking of the new Voter Model and PageRank are correlated to each other with this probability. The ranking of the new Voter Model becomes the same as that of PageRank if we assume that the network is undirectional, but since both our new model and PageRank use directional network, the ranking results are not the same. The results indicate that the correlation varies with the uniform adoption probability. There is little correlation between the extended Voter Model and PageRank when the uniform adoption probability is small and the high ranked nodes are different, but the correlation becomes larger when both the uniform adoption and the random surfer jump probabilities become larger. We found that the ranking becomes stable for the uniform adoption probability in the range of 0.15 and 0.35 and the self correlation within the extended Voter Model is high in this region, which is consistent with the report that the recommended value for the random surfer jump is 0.15.

The paper is organized as follows. We briefly explain the standard Voter Model and revisit PageRank in sections 2 and 3, respectively. Then we explain our new Voter Model, the extended Voter Model with uniform adoption, in section 4. Experimental results that describe various correlation results are detailed in section 5. Finally we summarize our conclusion in section 7.

2 Voter Model

In this section, according to the work [6], we first consider the diffusion of opinions in a social network represented by an undirected (bidirectional) graph $G = (V, E)$ with self-loops. Here, V and E ($\subset V \times V$) are the sets of all the nodes and links in the network, respectively. For a node $v \in V$, let $\Gamma(v)$ denote the set of neighbors of v in G, that is, $\Gamma(v) = \{u \in V; (u, v) \in E\}$. Note that $v \in \Gamma(v)$.

According to the work [6], we recall the definition of the basic voter model with two opinions on network G. In the voter model, each node of G is endowed with two states; opinions 1 and 2. The opinions are initially assigned to all the nodes in G, and the evolution process unfolds in discrete time-steps $t = 1, 2, 3, \cdots$ as follows: At each time-step t, each node v picks a random neighbor u and adopts the opinion that u holds at time-step $t - 1$.

More formally, let $f_t : V \to \{1, 2\}$ denote the opinion distribution at time-step t, where $f_t(v)$ stands for the opinion of node v at time-step t. Then, $f_0 : V \to \{1, 2\}$ is the initial opinion distribution, and $f_t : V \to \{1, 2\}$ is inductively defined as follows: For any $v \in V$,

$$\begin{cases} f_t(v) = 1, \text{ with probability } \frac{n_{t-1}(1,v)}{n_{t-1}(1,v) + n_{t-1}(2,v)}, \\ f_t(v) = 2, \text{ with probability } \frac{n_{t-1}(2,v)}{n_{t-1}(1,v) + n_{t-1}(2,v)}, \end{cases}$$

where $n_t(k, v)$ is the number of v's neighbors that hold opinion k at time-step t for $k = 1, 2$.

3 PageRank Revisited

We revisit PageRank [2, 11]. For a given Web network (directed graph), we identify each node with a unique integer from 1 to $|V|$. Then we can define the adjacency matrix $\boldsymbol{A} \in \{0, 1\}^{|V|\times|V|}$ by setting $a(u, v) = 1$ if $(u, v) \in E$; otherwise $a(u, v) = 0$. A node can be self-looped, in which case $a(u, u) = 1$. For each node $v \in V$, let $F(v)$ and $B(v)$ denote the set of child nodes of v and the set of parent nodes of v, respectively, $F(v) = \{w \in V;\ (v, w) \in E\}$, $B(v) = \{u \in V;\ (u, v) \in E\}$. Note that $v \in F(v)$ and $v \in B(v)$ for a node v with a self-loop.

Then we can consider the row-stochastic transition matrix $\boldsymbol{P}$, each element of which is defined by $p(u, v) = a(u, v)/|F(u)|$ if $|F(u)| > 0$; otherwise $p(u, v) = z(v)$, where $\boldsymbol{z}$ is some probability distribution over pages, i.e., $z(v) \geq 0$ and $\sum_{v\in V} z(v) = 1$. This model means that from dangling Web pages without out-links ($F(u) = \emptyset$), a random surfer jumps to page v with probability $z(v)$. The vector $\boldsymbol{z}$ is referred to as a personalized vector because we can define $\boldsymbol{z}$ according to user's preference.

Let $\boldsymbol{y}$ denote a vector representing PageRank scores over pages, where $y(v) \geq 0$ and $\sum_{v \in V} y(v) = 1$. Then using an iteration-step parameter t, PageRank vector $\boldsymbol{y}$ is defined as a limiting solution of the following iterative process,

$$\boldsymbol{y}_t^T = \boldsymbol{y}_{t-1}^T \left((1-\beta)\boldsymbol{P} + \beta \boldsymbol{e}\boldsymbol{z}^T \right) = (1-\beta)\boldsymbol{y}_{t-1}^T \boldsymbol{P} + \beta \boldsymbol{z}^T, \tag{1}$$

where $\boldsymbol{a}^T$ stands for a transposed vector of $\boldsymbol{a}$ and $\boldsymbol{e} = (1, \cdots, 1)^T$. In the Equation (1), β is referred to as the uniform jump probability. This model means that with the probability β, a random surfer also jumps to some page according to the probability distribution $\boldsymbol{z}$. The matrix $((1-\beta)\boldsymbol{P} + \beta \boldsymbol{e}\boldsymbol{z}^T)$ is referred to as a Google matrix. The standard PageRank method calculates its solution by directly iterating Equation (1), after initializing $\boldsymbol{y}_0$ adequately. One measure to evaluate its convergence is defined by

$$\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|_{L1} \equiv \sum_{v \in V} |y_t(v) - y_{t-1}(v)|. \tag{2}$$

Note that any initial vector $\boldsymbol{y}_0$ can give almost the same PageRank scores if it makes Equation (2) almost zero because the unique solution of Equation (1) is guaranteed.

4 Voter Model with uniform adoption

We propose an extended Voter Model with uniform adoption on a directed graph $G = (V, E)$ with self-loops for K opinions. Let $m_t(k, v)$ be the number of v's parents that hold opinion k at time-step t for $k = 1, 2, ..., K$. In addition, just like the personalized vector employed in PageRank, we introduce some probability distribution $\boldsymbol{z}$ over nodes. Let $m_t(k)$ be the weighted share of opinion k at time-step t given by

$$m_t(k) = \sum_{\{v \in V;\, f_t(v) = k\}} z(v), \tag{3}$$

then $f_t : V \to \{1, 2, ..., K\}$ is inductively defined as follows, given an initial opinion distribution $f_0 : V \to \{1, 2, ..., K\}$. For any $v \in V$,

$$f_t(v) = k, \quad \text{with probability } (1-\alpha)\frac{m_{t-1}(k, v)}{\sum_{k=1}^{k=K} m_{t-1}(k, v)} + \alpha m_{t-1}(k). \tag{4}$$

This model indicates that the opinion of each node $v \in V$ is influenced by its parents nodes $B(v)$ with probability $(1-\alpha)$ and by any other node $u \in V$ with probability α according to $\boldsymbol{z}$. Hereafter, α is referred to as the uniform adoption probability and the extended Voter Model with uniform adoption is referred to as VM for short[2].

Now we consider estimating the expected influence degree of node $u \in V$, which is defined as the expected number of nodes influenced by u's initial opinion $f_0(u)$. Note that the following definition does not depend on which opinion u holds initially. We denote the expected influence degree of node u at time-step t by $x_t(u)$. Let $\boldsymbol{h}_u \in \{0, 1\}^{|V|}$ be a vector whose u-th element is 1 and other elements are 0, and $\boldsymbol{Q}$ the column-stochastic

[2] We call it as the extended VM when we have to make distinction from the standard VM.

transition matrix, each element of which is defined by $q(u,v) = a(u,v)/|B(v)|$. Here note that $B(v) \neq \emptyset$ for any node $v \in V$ because of the existence of self-loop. From the definition of our model, we can calculate $x_1(u)$ as follows.

$$x_1(u) = (1-\alpha) \sum_{v \in F(u)} |B(v)|^{-1} + |V|\alpha z(u) = \boldsymbol{h}_u^T \left((1-\alpha)\boldsymbol{Q} + \alpha \boldsymbol{z}\boldsymbol{e}^T\right)\boldsymbol{e}. \tag{5}$$

Each element of the vector $\boldsymbol{h}_u^T((1-\alpha)\boldsymbol{Q}+\alpha\boldsymbol{z}\boldsymbol{e}^T)$ is the probability that the corresponding node v is influenced by the node u with one time-step. Thus from the independence property of the opinion diffusion process, we can calculate $x_t(u)$ as follows.

$$x_t(u) = \boldsymbol{h}_u^T \left((1-\alpha)\boldsymbol{Q} + \alpha \boldsymbol{z}\boldsymbol{e}^T\right)^t \boldsymbol{e}. \tag{6}$$

Here since the vector $\boldsymbol{h}_u$ works for selecting the u-th element, we can obtain the vector consisting of the expected influence degree at time-step t as follows:

$$\boldsymbol{x}_t = \left((1-\alpha)\boldsymbol{Q} + \alpha \boldsymbol{z}\boldsymbol{e}^T\right)^t \boldsymbol{e} = \left((1-\alpha)\boldsymbol{Q} + \alpha \boldsymbol{z}\boldsymbol{e}^T\right)^{t-1} \boldsymbol{x}_{t-1} \tag{7}$$

Moreover, since $((1-\alpha)\boldsymbol{Q}+\alpha\boldsymbol{z}\boldsymbol{e}^T)$ becomes the column-stochastic transition matrix, we can consider a stationary vector defined by $\boldsymbol{x} = \lim_{t\to\infty} \boldsymbol{x}_t$.

For the sake of technical convenience, we perform scaling to the vector $\boldsymbol{x}$ defined by $\boldsymbol{x} \leftarrow \boldsymbol{x}/|V|$. Then, similarly to PageRank calculation process defined in Equation (1), we can obtain the expected influence vector at time-step t as follows after initializing vector to $\boldsymbol{x}_0 = \boldsymbol{e}/|V|$:

$$\boldsymbol{x}_t = \left((1-\alpha)\boldsymbol{Q} + \alpha \boldsymbol{z}\boldsymbol{e}^T\right)\boldsymbol{x}_{t-1} = (1-\alpha)\boldsymbol{Q}\boldsymbol{x}_{t-1} + \alpha\boldsymbol{z}. \tag{8}$$

We can employ the same convergence measure defined by Equation (2), just by replacing the vector $\boldsymbol{y}$ with $\boldsymbol{x}$. Here, we note that in case of undirected networks with self-loops Equations (1) and (8) become completely equivalent since there exist no dangling nodes. Note also that in this case, our extended VM reduces to the standard VM by setting $\alpha = 0$. On the other hand, in case of directed networks with self-loops, Equations (1) and (8) give different vector sequences, and we empirically evaluate their differences with special emphasis on their stationary vectors.

5 Experiments

In this section, we evaluate the effects of 1) the uniform adoption probability α and 2) community structure, and examine the relation between VM and PR by extensive experiments using four real networks.

5.1 Experimental settings

In our experiments, we employ the Pearson correlation coefficients as our basic evaluation measure. For the sake of convenience, we recall its definition: given two vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, the correlation coefficient $C(\boldsymbol{x},\boldsymbol{y})$ is defined as follows.

$$C(\boldsymbol{x},\boldsymbol{y}) = \frac{(\boldsymbol{x}-\bar{x}\boldsymbol{e})^T(\boldsymbol{y}-\bar{y}\boldsymbol{e})}{\sqrt{(\boldsymbol{x}-\bar{x}\boldsymbol{e})^T(\boldsymbol{x}-\bar{x}\boldsymbol{e})}\sqrt{(\boldsymbol{y}-\bar{y}\boldsymbol{e})^T(\boldsymbol{y}-\bar{y}\boldsymbol{e})}}, \tag{9}$$

where $\bar{x}$ and $\bar{y}$ stand for the average element values of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and recall that $\boldsymbol{e}$ is a vector defined by $\boldsymbol{e} = (1, \cdots, 1)^T$.

As mentioned earlier, we focus on evaluating the vectors of the expected influence degree, each of which is the stationary vector defined as a limiting solution of Equation (8) in VM. In our experiments, the personalized vector $\boldsymbol{z}$ is set to uniform one, i.e., $\boldsymbol{z} = (1/|V|, \cdots, 1/|V|)^T$. Based on Equation (2), the convergence criterion to obtain the stationary vectors is set to $||\boldsymbol{x}_t - \boldsymbol{x}_{t-1}||_{L1} < 10^{-12}$ in case of VM, and $||\boldsymbol{y}_t - \boldsymbol{y}_{t-1}||_{L1} < 10^{-12}$ in case of PR.

Our evaluation consists of two series of experiments. In the first series of experiments, we evaluate the effects of the uniform adoption probability on the expected influence degree. In the second series of experiments, we evaluate the effects of network's community structure on the expected influence degree. Now, we explain our method of rewiring the originally observed network to change its community structure. The rewired network is constructed just by randomly rewiring links of the original network according to some probability p without changing the degree of each node [14]. More specifically, by arbitrarily ordering all links except for self-loops 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., $e = (u, v)$. From the list L_E, we can produce two node lists, i.e., the *from*-part node list L_F and the *to*-part node list L_T. Thus, by swapping two elements of the node list L_T with the probability p so as not to produce multiple-links, we can obtain a partially reordered node list L'_T. Then, by concatenating L'_T with the other node list L_F, we can produce a link list for a rewired network. Namely, 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 rewired network is $L'_E = ((L_F(1), L'_T(1)), \cdots, (L_F(|E|), L'_T(|E|)))$.

5.2 Network Data

In our experiments, we employed four sets of real networks, which exhibit many of the key features of social networks. Below we describe the details of these network data.

The first one is a reader network of "Ameba"[3] that is a Japanese blog service site. Blogs are personal on-line diaries managed by easy-to-use software packages, and have rapidly spread through the World Wide Web [7]. Each blog of "Ameba" can have the *reader list* that consists of the hyperlinks to the blogs of the reader bloggers. Here, a reader link from blog X to blog Y is generated when blog Y registers blog X as her favorite blog. Thus, a reader network can be regarded as a social network. We crawled the reader lists of $117,374$ blogs of the Ameba blog service site in June 2006, and collected a large connected network. This network had $56,604$ nodes and $1,071,080$ directed links. We refer to this network as the Ameblo network.

Second one is a trackback network of blogs used in [9]. Bloggers 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" [4], where a blogger can recruit trackbacks of other bloggers

[3] http://www.ameba.jp/

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

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. We refer to this network data as the Blog network.

The third one is a fan network of "@cosme"[5] that is a Japanese word-of-mouth communication site for cosmetics. Each user page of "@cosme" can have *fan links*. Here, a fan link from user X to user Y is generated when user Y registers user X as her favorite user. Thus, a fan network can be regarded as a social network. We traced up to ten steps in the fan links from a randomly chosen user in December 2009, and collected a large connected network [6]. This network had $45,024$ nodes and $546,930$ directed links. We refer to this network as the Cosme network.

Last we employed a network derived from the Enron Email Dataset [10]. We first extracted the email addresses that appeared in the Enron Email Dataset as senders and recipients. We regarded each email address as a node, and constructed a directed network obtained by linking two email addresses u and v if u sent an email to v Next, we extracted its maximal strongly connected component. We refer to this strongly connected bidirectional network as the Enron network. This network had $4,254$ nodes and $44,314$ directed links. We refer to this dataset as the Enron network dataset.

Table 1: Basic statistics of networks.

network	$\|V\|$	$\|E\|$	$C(\boldsymbol{B}, \boldsymbol{F})$
Ameblo	56,604	1,071,080	0.61350
Blog	12,047	79,920	0.74377
Cosme	45,024	546,930	0.51940
Enron	19,654	377,612	0.54929

Table 1 shows the basic statistics of the Ameblo, Blog, Cosme and Enron networks. Here, $C(\boldsymbol{B}, \boldsymbol{F})$ denotes the Peason correlation coefficients between the in-degree vector $\boldsymbol{B}$, each element of which is $|B(v)|$, and the out-degree vector $\boldsymbol{F}$, each element of which is $|F(v)|$. From this table, we consider that each network has an intrinsic characteristics as a directed network because $C(\boldsymbol{B}, \boldsymbol{F})$ is reasonably smaller than 1.

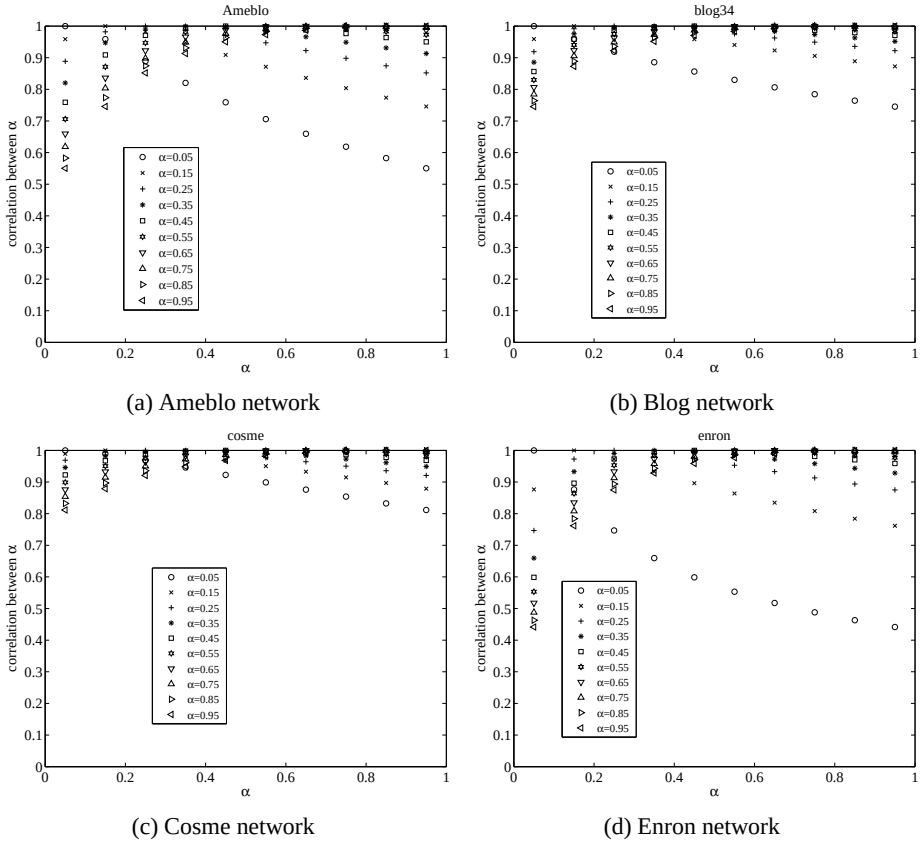
5.3 Effects of uniform adoption probability

As the first series of experiments, we evaluated the effects of the uniform adoption probability change on the expected influence degree. Here, let $\boldsymbol{x}(\alpha)$ be the stationary vector defined as a limmiting solution of Equation (8) for VM with α. In order to evaluate the effects of different uniform adoption probabilities, we calculated the correlation coefficients $C(\boldsymbol{x}(\alpha), \boldsymbol{x}(\alpha'))$ with respect to each pair of the uniform adoption probabilities, α

[5] `http://www.cosme.net/`

[6] We further tried this collection procedure twice, and compared the resulting networks. Then, we found that they overlapped 99.5%.

and α' (self correlation). In Fig.1, we plot $C(\boldsymbol{x}(\alpha), \boldsymbol{x}(\alpha'))$ with respect to α, where each result with different α' is shown by a different marker. Here we changed both the values of α and α' from 0.05 to 0.95 with an increment of 0.1.



(a) Ameblo network (b) Blog network

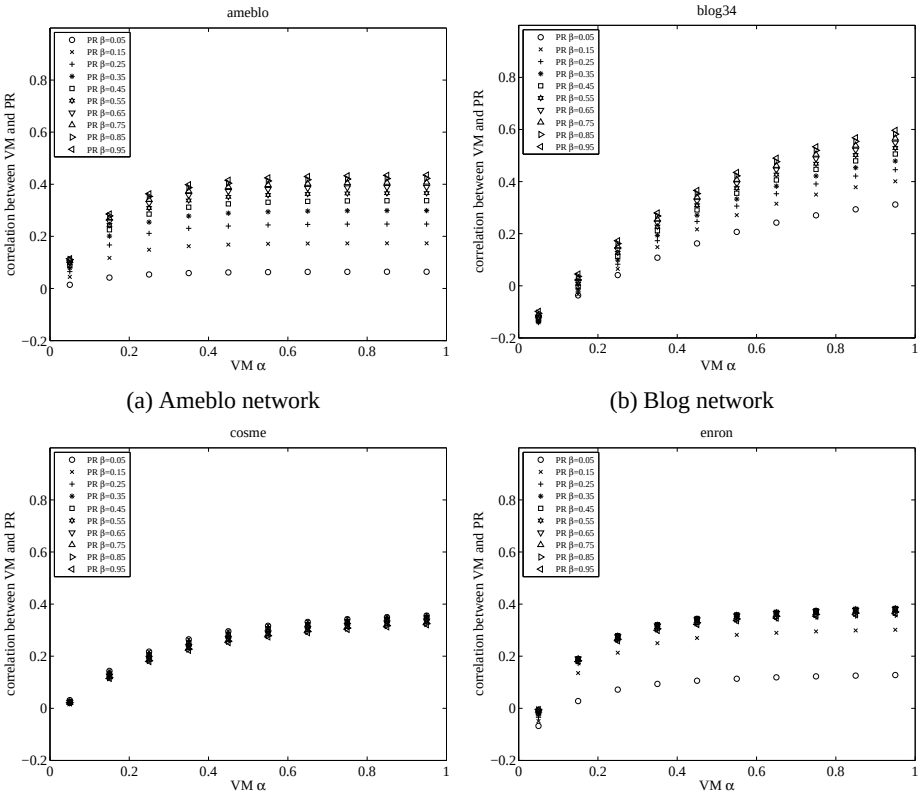
(c) Cosme network (d) Enron network

Fig. 1: The correlation coefficient between VMs with different α

From Fig.1, we can observe the following smilar characteristics of VM for all of the four networks. First, the correlation coefficients $C(\boldsymbol{x}(\alpha), \boldsymbol{x}(\alpha'))$ for any pair of α and α' are relatively high. Second, $C(\boldsymbol{x}(\alpha), \boldsymbol{x}(\alpha'))$ in the range of $0.15 \leq \alpha \leq 0.35$ shows especially high values regardless of α. This suggests that we can recommend to employ this range of α because this would give a stable (and thus, representative) value of the expected influence degree for VM. Incidentally, it is reported that the uniform jump probability β in PR is frequently used at $\beta = 0.15$ [2, 11]. Third, we can see that when $\alpha = 0.05$, $C(\boldsymbol{x}(\alpha), \boldsymbol{x}(\alpha'))$ decreases almost linearly as α' increases, while it decreases very little for small values of α' and only modestly for large values of α' when $\alpha = 0.95$.

Similarly to the above, let $\boldsymbol{y}(\beta)$ be the stationary vector defined as a limmiting solution of Equation (1) for PR with β. In order to examine the relation between VM and

PR, we calculated the correlation coefficients $C(\boldsymbol{x}(\alpha), \boldsymbol{y}(\beta))$ with respect to each pair of the uniform adoption probability α and the uniform jump probability β. In Fig.2, we plot $C(\boldsymbol{x}(\alpha), \boldsymbol{y}(\beta))$ with respect to α, where each result with different β is shown by a different marker. Here we also changed the values of β from 0.05 to 0.95 with an increment of 0.1.



(a) Ameblo network (b) Blog network

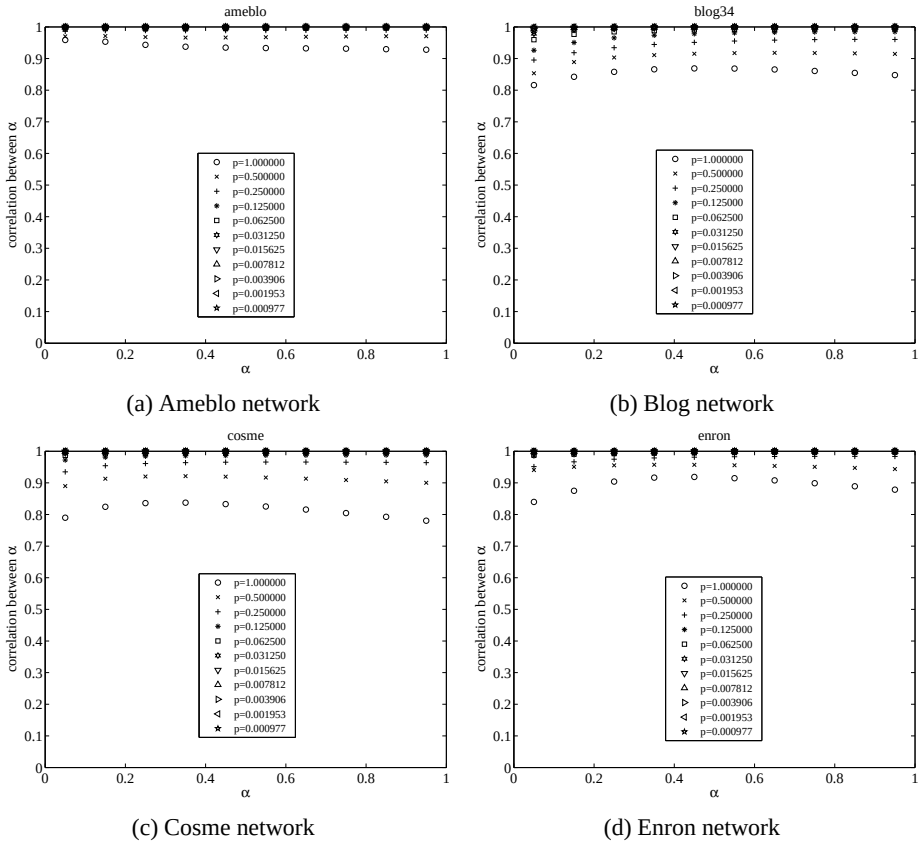
(c) Cosme network (d) Enron network

Fig. 2: The correlation coefficient between VM with α and PR with β

From Fig.2, we can observe the following similar relationships between VM and PR for all of the four networks. First, when α is small, there exists almost no correlation between the expected influence degree and the PageRank score. Second, for any β, $C(\boldsymbol{x}(\alpha), \boldsymbol{y}(\beta))$ generally increses as α increases, although their rates of increase depend on β as well as the network. Third, the maximum values of $C(\boldsymbol{x}(\alpha), \boldsymbol{y}(\beta))$ are attained at $\alpha = 0.95$. Incidentally, these maximum values are somewhat smaller than the correlation coefficients between in- and out-degree vectors, $C(\boldsymbol{B}, \boldsymbol{F})$, shown in Table 1, but their relative values are consistent between the two.

5.4 Effects of community structure

As the second series of experiments, we evaluated the effects of the community structure change on the expected influence degree. To this end, we constructed the 11 rewired networks from each of the original four networks using the rewiring probability $p = 2^{-k}$ ($k = 0, 1, \cdots, 10$) so that each network has a different community structure with different degree from the original one's (see the rewiring method in Section 5.1). Now, let $\boldsymbol{x}(\alpha, p)$ be the stationary vector calculated from the network rewired with probability p for VM with α. In order to evaluate the effects of different community structure, we calculated the correlation coefficients $C(\boldsymbol{x}(\alpha), \boldsymbol{x}(\alpha, p))$ with respect to each pair of the uniform adoption probability α and the rewiring probability p. In Fig.3, we plot $C(\boldsymbol{x}(\alpha), \boldsymbol{x}(\alpha, p))$ with respect to α, where each result with different p is shown by a different marker. Again we changed the values of α from 0.05 to 0.95 with an increment of 0.1.



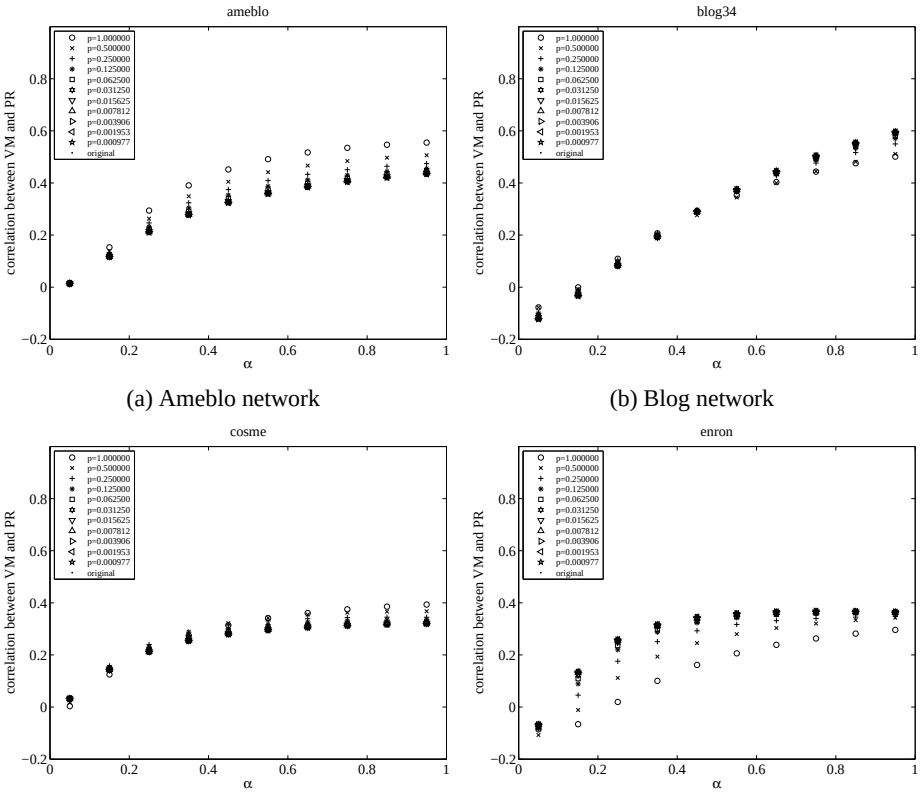
(a) Ameblo network (b) Blog network

(c) Cosme network (d) Enron network

Fig. 3: The correlation coefficient of VM between the original network and the rewiring network with p

From Fig.3, we can observe the following smilar characteristics of VM for all of the four networks. First, the correlation coefficients $C(\boldsymbol{x}(\alpha), \boldsymbol{x}(\alpha, p))$ for any pair of α and p are relatively high. Second, in comparison to Fig.1, there exist almost no ranges for α where $C(\boldsymbol{x}(\alpha), \boldsymbol{x}(\alpha, p))$ gives especially high values for all values of p. Third, $C(\boldsymbol{x}(\alpha), \boldsymbol{x}(\alpha, p))$ monotonically decreases as p increases. Overall, this experimental results suggest that the expected influence degree is not much affected by the community structure although the effect is more for a network with less community structure.

Similarly to the above, let $\boldsymbol{y}(\beta, p)$ be the stationary vector calculated from the network rewired with probability p for PR with β. In order to examine the relation between VM and PR in terms of community structure, we calculated the correlation coefficients $C(\boldsymbol{x}(\alpha), \boldsymbol{y}(\beta, p))$ with respect to each pair of the uniform adoption probability α and the rewiring probability p by setting $\beta = \alpha$. In Fig.4, we plot $C(\boldsymbol{x}(\alpha), \boldsymbol{y}(\alpha, p))$ with respect to α, where each result with different p is shown by a different marker.



(a) Ameblo network (b) Blog network

(c) Cosme network (d) Enron network

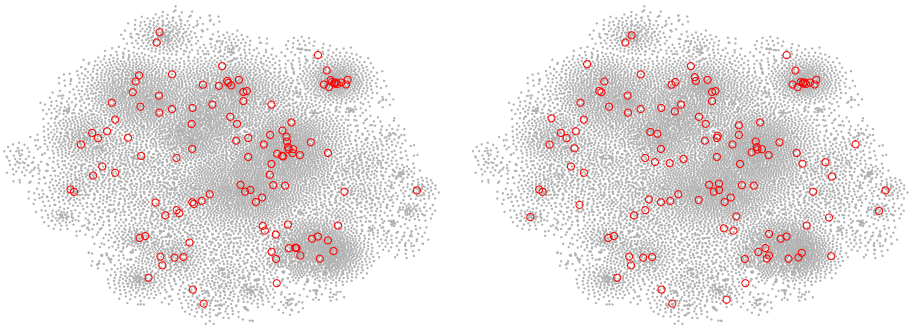
Fig. 4: The correlation coefficient between VM with α and PR with $\beta(=\alpha)$ and p

From Fig.4, we can see that for all of the four networks, each plotting result is very similar to the corresponding one appearing in Fig.2. Namely, the correlation coefficients $C(\boldsymbol{x}(\alpha), \boldsymbol{y}(\alpha, p))$ for any pair of α and p are relatively small. Further, this experimental results also suggest that the expected influence degree is not much affected by the community structure. As an interesting distinction, $C(\boldsymbol{x}(\alpha), \boldsymbol{y}(\alpha, p))$ is large when p is large for the Ameblo and Cosme networks, but a reverse tendency can be observed for the Blog and Enron networks. Clarifying this reason is left for our future work.

5.5 Visual analyses

We further analyzed the effects of the uniform adoption probability on the expected influence degree by visualizing the original networks. More specifically, we embeded the nodes in each network into a 2-dimensinal space by using the cross-entropy method [18], and plotted them as points. Then, we emphasized the highly influential nodes that have the expected influence degree within the top 1 % by using (red) circles. In the following experiments we only show the results using the Blog network as an example, but similar results were obtained for the other networks.

Fig.5 are the visualization results for two different values of α. Here we set α to 0.25 and 0.95 becauase they are considered to give the most and the least representative values for the expected influence degree as discussed in Section 5.3. From Fig.5, we can see that the highly influential nodes scatter around the entire the network for both α values. This partly explains the reason why the expected influence degree is not much affected by the community structure. This figure also shows that these two visualization results are close to each other.

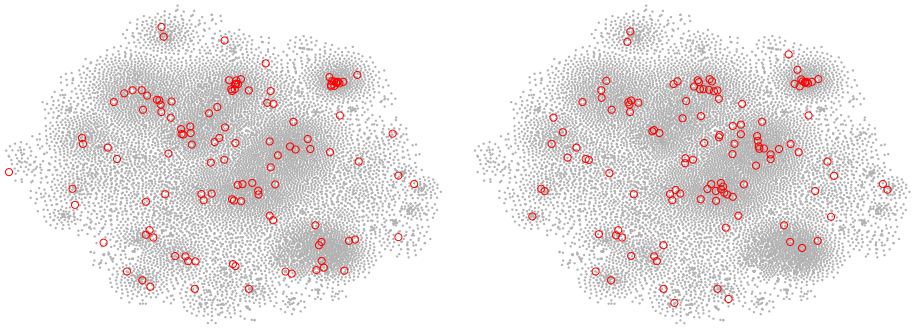


(a) Blog network at $\alpha = 0.25$ (b) Blog network at $\alpha = 0.95$

Fig. 5: Visualization of Networks (VM)

We also analyzed the resullts of PR to see if there is any difference between VM and PR. Fig.6 are the visualization results for PR, and the (red) circles are again the highly ranked nodes that have the top 1 % PageRank score. Here we set β to 0.25 and 0.95, the same as α. From Fig.6, we can also see that the highly ranked nodes scatter around

entire the network for both β values. Although we see that these nodes are different from the results of VM, but there is no clear difference bewteen the results of different β values.



(a) Blog network at $\beta = 0.25$ (b) Blog network at $\beta = 0.95$

Fig. 6: Visualization of Networks (PR)

6 Discussion

In this section, we discuss further extensions to our VM by employing majority rules and non-uniform adoption probability.

In our VM, with probability $(1 - \alpha)$, the opinion of each node $v \in V$ is influenced by choosing one of its parents nodes $B(v)$ and by any other node $u \in V$ with probability α according to $\boldsymbol{z}$. That is, our VM deals with all neighbor nodes equally by chosing one of neighbors at random. In a situation where a node decides its opinion considering the opinions of more than one neighbor, the q-voter model is one of the basic stochastic models [3]. In this model, the opinion of each node $v \in V$ is influenced by its chosen q parent nodes when their opinion is the same. Similarly, our VM can be further extended by adding some majority rules. We can also extend our VM with non-uniform adoption probability, that is, it might be natural to assume that not all friends or acquaintances have the same influence on a given node. To this end, we can introduce the weighted transition matrix $\boldsymbol{Q'}$ whose each element is defined by $q'(u, v) = w(u, v)/\sum_{u' \in V} w(u', v)$. Here, $w(u, v)$ is the weight over the link from a node $u \in V$ to a node $v \in V$ and $w(u, v) > 0$ if $a(u, v) = 1$; otherwise $w(u, v) = 0$. By using the column stochastic transition matrix $\boldsymbol{Q'}$, we can revise the Equation (8) as follows.

$$\boldsymbol{x}_t = \left((1 - \alpha)\boldsymbol{Q'} + \alpha \boldsymbol{z}\boldsymbol{e}^T\right)\boldsymbol{x}_{t-1} = (1 - \alpha)\boldsymbol{Q'}\boldsymbol{x}_{t-1} + \alpha \boldsymbol{z}. \tag{10}$$

In future, we plan to analyze these further extended models.

7 Conclusion

We addressed in this paper the problem of estimating the influential nodes in a social network, and focused on a particular class of information diffusion model, a model for opinion propagation. The popular model for opinion propagation is the Voter model in which the main assumption is that people change their opinion based on their direct neighbors, i.e. via local interaction. We extended this model to include the fact that people's opinion is also affected by the overall opinion distribution of the whole society. The new model is called the Voter Model with uniform adoption (the extended VM). It assumes that the network is directional because the people to people relation is directional.

The uniform adoption implies the random opinion adoption of all nodes in the network. We came to notice that this mechanism is the same as the random surfer jump of the well known PageRank algorithm. This motivated us to investigate the relationship between the extended VM and PageRank. We mathematically derived the ranking vector of the extended VM and compared it with that of PageRank, and explored how the two models are related by a series of extensive experiments using four real world social networks. The both models assume a directed network and give different rankings because the adjacency matrix is asymmetric. However, if we assume an undirected network in which the adjacency matrix is symmetric, the both models become identical and should give the same ranking. We investigated the effects of the uniform adoption probability on node ranking and how the ranking of the extended VM and PageRank are correlated to each other with this probability. The results indicate that the correlation varies with the uniform adoptiopn probability. The correlation is very small when the uniform adoption probability is small, but it becomes larger when both the uniform adoption and the random surfer jump probabilities become larger. However, the visualization rerults do not indicate the clear difference of the rankings between the different values of the uniform adoption probability. We also investigated how the different community structure affects the correlation, but did not see the strong effects. We found that the ranking becomes stable for the uniform adoption probability in the range of 0.15 and 0.35 and the self correlation within the extended Voter Model is high in this region. It is interesting to note that the reported recommended value for the random surfer jump of PageRank is 0.15, which is similar to our finding for the uniform adoption probability.

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-10-4053, and JSPS Grant-in-Aid for Scientific Research (C) (No. 20500147).

References

1. Agarwal, N., and Liu, H. (2008). Blogosphere: Research issues, tools, and applications. *SIGKDD Explorations* 10:18–31.

2. Brin, S. and Page, L. (1998) The anatomy of a large scale hypertextual Web search engine, In *Proceedings of the Seventh International World Wide Web Conference*, (pp. 107–117).
3. Castellano, C.; Munoz, M. A.; and Pastor-Satorras, R. (2009). Nonlinear q-voter model. *Physical Review E* 80:041129.
4. Crandall, D.; Cosley, D.; Huttenlocner, D.; Kleinberg, J.; and Suri, S. (2008). Feedback effects between similarity and sociai influence in online communities. In *Proceedings of KDD 2008*, (pp. 160–168).
5. Domingos, P., and Richardson, M. (2001). Mining the network value of customers. In *Proceedings of KDD 2001*, (pp. 57–66).
6. Even-Dar, E., and Shapria, A. (2007). A note on maximizing the spread of influence in social networks. In *Proceedings of WINE 2007*, (pp. 281–286).
7. 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).
8. 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).
9. Kimura, M.; Saito, K.; Nakano, R.; and Motoda, H. (2010). Extracting influential nodes on a social network for information diffusion. *Data Mining and Knowledge Discovery* 20:70–97.
10. Klimt, B., and Yang, Y. (2004). The enron corpus: A new dataset for email classification research. In: Proceedings of the 2004 European Conference on Machine Learning (ECML'04). (pp. 217–226).
11. Langville, A. N. and Meyer, C. D. (2005). Deeper inside PageRank, *Internet Mathematics*, **1:3** 335–380.
12. Leskovec, J.; Adamic, L. A.; and Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Transactions on the Web* 1:5.
13. Liggett, T. M. (1999). *Stochastic interacting systems: contact, voter, and exclusion processes*. New York: Springer.
14. Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167–256.
15. Richardson, M., and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of KDD 2002*, (pp. 61–70).
16. Sood, V., and Redner, S. (2005). Voter model on heterogeneous graphs. *Physical Review Letters* 94:178701.
17. Wu, F., and Huberman, B. A. (2008). How public opinion forms. In *Proceedings of WINE 2008*, (pp. 334–341).
18. 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).
19. Yang, S.; Chen, W.; and Wang, Y. (2009). Efficient influence maximization in social networks. In *Proceedings of KDD 2009*, (pp. 199–208).