Don't Count The Number of Friends When You Are Spreading Information in Social Networks

Hyoungshick Kim
Department of Computer Science and Engineering
Sungkyunkwan University
Suwon, Korea
hyoung@skku.edu

ABSTRACT

The problem of spreading information in social networks is a topic of considerable recent interest, but the conventional influence maximisation problem which selects a set of any arbitrary k nodes in a network as the initially activated nodes might be inadequate in a real-world social network – cyberstalkers try to initially spread a rumour through their neighbours only rather than arbitrary users selected from the entire network. To consider this more practical scenario, Kim and Eiko [16] introduced the optimisation problem to find *influential neighbours* to maximise information diffusion. We extend this model by introducing several important parameters such as user propagation rate on his (or her) neighbours to provide a more general and practical information diffusion model. We performed intensive simulations on several real-world network topologies (emails, blogs, Twitter and Facebook) to develop more effective information spreading schemes under this model. Unlike the results of previous research, our experimental results shows that information can be efficiently propagated in social networks using the propagation rate alone, even without consideration of the "number of friends" information. Moreover, we found that the naive random spreading would be used to efficiently spread information if k increases sufficiently (e.g. $k = 4$).

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services; J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Human Factors, Measurement

Keywords

Information diffusion, Information dissemination, Online social networks, Viral marketing, Influential neighbours, Propagation rate

IMCOM (ICUIMC)'14 2014, January 9–11, 2014, Siem Reap, Cambodia.
Copyright 2014 ACM 978-1-4503-2644-5...$15.00.

1. INTRODUCTION

In the field of social network analysis, a fundamental problem is to develop an epidemiological model and then to find an efficient way to spread (or prevent) information and ideas through the model. It seems natural that many people are often influenced by opinions of their friends. This is called the "word of mouth" effect and has for long been recognised as a powerful force affecting product recommendation. Recent advances in the theory of networks have provided us with the mathematical and computational tools to understand them better. For example, in the *Independent Cascade* (IC) model proposed by Goldenberg et al. [8], (1) some non-empty set of nodes are initially *activated* (or influenced); (2) at each successive step, the influence is propagated by activated nodes independently activating their inactive neighbours based on the *propagation probabilities* of the adjacent edges. Here, activated nodes mean the nodes which have adopted the information or have been infected. This models how a piece of information will likely spread through a network over time. It enables us to investigate what sort of information spreading strategies might be effective under certain conditions.

This model is also highly relevant to security. For example, cyberstalkers might be interested in spreading malicious rumours, gossip, news or pictures through social networks to damage the reputation of their victims (e.g. celebrity, political party, company or country). The same model works in social media campaign where spammers and propagandists want to share their advertisements on online social networks; fake accounts are often created and they can be used to amplify advertising campaigns using social media [20].

Thus far, however, the models and analytic tools used to analyse epidemics have been somewhat limited. Most previous studies aimed to analyse the process of information diffusion by choosing a set of any arbitrary k nodes in a network as the initially activated nodes from a bird's eye perspective based on the full control of nodes in the network and/or complete knowledge of the network topology, which may indeed be unacceptable in many real life networks since there is no such central entity (except the service provider itself).

From the point of view of an individual user who wants to efficiently spread a piece of information (or a rumour) through a network, a more reasonable model would not assume the knowledge about the entire network topology. Kim and Yoneki [16] recently introduced the problem called *Influential neighbour selection* (INS) where a spreader s spreads a piece of information (e.g. rumour) through the carefully chosen k of his (or her) neighbours instead of a set of any arbitrary k nodes in a network. Under this model, each user can only communicate with the user's immediate neighbours and has no knowledge about the global network topology except for his (or her) own connections. However, their work has two limitations: (1) it does not model users with varying levels of propagation rate on their neighbours, as information can always be propagated to neighbouring nodes with the same constant probability. Naturally, in real-world online social network services such as Twitter[1] or Facebook[2], each user has a different propagation rate for his (or her) neighbours on spreading information in a network according to the user's role such as opinion formers, leaders or followers [3]; (2) their experimental results were limited to undirected graphs with synthetic, rather than real, parameter values which were chosen in a somewhat ad-hoc manner.

In this paper, we extend the epidemiological model by introducing several parameters (*user propagation weight*, *decay factor* and *content interestingness* – see their formal definitions in Section 2) to provide a more general and practical information diffusion model. This gives much finer granularity than the previous model [16]. Under this realistic model, we empirically evaluated the performance of four reasonable spreading schemes from the simple random neighbour selection to a sophisticated neighbour selection scheme using both the "number of friends" and "user propagation weight" each neighbour has. To measure the performance of these schemes, we use the conventional *Independent Cascade* (IC) model [8], which is widely used for the analysis of information diffusion [8, 15, 10]. Our experimental results show that the scheme to select neighbours with a high propagation rate produced the best overall results, even without consideration of the "number of friends". We also found that even the naive random neighbour selection would be used to efficiently spread information if k increases sufficiently (e.g. $k = 4$). These results show that it is very difficult to prevent the spread of negative information (e.g. rumour) in social networks. For example, it might not be helpful to simply hide the "number of friends" each user has.

The rest of this paper is organised as follows. In Section 2 we formally define the Influential Neighbour Selection (INS) problem and notations. Then, we present the four reasonable neighbour selection schemes in Section 3. In Section 4, we evaluate the performance of the proposed schemes using real-world network topologies, and recommend how they should be used depending on the conditions. Some related work is discussed in Section 5. Finally, we conclude in Section 6.

2. INFLUENTIAL NEIGHBOUR SELECTION PROBLEM

In this section, we begin with the definition of the *Independent Cascade* (IC) model [8], and then introduce the *Influential Neighbour Selection* (INS) problem, which will be used in the rest of the paper.

We model an *influence network* as a directed graph $G = (V, E)$ consisting of a set of nodes V and a set of ordered pairs of nodes E called the edge set representing the communication links between node pairs. A directed edge (u, v) from node u to node v of the graph G is associated with a *propagation probability* λ which is the probability that v is activated by u through the edge in the next time step if u is activated. Here, v is said to be a *neighbour* (or successor) of node u. For node $u \in V$, we use $N(u)$ to denote the set of u's neighbours. The *out-degree* of node u is denoted as $d(u) = |N(u)|$. This metric could be used simply in estimating the node u's influence on information propagation.

In IC model [8], we assume that the time during which a network is observed is finite, from 1 until t; without loss of generality, the time period is divided into fixed discrete steps $\{1, \ldots, t\}$. Let $S_i \subseteq V$ be the set of nodes that are activated at the time step i. We consider the dynamic process of information diffusion starting from the set of nodes $S_0 \subseteq V$ that are initially activated until the time step t as follows: At each time step i where $1 \leq i \leq t$, every node $u \in S_{i-1}$ may activate its inactivated neighbours $v \in V \setminus S_{i-1}$ with an independent probability of λ. The process ends after the time step t with S_t. A conventional *Influential Maximisation* (IM) problem is to find a set S_0 of k nodes with the maximum number of activated nodes after the time step t for a budget constraint k.

The *Influential Neighbour Selection* (INS) problem [16] is a variant of the IM problem; given a spreader $s \in V$ and a budget constraint k, we aim to maximise the number of activated nodes in a network after the time step t by selecting s's $\min(k, d(s))$ neighbours only (rather than any subset of k nodes) as the set of nodes $S_0 \subseteq V$ that are initially activated. There are three limitations unlike the conventional IC mode: (1) each node only communicates with its immediate neighbours; (2) each node has no knowledge about the global network topology except for its own connections and (3) each message size is bounded to $O(\log |V|)$ bits.

Here we additionally introduce the three important parameters (user propagation weight ω, decay factor γ and content interestingness ϕ) to establish a more general and practical information propagation model.

The user propagation weight ω represents each user's average propagation rate to his (or her) neighbours. Given a user u, $\omega(u)$ is defined as $\tau(u)/(\rho(u)/d(u))$ where $\tau(u)$ and $\rho(u)$ are the number of u's posts shared by u's neighbours and the number of u's all posts, respectively. For example, if a user u with 1,000 neighbours wrote 10 posts and gets 100 shares, $\omega(u)$ is $100/(10 \cdot 1000) = 0.01$.

Also, previous studies [27, 12] showed that propagation probability λ can be greatly changed with the content of information (content interestingness ϕ). Naturally, higher content interestingness ϕ of a piece of information may facilitate higher propagation for the information through a network. Therefore we need to consider this too.

Finally, we define the decay factor γ at hop N as the ratio between the propagation probability at hop N and the propagation probability at hop $N-1$. In practice, the propa-

[1] http://www.twitter.com/
[2] http://www.facebook.com

gation probability might decay exponentially as the cascades spreads away from the information source and one possible explanation would be that the freshness of the information would drop as the time goes on.

Therefore, given an edge $(u, v) \in E$, a spreader $s \in V$ and a piece of information r, $\lambda(u, v, s, r)$ is finally defined as follows:

$$\lambda(u, v, s, r) \quad = \quad \min\{\omega(u) \cdot \phi(r) \cdot \gamma^{\delta(u,s,r)-1}, 1\} \qquad (1)$$

where $\delta(u, s, r)$ is the number of times the information r is to be relayed from s to u.

For example, when $\phi(r) = 0.0136$, $\delta(u, s, r) = 3$ and $\gamma = 0.2$, a user u with $\omega(u) = 1$ would activate his (or her) neighbour v with the probability of about 0.0005 ($\approx 1 \cdot 0.0136 \cdot (0.2)^2$).

3. NEIGHBOUR SELECTION SCHEMES

For the `INS` problem described in Section 2, we basically use a greedy strategy to select the influential neighbours.

Assume that a spreader $s \in V$ wants to spread a piece of information r through the network $G = (V, E)$ by sharing r with its $\min(k, d(s))$ neighbours only. Node s first tries to assess the influence of information diffusion for each neighbour $v \in N(s)$, respectively, by collecting the information about v. We note that neighbours' influence should be estimated based on s's local information only, rather than the whole network. As online social networks such as Facebook typically provide APIs to get the neighbourhood information about user, s might automatically collect the information about its own neighbours. After estimating the neighbours' influences, s selects the top $\min(k, d(s))$ nodes with the highest estimated values from $N(s)$ as the most influential neighbours for information diffusion; that is, for the `IC` model in Section 2, they are chosen as the set of initially activated nodes $S_0 \subseteq V$.

For the purpose of influence estimation, we test the following four selection schemes based on the "number of friends" and "user propagation weight" each user has.

- `Random` selection: Pick $\min(k, d(s))$ nodes randomly from $N(s)$. This scheme is very simple and efficient – the spreader s does not need any knowledge of the network topology.
- `Degree` selection: Pick the $\min(k, d(s))$ highest-degree nodes from $N(s)$. This scheme requires the degree knowledge of neighbours.
- `Propagation`-weight selection: Pick the $\min(k, d(s))$ highest user propagation weight nodes from $N(s)$. This scheme requires the user propagation weight knowledge of the nodes. To calculate $\omega(v)$ for s's neighbour $v \in N(s)$, the information about $\tau(v)$, $\rho(v)$ and $d(v)$ is required where $\tau(v)$ and $\rho(v)$ are the number of v's posts shared by v's neighbours and the number of v's all posts, respectively.
- `Hybrid` selection: Pick the $\min(k, d(s))$ nodes $v \in V$ with the highest *weighted* node degree $\omega d(v)$ which is defined as $\omega d(v) = \omega(v) \cdot d(v)$. At the first glance, this scheme requires the knowledge about both the degree and the user propagation weight of neighbours. In fact, however, this scheme can be simply implemented without the knowledge about node degree since $\omega(v) \cdot d(v)$ is calculated as $\tau(v)/\rho(v)$ – $d(v)$ is removed in the calculation.

The expected communication costs of all these schemes are $O(\kappa)$ where κ is the average out-degree in the graph.

Here we do not consider the other metrics (e.g. [26]) to estimate node centrality based on localized information alone since previous work [16] already showed that these metrics are not significantly effective for the `INS` problem compared with node degree.

4. EXPERIMENTAL RESULTS

In this section, we analyse the performance of the selection schemes presented in Section 3 on several real-world networks.

We summarise the properties of the networks used in experiments in Table 1. For **Facebook**, we used a dataset crawled in early 2008 of 26,701 nodes and 251,249 edges representing a regional sub-network of Facebook. For **Twitter**, we used a graph[3] consisting of mentions and retweets of some part of the Twitter network. The three notations κ, $\mathcal{D}$, and $\mathcal{C}$ represent the "average degree", "network diameter", and "number of connected components" (or "number of weakly connected components" for **Twitter**), respectively. The diameter of a network ($\mathcal{D}$) is the maximum distance between nodes in the network [21]; the diameter of a disconnected network is taken as infinite (inf).

Table 1: Summary of datasets used.

Network	Type	$\lvert V\rvert$	$\lvert E\rvert$	κ	$\mathcal{C}$	$\mathcal{D}$
Email [11]	Undirected	1,134	5,453	9.62	1	8
Blog [1]	Undirected	1,224	16,718	27.32	2	inf
Twitter	Directed	3,656	188,712	51.62	171	inf
Facebook	Undirected	26,701	251,249	18.82	1	15

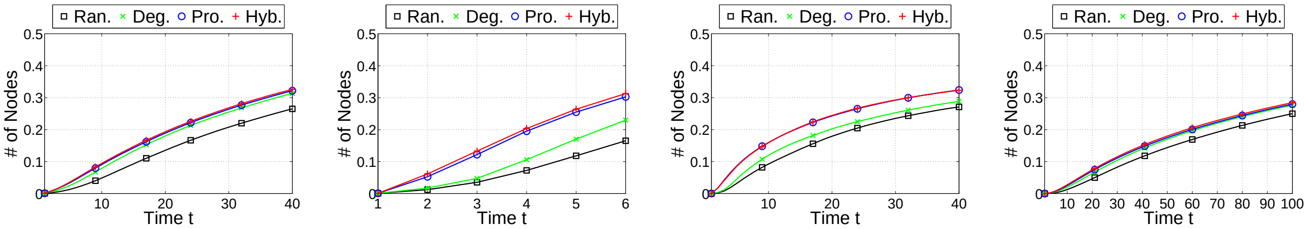
In this paper, our research interest is finding the best neighbour selection scheme to maximise information diffusion. We use the `IC` model in Section 2 to evaluate the performance of the schemes presented in Section 3 with varying the number of initially activated neighbours k. The propagation probability $\lambda(u, v, s, r)$ on an edge $(u, v) \in E$ is defined with the spreader $s \in V$ and a piece of information r described in Section 2.

For more realistic simulations of information propagation, given a user u, $\omega(u)$ is randomly drawn as a positive number from a normal distribution with mean 1 so that the Spearman's rank correlation coefficient between the "user propagation weights" and "numbers of neighbours" for all users is about 0.549. Note that we obtained this from real data – the Spearman's rank correlation coefficient between the "ranking by followers" and "ranking by retweets" for all users in Twitter is 0.549 [3]. In practice, $\omega(u)$ is usually

[3] `http://wiki.gephi.org/index.php/Datasets`

Table 2: The two-sample Kolmogorov-Smirnov test results ($\alpha = 0.05$) for the comparison of performance of the neighbour selection schemes.

Network	`Random` vs `Degree`		`Degree` vs `Propag.`		`Propag.` vs `Hybrid`	
	Test result	p-value	Test result	p-value	Test result	p-value
Email	**Different**	0.0034	Same	0.8567	Same	0.9762
Blog	**Different**	0.0000	**Different**	0.0000	Same	0.8110
Twitter	Same	0.1907	**Different**	0.0013	Same	1.0000
Facebook	**Different**	0.0152	Same	0.9885	Same	0.9954



(a) **Email** (b) **Blog** (c) **Twitter** (d) **Facebook**

Figure 1: Changes in the ratio of the average number of activated nodes to the total number of nodes in the network over time t.

much less than 1 but the mean of 1 might be fine to evaluate the performance of neighbour selection schemes since we also consider the content interestingness ϕ which is also much less than 1; given a piece of information r, $\phi(r)$ is randomly drawn from a normal distribution with mean 0.0136 and standard deviation 0.0501 according to real data [27]. We also set $\gamma = 0.2$ according to the mean of decay factors observed in [27].

In each simulation run, we randomly pick a spreader s for each of the networks in Table 1 and then select its k neighbours according to a selection criterion presented in Section 3. With fixed k, we repeated this 500 times to minimise the bias of the test samples (randomly selected spreaders); we measure the ratio of the average number of activated nodes per test sample to the total number of nodes in the network. For example, with $k = 1$, Figure 1 shows how these values are changed over time t under the `IC` model. Here, we use the different ranges of the time duration on the x-axis since the sizes of networks are totally different (see the number of nodes in each of the networks in Table 1). We performed the simulations in **Email**, **Blog**, **Twitter** and **Facebook**, respectively, after the 40th, 6th, 40th and 70th time steps to cover about a third size of of each network.

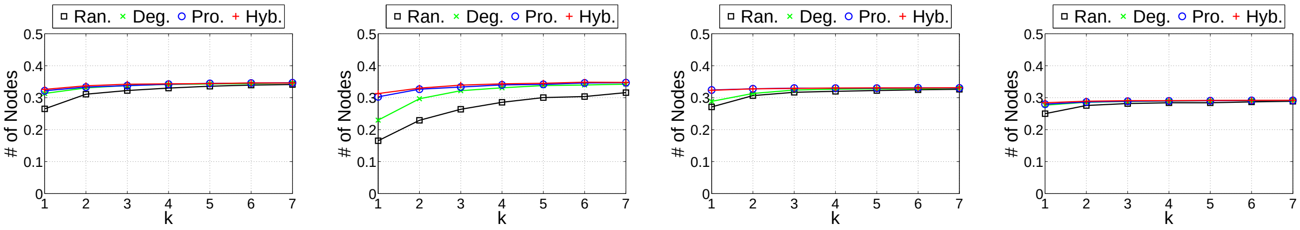
From this figure, we can see that `Hybrid` and `Propagation`-weight selection schemes outperformed other schemes in any network topology: When we use these schemes in **Blog** and **Twitter**, the ratios of the average number of activated nodes to the total number of nodes are over 0.3 while `Random` and `Degree` selection schemes are not (those are not particularly effective in spreading a piece of information for **Blog**). The two-sample Kolmogorov-Smirnov test [19] with $\alpha = 0.05$ was used to compare the performance of neighbour selection schemes in a statistically significant manner. We tested whether the distributions of the numbers of the activated nodes after the final time step between schemes at each network are statistically different. Table 2 shows the results of testing on each network topology in Table 1.

From Table 2, we can see that there is a significant gap between `Random` selection scheme and the other schemes. Also, `Propagation`-weight selection scheme significantly increases the number of activated nodes on average compared with `Degree` selection scheme in **Blog** and **Twitter**. We surmise that the differences of underlying network topologies may explain this. The average node degrees of **Blog** and **Twitter** are relatively large (27.32 and 51.62, respectively) while those of **Email** and **Facebook** are quite small (9.62 and 18.82, respectively). This shows that we would not recommend using `Degree` selection scheme when the average node degree is large. Interestingly, the number of activated nodes of `Hybrid` and `Propagation`-weight selection schemes were not significantly different for all network topologies although the `Hybrid` scheme is slightly greater than the `Propagation`-weight selection scheme in the average number of activated nodes. This implies that we can effectively spread information using the `Hybrid` scheme, even without consideration of the "number of friends" information.

We now discuss how the performance of the different neighbour selection schemes may change with the number of initially activated nodes k. To accelerate the speed of information diffusion, a possible straightforward approach is to increase the number of initially activated neighbours k. Probably, we can imagine that even the naive `Random` selection scheme can also be used to efficiently disseminate a piece of information if k increases sufficiently. In this context, our goal should be interpreted to find the minimum k for each scheme to converge to an optimal solution for information diffusion over time.

Table 3: The two-sample Kolmogorov-Smirnov test results ($\alpha = 0.05$) to analyse the performance of neighbour selection schemes by varying the size of the number of initially activated neighbours k.

Network	`Random`		`Degree`		`Propagation`		`Hybrid`	
	min k	p-value	min k	p-value	min k	p-value	min k	p-value
Email	**3**	0.4493	**2**	0.0555	**2**	0.2184	**2**	0.1658
Blog	**4**	0.2829	**2**	0.0909	**2**	0.9307	**2**	0.8970
Twitter	**2**	0.1436	**2**	0.1907	1	0.2491	1	0.4031
Facebook	**3**	0.6019	1	0.1907	1	0.4031	1	0.4493



(a) **Email** (b) **Blog** (c) **Twitter** (d) **Facebook**

Figure 2: Changes in the ratio of the average number of activated nodes to the total number of nodes in the network with the number of initially activated neighbours k.

To demonstrate the effects of the number of initially activated neighbours k, we first analyse the ratio of the average number of activated nodes with k ranging from 1 to 7 in **Email**, **Blog**, **Twitter** and **Facebook**, respectively, after the 40th, 6th, 40th and 70th time steps. The experimental results are shown in Figure 2.

From this figure, we can see that the effects of k may not be linear: the average number of activated nodes in all networks are still below 0.4 even for $k = 7$. When we use `Hybrid` and `Propagation`-weight selection schemes, k might not be an important factor for the information diffusion. However, `Random` and `Degree` selection schemes are rather affected by k although the effects of k are still inherently limited. The ratios of activated nodes in all networks except for **Blog** show almost the same pattern — the curves commonly have gentle slope from $k = 3$. As a selective strategy is at least as effective as random selection, we can always expect that it is enough to have three neighbours who can share the information regardless of the selection method used. The same conclusion was reached by using the two sample Kolmogorov-Smirnov test to analyse the average number of activated nodes in samples with a different k (see the results in Table 3).

In summary, our suggestion is to use `Hybrid` and `Propagation`-weight selection schemes with a small k. However, if we increase k sufficiently (e.g. $k = 4$), `Random` selection scheme would also perform well in spreading information.

5. RELATED WORK

Influential Maximisation (`IM`) problem has received increasing attention given the increasing popularity of online social networks, such as Facebook and Twitter, which have provided great opportunities for the diffusion of information, opinions and adoption of new products.

The `IM` problem was originally introduced for marketing purposes by Domingos and Richardson [7]: The goal is to find a set of k initially activated nodes with the maximum number of activated nodes after the time step t. Kempe et al. [15] formulated this problem under two basic stochastic influence cascade models: the *Independent Cascade* (`IC`) model [8] and the *Linear Threshold* (`LT`) model [15]. In the `IC` model each edge has a propagation probability and influence is propagated by activated nodes independently activating their inactive neighbours based on the edge propagation probabilities. In the `LT` model, each edge has a weight, each node has a threshold chosen uniformly at random, and a node becomes activated if the weighted sum of its active neighbours exceeds its threshold. Kempe et al. [15] showed that the optimisation problem of selecting the most influential nodes is NP-hard for both models and also proposed a greedy algorithm that provides a good approximation ratio of 63% of the optimal solution. However, their greedy algorithm relies on the Monte-Carlo simulations on influence cascade to estimate the influence spread, which makes the algorithm slow and not scalable.

A number of papers in recent years have tried to overcome the inefficiency of this greedy algorithm by improving the original greedy algorithm [18, 5] or proposing new algorithms [17, 5, 4]. For example, Leskovec et al. [18] proposed the *Cost-Effective Lazy Forward* (`CELF`) scheme in selecting new seeds to significantly reduce the number of influence spread evaluations, but it is still slow and not scalable to large graphs, as demonstrated in [4]. Kimura and Saito [17] proposed shortest-path based heuristic algorithms to evaluate the influence spread. Chen et al. [5] proposed two faster greedy algorithm called *MixedGreedy* and *DegreeDiscount* algorithms for the `IC` model where the propagation probabilities on all edges are the same; MixedGreedy is to remove the edges that have no contribution to propagate influence, which can reduce the computation on the unnecessary edges; DegreeDiscount assumes that the influence spread increases with node degree. Chen et al. [4] proposed the *Maximum Influence Arborescence* (MIA) heuristic based on local tree structures to reduce computation costs. Wang et al. [25] proposed a community-based greedy algorithm for identifying most influential nodes. The main idea is to divide a social network into communities, and estimate the influence spread in each community instead of the whole network topology. Several studies design machine learning algorithms to generate reasonable influence graphs by studying practical influence cascade model parameters from real datasets [2, 24, 23, 9].

More recently, as a variant of the conventional IM problem, Kim and Yoneki [16] introduced the problem called *Influential neighbour selection* (INS) to select the most influential neighbours of a node rather than the most influential arbitrary nodes in a network. They used the IC model for performance evaluation. In this paper, we extend this model by introducing several parameters (user propagation weight, decay factor and content interestingness) to provide a more general and practical information diffusion model.

Many studies mentioned that the levels of information sharing activity varied greatly between users in social networks. Romero et al. [22] argued that a majority of Twitter users might be passive, not engaging in creating and sharing information. Cha et al. [3] found that users with many followers do not necessarily influential in terms of spawning retweets or mentions – the Spearman's rank correlation coefficient between the "ranking by followers" and "ranking by retweets" for all users is 0.549. Zhou et al. [27] showed that in Twitter, the content of a tweet might be an important factor in determining the "retweet rate" – the mean retweet rate is 0.0136 but standard deviation is as high as 0.0501. Also, they observed that cascades tend to be wide not too deep indicates the retweet rate may decay as the cascades spreads away from the source – the mean of decay factors are all about 0.2. In this paper, we used those observed in real datasets as the parameter values for the simulations.

The INS problem might be applied to a wide range of social-based forwarding schemes [13, 6, 14]. It has mainly been proposed for Delay Tolerant Networks (DTNs), where the connection between nodes in the network frequently changes over time: the basic idea is to use node centrality for relay selections, and the forwarding strategy is to forward messages to nodes which are more central than the current node.

6. CONCLUSIONS

When faced with epidemic threats, it's important to understand the characteristics of epidemic spreading to stop or mitigate them. Kim and Eiko [16] introduced the optimisation problem to find *influential neighbours* to maximise information diffusion. We have extended their work by introducing several parameters (*user propagation weight*, *decay factor* and *content interestingness*) to provide a more general and practical information diffusion model.

We presented four neighbour selection schemes (Random, Degree, Propagation-weight and Hybrid selection) and explored their feasibility. We compared these selection schemes by computing the ratio of the average number of activated nodes to the total number of nodes in the network. We discussed which selection methods are generally recommended under which conditions. In summary,

- the best neighbour selection schemes in general are Hybrid and Propagation-weight selection schemes. For all tested networks (**Email**, **Blog**, **Twitter** and **Facebook**), it is enough to set $k = 1$ or 2.
- the simulation results of Hybrid and Propagation-weight selection schemes were not significantly different for all network topologies. This implies that we can effectively spread information using the Hybrid scheme, even without consideration of the "number of friends" information – the Hybrid scheme can be simply implemented without the knowledge about node degree since the contribution of node degree is removed in the calculation.
- Degree selection scheme performed well on network topologies with a small average degree; however, in network topologies with a large average degree, there is a significant gap between Degree and Propagation-weight selection schemes.
- if we increase k sufficiently (e.g. $k = 4$), Random selection scheme would also perform well in spreading information in social networks. This implies that it is very difficult to develop methods to prevent the spread of negative information (e.g. rumour) in real-world social networks. For example, it might not be helpful to simply hide the "number of friends" each user has.

As an extension to this work, we plan to consider a theoretical study to formally generalize and verify our results. We will also employ practical techniques to reduce the spread of information (e.g. rumour) by carefully monitoring users with a high "user propagation weight" or "number of friends".

Another interesting problem is to develop a more general model for information diffusion. We may consider not only a user's neighbours but also neighbours of neighbours as the candidate space of the initially activated nodes. In other words, we can extend the concept of the INS problem by expanding the set of the initially activated nodes with the distance from an information source node.

7. REFERENCES

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pages 36–43, New York, NY, USA, 2005. ACM.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 7–15, New York, NY, USA, 2008. ACM.

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. 2010.

[4] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1029–1038, New York, NY, USA, 2010. ACM.

[5] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 199–208, New York, NY, USA, 2009. ACM.

[6] E. M. Daly and M. Haahr. Social network analysis for routing in disconnected delay-tolerant MANETs. In *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*, MobiHoc '07, pages 32–40, New York, NY, USA, 2007. ACM.

[7] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 57–66, New York, NY, USA, 2001. ACM.

[8] 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*, pages 211–223, 2001.

[9] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 241–250, New York, NY, USA, 2010. ACM.

[10] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501, New York, NY, USA, 2004. ACM.

[11] R. Guimerà, L. Danon, D. A. Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Physical Review E*, 68(6), 2003.

[12] L. K. Hansen, A. Arvidsson, F. A. Nielsen, E. Colleoni, and M. Etter. Good friends, bad news - affect and virality in twitter. In *Future Information Technology*, volume 185 of *Communications in Computer and Information Science*, pages 34–43. Springer Berlin Heidelberg, 2011.

[13] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, WDTN '05, pages 244–251, New York, NY, USA, 2005. ACM.

[14] P. Hui, J. Crowcroft, and E. Yoneki. Bubble rap: social-based forwarding in delay tolerant networks. In *Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*, MobiHoc '08, pages 241–250, New York, NY, USA, 2008. ACM.

[15] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, New York, NY, USA, 2003. ACM.

[16] H. Kim and E. Yoneki. Influential neighbours selection for information diffusion in online social networks. In *IEEE ICCCN 2012, Munich, Germany*, 2012.

[17] M. Kimura and K. Saito. Tractable models for information diffusion in social networks. In *Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases*, PKDD'06, pages 259–271, Berlin, Heidelberg, 2006. Springer-Verlag.

[18] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 420–429, New York, NY, USA, 2007. ACM.

[19] F. J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[20] P. Metaxas, E. Mustafaraj, and D. Gayo-Avello. How (not) to predict elections. In *Proceedings of the 3rd international conference on Privacy, Security, Risk and Trust (PASSAT) and the 3rd international conference on Social Computing (SocialCom)*, pages 165–171, 2011.

[21] H. Per and H. Frank. Eccentricity and Centrality in Networks. *Social Networks*, 17(1):57–63, 1995.

[22] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. Influence and passivity in social media. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 113–114, New York, NY, USA, 2011. ACM.

[23] K. Saito, M. Kimura, K. Ohara, and H. Motoda. Selecting information diffusion models over social networks for behavioral analysis. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III*, ECML PKDD'10, pages 180–195, Berlin, Heidelberg, 2010. Springer-Verlag.

[24] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 807–816, New York, NY, USA, 2009. ACM.

[25] Y. Wang, G. Cong, G. Song, and K. Xie. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1039–1048, New York, NY, USA, 2010. ACM.

[26] K. Wehmuth and A. Ziviani. Daccer: Distributed assessment of the closeness {CEntrality} ranking in complex networks. *Computer Networks*, 57(13):2536–2548, 2013.

[27] Z. Zhou, R. Bandari, J. Kong, H. Qian, and V. Roychowdhury. Information resonance on twitter: watching iran. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 123–131, New York, NY, USA, 2010. ACM.