Influential Neighbours Selection for Information Diffusion in Online Social Networks

Hyoungshick Kim
University of British Columbia
Email: hyoung@ece.ubc.ca

Eiko Yoneki
University of Cambridge
Email: Eiko.Yoneki@cl.cam.ac.uk

***Abstract*—The problem of maximizing information diffusion through a network is a topic of considerable recent interest. A conventional problem is to select a set of any arbitrary k nodes as the initial influenced nodes so that they can effectively disseminate the information to the rest of the network. However, this model is usually unrealistic in online social networks since we cannot typically choose arbitrary nodes in the network as the initial influenced nodes. From the point of view of an individual user who wants to spread information as much as possible, a more reasonable model is to try to initially share the information with only some of its neighbours rather than a set of any arbitrary nodes; but how can these neighbours be effectively chosen?**

We empirically study how to design more effective neighbours selection strategies to maximize information diffusion. Our experimental results through intensive simulation on several real-world network topologies show that an effective neighbours selection strategy is to use node degree information for short-term propagation while a naive random selection is also adequate for long-term propagation to cover more than half of a network. We also discuss the effects of the number of initial activated neighbours. If we particularly select the highest degree nodes as initial activated neighbours, the number of initial activated neighbours is not an important factor at least for long-term propagation of information.

I. Introduction

In the field of social networks analysis, a fundamental problem is to develop an epidemiological model and then to find an efficient way to spread (or prevent) information, ideas, and infectious disease 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. [1], some non-empty set of nodes are initially *activated* (or influenced). 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.

Thus far, however, the models and analytic tools used to analyse epidemics have been somewhat limited. Most previous studies aimed to analyse the characteristics of information diffusion by choosing a set of any arbitrary k nodes in a network as the initial activated nodes. However, this model has assumed 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: there is no central node to communicate with any nodes in a network and/or to maintain global knowledge of the network topology.

From the point of view of an individual user who wants to efficiently spread information through a network, a more reasonable model is to choose the user's k neighbours as the initial activated nodes instead of a set of any arbitrary k nodes. This model is motivated by a practical scenario in online social networking services (e.g. Facebook or Twitter) — when a user u wants to advertise new information (or events), what is the best way to propagate the information through a network? Probably, the user u can ask u's neighbours who seem to be enormously influential in the network (e.g. users with many neighbours) to post this information in order to propagate this to their neighbours again. That is, in this paper, we seek to answer a simple question: "How can we select k neighbours to maximize information diffusion in a decentralized fashion?" Here, we assume that each user can only communicate with the user's immediate neighbours and has no knowledge about the global network topology except for its own connections.

We empirically study this problem through intensive simulation experiments on several real-world network topologies. In particular, we evaluated the performance of four reasonable selection schemes from a simple random selection strategy to other complicated selection strategies which take advantage of the knowledge of local connectivity such as node degree. To measure the performance of neighbours selection schemes, we use the *Independent Cascade* (`IC`) model [1], which is widely used for the analysis of information diffusion [1], [2], [3].

Our experimental results show that the strategies of using local connectivity of nodes produce similar results for a given budget. Thus more obvious recommendation would be to select high degree neighbours for short-term propagation since the other strategies based on local connectivity may incur a significant communication overhead without the performance improvement. Also, even a straightforward (naive) selection method (e.g. sharing information with their neighbours randomly) can be effective enough to spread information for long-term propagation to cover more than half of a network or large networks. Probably, in these environments, there is no good neighbour to maximize information diffusion.

The rest of this chapter is organised as follows. In Section II

we formally define the Influential Neighbours Selection (INS) problem and notation. Then, we present the four reasonable neighbours selection strategies in Section III. In Section IV, we evaluate the performance of the proposed strategies using real-world network topologies, and recommend how they should be used depending on the conditions. Some related work is discussed in Section V. Finally, we conclude in Section VI.

II. Model and Problem Formulation

In this section, we begin with the definition of the *Independent Cascade* (IC) model [1], and then introduce the *Influential Neighbours Selection* (INS) problem, which will be used in the rest of the paper.

We model an *influence network* as an undirected graph $G = (V, E)$ where V denotes the node set and E the edge set representing the communication links between node pairs. Each edge (u, v) of the graph G is associated with a *propagation probability* $\lambda(u, v)$, which is formalized by function $\lambda : E \rightarrow [0, 1]$. For simplicity, in this paper, we use a constant propagation probability λ for all edges.

For a pair of nodes u and $v \in V$, $\delta(u, v)$ denotes the number of hops on the shortest path between u and v; if u is not connected to v, $\delta(u, v) = \inf$. For node $u \in V$, we use $N_h(u)$ to denote the set of nodes within h distance from u. More precisely, $N_h(u) = \{v \in V \setminus u : \delta(u, v) \leq h\}$. When $h = 1$, we use $N(u)$ instead of $N_1(u)$ to particularly denote u's *neighbour set*.

The *degree* and the *clustering coefficient* of node u are denoted as $d(u) = |N(u)|$ and $c(u)$, respectively. The clustering coefficient of node u measure the probability of neighbours of node u to be neighbours to each other as well. This is calculated for u as the fraction of permitted edges between the neighbours of u to the number of edges that could possibly exist between these neighbours: $c(u) = 2 \cdot \Delta / (d(u) \cdot (d(u) - 1))$ where Δ is the number of the edges between the neighbours of node u. These two metrics $d(u)$ and $c(u)$ can often be used to analyse the u's local connectivity pattern.

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:

In IC model [1], 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 $\lambda(u, v)$. The process ends after the time step t with S_t. A conventional *Influential Maximization* (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 Neighbours Selection* (INS) problem is a variant of the IM problem; for a node $u \in V$ and a budget constraint k, we aim to maximize the number of activated nodes in a network after the time step t by selecting u's k neighbours rather than any subset of k nodes as the set of nodes $S_0 \subseteq V$ that are initially activated.

In this paper, we particularly consider a decentralized version of the INS problem to simulate users in online social networks such as Facebook or Twitter. That is, (1) each node only communicates with its immediate neighbours. Formally, a node $u \in V$ can only communicate with $v \in N(u)$; (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.

III. Neighbours Selection Criteria

We present the general framework of the INS problem for an online social network $G = (V, E)$ as follows.

Assume that a node $u \in V$ has some piece of information and wants to efficiently spread this information through the network G by sharing this with its $\min(k, d(u))$ neighbours only. Node u first tries to assess the influence of information diffusion for each neighbour $v \in N(u)$, respectively, by collecting the information about v. We note that v's influence should be estimated based on each node's local information only, rather than the whole network since u cannot build up the global network topology. As online social networks such as Facebook typically provide APIs to get the neighbourhood information about user, u might automatically collect the information about its own neighbours. Although users' personal information cannot be accessed by outsiders with the user's privacy preference settings, most users typically expose their degree and/or neighbourhood information to at least their neighbours and therefore u can easily collect the information of its neighbours. After collecting the information about neighbours, node u estimates their neighbours' influences and then selects the top $\min(k, d(u))$ nodes with the highest estimated values from $N(u)$ as the most influential neighbours for information diffusion; that is, for the IC model in Section II, 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 strategies based on local connectivity pattern such as node degree and/or clustering coefficient:

1) **Random** selection: Pick $\min(k, d(u))$ nodes randomly from $N(u)$.
   - This strategy is very simple and efficient: The user u does not need any knowledge of the network topology. The expected communication cost is $O(1)$.
2) **Degree** selection: Pick the $\min(k, d(u))$ highest-degree nodes from $N(u)$.
   - This strategy requires the degree knowledge of neighbours. The expected communication cost is $O(\kappa)$ where κ is the average degree in the graph.
3) **Volume** selection: Pick the $\min(k, d(u))$ highest volume centrality nodes from $N(u)$. Here, the volume centrality is defined as the sum of degree of all $w \in N_h(v)$. This metric was recently proposed by Wehmuth and Ziviani [4]. They experimentally showed that this metric is

highly correlated with the traditional *closeness* centrality which measures how quickly a node can communicate with all other nodes in a network. Closeness centrality is calculated for a node u as the average shortest path length to all other nodes in the network.

- This strategy requires the degree knowledge of the nodes within the distance h from v and the expected communication cost is $O(\kappa^{(h+1)})$. To calculate the volume centrality for each $v \in N(u)$, $\sum_{i=0}^{h} \kappa^i$ messages are required where κ is the average degree in the graph.

4) **Weighted**-volume selection: Pick the $\min(k, d(u))$ highest *weighted*-volume centrality nodes from $N(u)$. We extend the volume centrality metric to improve centrality estimation accuracy by additionally considering the relative weights of both the distance $\delta(v, w)$ between v and $w \in N_h(v)$ and the w's individual clustering tendency $c(w)$. As we might expect, closer nodes' connectivity has more contributions than far away nodes, so it would be preferable that the relative importance to local connectivity decreases as the distance from the node that we want to estimate (see Figure 1). All other things being equal, central nodes with low clustering coefficients may also be characterized as 'hubs' since they are actually linking neighbouring network parts that would be otherwise disconnected (see Figure 2).

- This strategy requires the degree and clustering coefficient knowledge of the nodes within the distance h from v and the expected communication cost is $O(\kappa^{(h+1)})$ which can be derived in the same way as above.

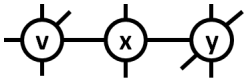


Fig. 1. An example to explain the relative weights of the distance $\delta(v, w)$ between v and $w \in N_h(v)$. In this example, we believe that x's connectivity has more contributions on v's centrality than y's connectivity.

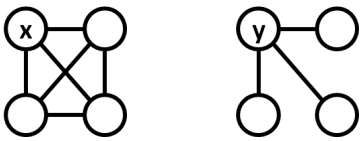


Fig. 2. An example to explain our centrality design philosophy for the relative weights of nodes' clustering tendencies. We compare a node with a high clustering coefficient when $c(x) = 1$ (left) and a node with a low clustering coefficient when $c(y) = 0$ (right). In this example, we believe that the y's role is more important in information diffusion than the x's role.

These functions are summarised in Table I. We will evaluate the performance and usefulness of these functions in Section IV.

IV. Experimental Results

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

Function	Influence of v	Cost
Ran.	1	$O(1)$
Deg.	$d(v)$	$O(\kappa)$
Vol.	$\sum_{w \in N_h(v)} d(w)$	$O(\kappa^{(h+1)})$
Wei.	$\sum_{w \in N_h(v)} d(w) \cdot (1 - c(w)) \cdot (1/2^{\delta(v,w)})$	$O(\kappa^{(h+1)})$

TABLE I
Summary of estimation functions.

We summarize the properties of the networks used in experiments in Table II. For **Facebook**, we particularly used a dataset crawled in early 2008 of 26,701 nodes and 251,249 edges representing a regional sub-network of Facebook. The three notations κ, $\mathcal{D}$, and $\mathcal{C}$ represent the "average degree", "network diameter", and "number of connected components", respectively. The diameter of a network ($\mathcal{D}$) is the maximum distance between nodes in the network [5]; the diameter of a disconnected network is taken as infinite (inf).

Network	$\lvert V \rvert$	$\lvert E \rvert$	κ	$\mathcal{C}$	$\mathcal{D}$
PGP [6]	10,680	24,316	4.55	1	24
Email [7]	1,134	5,453	9.62	1	8
Blog [8]	1,224	16,718	27.32	2	inf
Facebook	26,701	251,249	18.82	1	15

TABLE II
Summary of datasets used.

In order to show the usefulness of the node selection criteria proposed in Section III, we first calculate the Pearson correlation coefficients between closeness centrality and them, respectively: **degree**, **volume** and **weighted** influences (see Figure 3). Closeness centrality can be often applied to identify key nodes that are central in information dissemination processes [9].

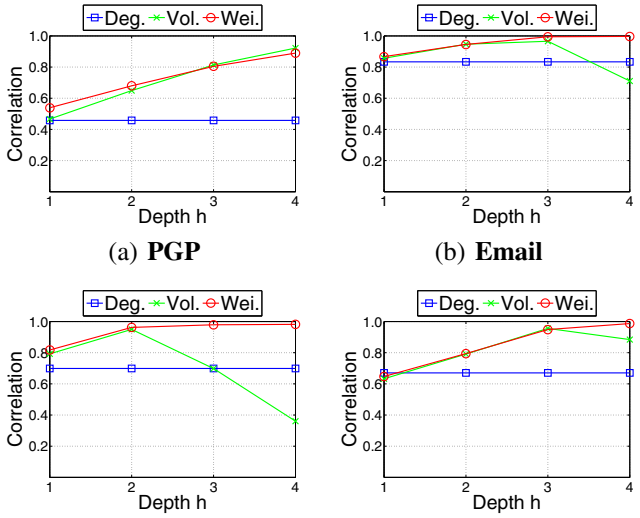


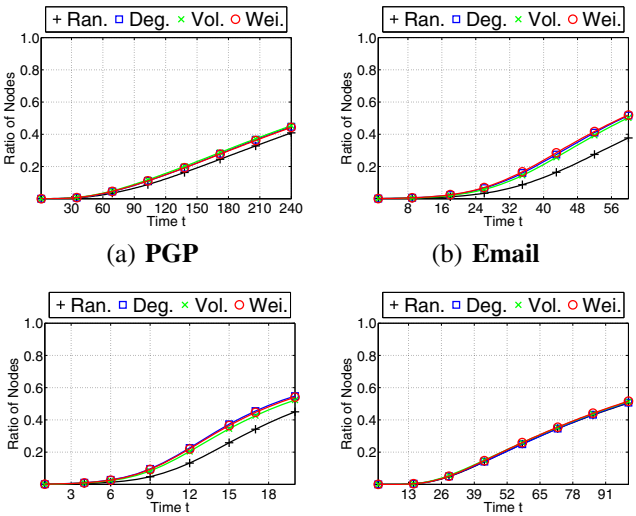
Fig. 3. The Pearson correlation coefficients between node influences in Section III and closeness centrality.

The **weighted**-volume and **volume** centrality values are more highly correlated with closeness centrality compared to **degree** if h is properly chosen. Moreover, as h increases,

the correlation coefficients with the **weighted**-volume centrality are significantly higher than those with the **volume** centrality [4] except for **PGP**. In particular, when $h = 4$, the **weighted**-volume centrality is almost correlated with closeness centrality although this trend appears to be rather weak in **PGP**. These results imply that the **volume** (if $h = 2$) or **weighted**-volume centrality (if $h \geq 2$) provides good approximations of closeness centrality.

In this paper, our research interest is finding the best selection strategy to maximize information diffusion. We use the `IC` model in Section II to evaluate the performance of the strategies presented in Section III with varying the number of initial activated neighbours k and a constant propagation probability λ on edges. We here set $h = 3$ for **weighted** and **volume** to give a good balance between accuracy of influence estimation and communication cost.

For simulation of `INS`, we randomly pick an information source node u for each of the networks in Table II and then select its k neighbours according to a selection criterion presented in Section III. With fixed k and λ, we repeated this 500 times to minimize the bias of the test samples (randomly selected information source nodes); 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 $k = 1$ and $\lambda = 0.01$, Figure 4 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 II).



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

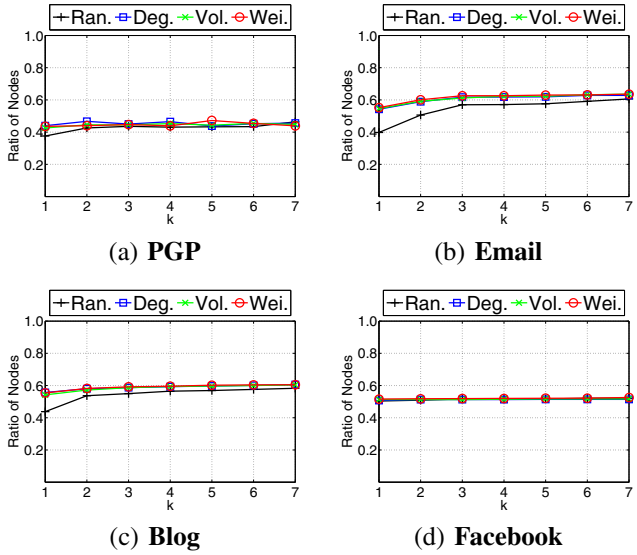
Fig. 4. Changes in the ratio of the average number of activated nodes to the total number of nodes in the network over time t.

From this figure, unlike the correlation coefficients with closeness centrality, we can see that all strategies except for **random** produce similar results with time t. The use of closeness centrality seems reliable in theory, but it may be effective less than thought. In practice, the **degree** selection strategy is at least as effective as **volume** and **weighted**. When we consider how expensive their costs ($O(\kappa^{(h+1)})$) are, we would not recommend using the **volume** and **weighted** selection methods.

Interestingly, there is a significant gap between **random** and the other strategies in **Email** and **Blog** while the **random** selection strategy is also comparable to the other strategies in **PGP** and **Facebook**. We surmise that the differences of underlying network topologies may explain this. The numbers of nodes of **Email** and **Blog** are relatively small (1,134 and 1,224, respectively) while those of **PGP** and **Facebook** are quite large (10,680 and 26,701, respectively). Surely, in a large network, the effects of initial activated nodes may be averaged over time to cover the many remaining nodes in the network.

Since the effectiveness of strategy choice remains rather limited, our research interest should naturally be shifted from choosing most influential neighbours for information diffusion to finding the optimal parameter values (e.g., k) for each strategy. To accelerate the speed of information diffusion, a possible straightforward approach is to increase the number of initial activated neighbours k. Probably, we can imagine that the naive **random** selection strategy can also be used to efficiently disseminate the user's information even for a small network such as **Email** and **Blog** if k increases sufficiently. In this context, our goal should be interpreted to find the minimum k for each strategy to achieve a reasonable level of information diffusion over time.

With the number of initial activated neighbours k ranging from 1 to 7, we discuss the effects of k. We divide the analysis into the two parts: 'long-term' and 'short-term' effects since they may be different in nature.



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

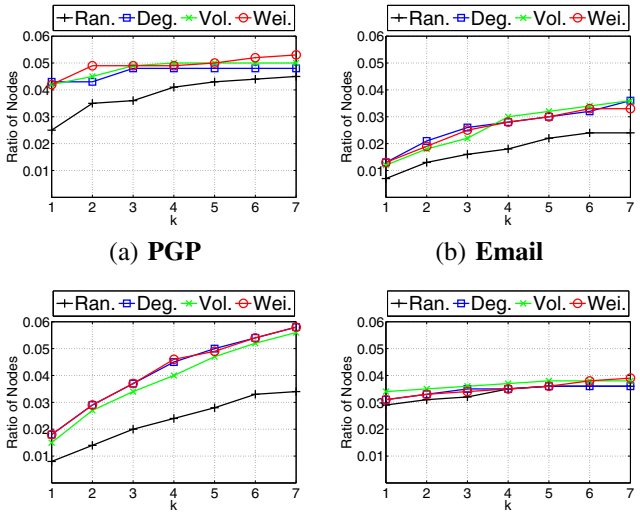
Fig. 5. Changes in the ratio of the average number of activated nodes to the total number of nodes in the network for a long-term with the number of initial activated neighbours k.

To demonstrate the *long-term* effects of k, we first analyse the ratio of the average number of activated nodes in **PGP**,

Email, **Blog**, and **Facebook**, respectively, after the 240th, 60th, 20th, and 100th time steps to cover more than half of each network. The experimental results are shown in Figure 5.

From this figure, we can see that the *long-term* effects of k may not be linear: the average number of activated nodes in all networks are still below 0.6 even for $k = 7$. When we use the **degree**, **volume** and **weighted** strategies, k is not an important factor in long-term propagation of information. This is natural enough; the relative importance of the number of the initial activated nodes is reduced over time. However, the **random** selection strategy is rather affected by k although the *long-term* effects of k are inherently limited. The ratios of activated nodes in all networks except for **Facebook** show almost the same pattern — the curves commonly have gentle slope from $k = 2$ or 3. As a selective strategy is at least as effective as random selection, we can always expect that it is enough to have two or three neighbours who can share the information regardless of the selection method used.

To discuss the *short-term* effects of k, we analyse the ratios of the average numbers of activated nodes in **PGP**, **Email**, **Blog**, and **Facebook**, respectively, after the 60th, 15th, 5th, and 20th time steps — the first quarter of the duration of the *long-term*. The experimental results are shown in Figure 6. For improved visualisation, we use the different range on the y-axis of this figure since the levels of the ratios of the average numbers of activated nodes in the networks are totally different from Figure 5.

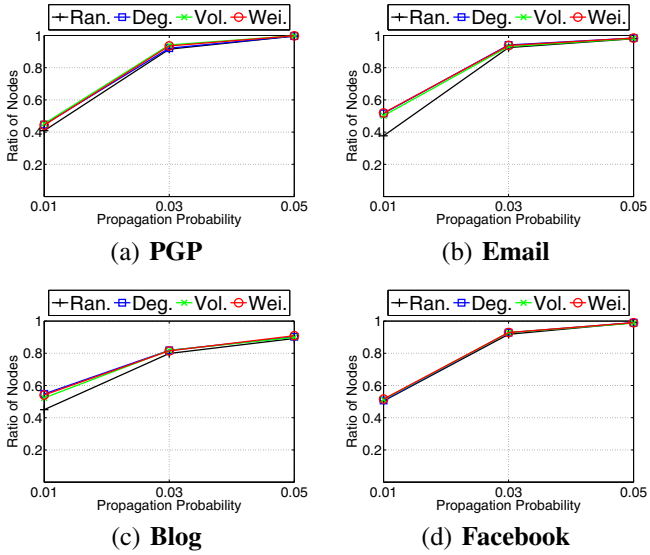


(a) **PGP** (b) **Email**

(c) **Blog** (d) **Facebook**

Fig. 6. Changes in the ratio of the average number of activated nodes to the total number of nodes in the network for a short-term with the number of initial activated neighbours k.

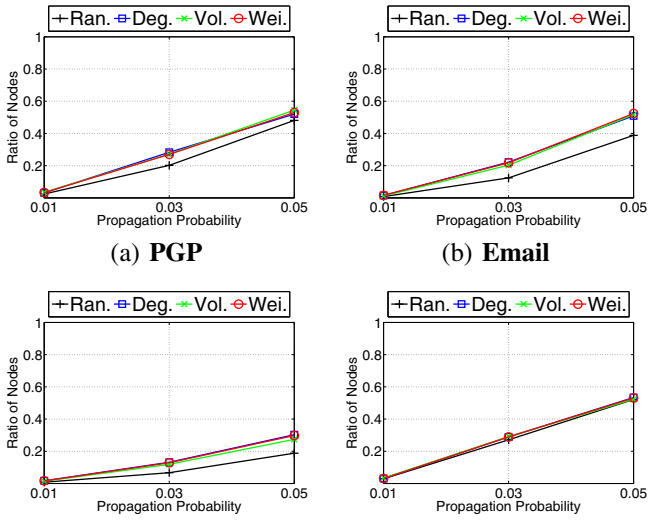
From this figure, we can see that the usefulness for short-term propagation of **degree**, **volume**, and **weighted** is better than the **random** selection for all networks except for a large network **Facebook**: the gap between them is clearly shown over k. That is, if one wishes to efficiently spread information for a short-term, one of the **degree**, **volume**, and **weighted** strategies should be carefully selected. Moreover, the choice of k might also be important in spreading information quickly. For example, in **Blog**, the ratio of the average number of activated nodes by each selection strategy increases linearly with k.



(a) **PGP** (b) **Email**

(c) **Blog** (d) **Facebook**

Fig. 7. Changes in the ratio of the average number of activated nodes to the total number of nodes in the network for a long-term with propagation probability λ.



(a) **PGP** (b) **Email**

(c) **Blog** (d) **Facebook**

Fig. 8. Changes in the ratio of the average number of activated nodes to the total number of nodes in the network for a short-term with propagation probability λ.

Finally, we discuss the effect of varying the propagation probability λ. In general, the speed of information diffusion is dramatically improved with λ. To demonstrate this we fix $k = 1$ and analyse the ratio of the average number of activated nodes for both of long-term and short-term propagation effects in the same manner as above. We here select $k = 1$ to

minimise the effects of k. The results are shown in Figure 7 and 8.

In these figures, as λ increases, so does the number of activated nodes in networks; this seems unsurprising.

The more interesting observation is that the gaps between **random** and the other strategies increases with λ for short-term propagation while they decrease reversely for long-term propagation. These results imply that the choice of selection strategy should be different based on the target duration (long-term vs. short-term) for information dissemination.

In summary, our suggestion is to use the **degree** strategy for short-term propagation but the **random** strategy for long-term propagation, respectively. Although the other centrality-based strategies **volume** and **weighted** produce similar results to those obtained by **degree**, they are not recommendable due to the incurring relatively high communication costs.

V. Related Work

Influential Maximization (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 [10]: 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. [11] formulated this problem under two basic stochastic influence cascade models: the *Independent Cascade* (IC) model [1] and the *Linear Threshold* (LT) model [11]. 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. [11] showed that the optimization 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 [12], [13] or proposing new algorithms [14], [13], [15]. For example, Leskovec et al. [12] 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 [15]. Kimura and Saito [14] proposed shortest-path based heuristic algorithms to evaluate the influence spread. Chen et al. [13] 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. [15] proposed the *Maximum Influence Arborescence* (MIA) heuristic based on local tree structures to reduce computation costs.

Wang et al. [16] 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 [17], [18], [19], [20].

In this paper, we use the IC model for the *Influential Neighbours Selection* (INS) problem as a variant of the IM problem to select the most influential neighbours of a node rather than the most influential arbitrary nodes in a network.

To estimate node influence of information diffusion in networks, we test the possibility of the local connectivity patterns of nodes rather than the simulations on influence cascade. Wehmuth and Ziviani [4] recently proposed a method to compute approximate closeness centrality, which uses only local information available at each node. Their study showed the possibility of the use of local connectivity to approximate closeness centrality. However, we showed the limitation of their centrality in applying this metric to the INS problem with on several real-network topologies.

The INS problem might be applied to a wide range of social-based forwarding schemes [21], [22], [23]. 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. Kim et al. [24] suggested some approximation methods to predict network centrality values for DTNs. Kim and Anderson [25] also proposed a model to measure the importance of a node by considering the time dimension.

VI. Conclusions

We introduced a new problem called the *Influential Neighbours Selection* (INS) problem to select a node's neighbours to efficiently disseminate its information. Previous studies had mainly aimed to develop solutions to select the most influential arbitrary nodes in a network for information diffusion. However, this model is not acceptable in many practical situations. For example, from the point of view of a user who wants to disseminate information through a network, it is desirable to consider to share the information with the user's neighbours only instead of any k nodes in a network; we empirically studied this through intensive simulation based on four real-world network topologies.

We presented four selection criteria from a simple random selection strategy to other complicated selection strategies which take advantage of the knowledge of local connectivity such as node degree and explored their feasibility. We compared these selection methods 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. We recommend using the **degree** selection strategy for short-term propagation but the **random** selection strategy for long-term propagation to cover more than half of a network, respectively. These strategies are thus amenable for large-scale, online and real-time computation. We also discussed the effects of the number of initial activated neighbours for each strategy. Interestingly, these effects may be rather limited. When we particularly used the **degree** selection strategy, the number of initial activated nodes is not an important factor at least in the `INS` problem for long-term propagation.

As part of this ongoing study, we plan to test community-based selection methods; if a user's neighbours are divided into several disjoint communities, we may improve the performance of information diffusion by selecting initially activated neighbours from different groups, respectively. 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.

ACKNOWLEDGEMENT

This research is part-funded by the EU grants for the RECOGNITION project (FP7-ICT 257756) and the EPSRC DDEPI Project, EP/H003959. We thank Ben Y. Zhao for his Facebook dataset.

REFERENCES

[1] 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*, pp. 211–223, 2001.

[2] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2003, pp. 137–146.

[3] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *WWW '04: Proceedings of the 13th international conference on World Wide Web*. New York, NY, USA: ACM, 2004, pp. 491–501.

[4] K. Wehmuth and A. Ziviani, "Distributed assessment of the closeness centrality ranking in complex networks," in *Proceedings of the Fourth Annual Workshop on Simplifying Complex Networks for Practitioners*, ser. SIMPLEX '12. New York, NY, USA: ACM, 2012, pp. 43–48.

[5] H. Per and H. Frank, "Eccentricity and Centrality in Networks," *Social Networks*, vol. 17, no. 1, pp. 57–63, 1995.

[6] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, "Models of social networks based on social distance attachment," *Physical Review E*, vol. 70, p. 056122, Nov 2004.

[7] 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*, vol. 68, no. 6, Dec 2003.

[8] 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*, ser. LinkKDD '05. New York, NY, USA: ACM, 2005, pp. 36–43.

[9] J. Tang, M. Musolesi, C. Mascolo, V. Latora, and V. Nicosia, "Analysing information flows and key mediators through temporal centrality metrics," in *Proceedings of the 3rd Workshop on Social Network Systems*, ser. SNS '10. New York, NY, USA: ACM, 2010, pp. 3:1–3:6.

[10] 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*, ser. KDD '01. New York, NY, USA: ACM, 2001, pp. 57–66.

[11] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proc. of ACM SIGKDD '03*, 2003, pp. 137–146.

[12] 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*, ser. KDD '07. New York, NY, USA: ACM, 2007, pp. 420–429.

[13] 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*, ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 199–208.

[14] 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*, ser. PKDD'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 259–271.

[15] 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*, ser. KDD '10. New York, NY, USA: ACM, 2010, pp. 1029–1038.

[16] 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*, ser. KDD '10. New York, NY, USA: ACM, 2010, pp. 1039–1048.

[17] 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*, ser. KDD '08. New York, NY, USA: ACM, 2008, pp. 7–15.

[18] 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*, ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 807–816.

[19] 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*, ser. ECML PKDD'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 180–195.

[20] 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*, ser. WSDM '10. New York, NY, USA: ACM, 2010, pp. 241–250.

[21] 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*, ser. WDTN '05. New York, NY, USA: ACM, 2005, pp. 244–251.

[22] 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*, ser. MobiHoc '07. New York, NY, USA: ACM, 2007, pp. 32–40.

[23] 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*, ser. MobiHoc '08. New York, NY, USA: ACM, 2008, pp. 241–250.

[24] H. Kim, J. Tang, R. Anderson, and C. Mascolo, "Centrality prediction in dynamic human contact networks," *Computer Networks*, vol. 56, no. 3, pp. 983–996, Feb. 2012.

[25] H. Kim and R. Anderson, "Temporal node centrality in complex networks," *Physical Review E*, vol. 85, p. 026107, Feb 2012.