Characterizing Collaboration in Social Network-enabled Routing

Manar Mohaisen[1], and Aziz Mohaisen[2]
[1] School of EEC Engineering, KoreaTech
330-708, Cheonan, R. of Korea
[e-mail: manar.subhi@koreatech.ac.kr]
[2] Computer Science and Engineering Department, SUNY Buffalo, NY, USA
[e-mail: mohaisen@buffalo.edu]
*Corresponding author: Aziz Mohaisen

Received July 1, 2015; revised January 21, 2016; revised February 12, 2016; accepted February 21, 2016; published April 30, 2016

Abstract

Connectivity and trust in social networks have been exploited to propose applications on top of these networks, including routing, Sybil defenses, and anonymous communication systems. In these networks, and for such applications, connectivity ensures good performance of applications while trust is assumed to always hold, so as collaboration and good behavior are always guaranteed. In this paper, we study the impact of differential behavior of users on performance in typical social network-enabled routing applications. We classify users into either collaborative or rational (probabilistically collaborative) and study the impact of this classification and the associated behavior of users on the performance of such applications, including random walk-based routing, shortest path based routing, breadth-first-search based routing, and Dijkstra routing. By experimenting with real-world social network traces, we make several interesting observations. First, we show that some of the existing social graphs have high routing costs, demonstrating poor structure that prevents their use in such applications. Second, we study the factors that make probabilistically collaborative nodes important for the performance of the routing protocol within the entire network and demonstrate that the importance of these nodes stems from their topological features rather than their percentage of all the nodes within the network.

Keywords: Social networks, collaboration, routing, random walks, adversarial behavior

A short version of this work appeared as a conference publication in proceeding of the 2012 International Conference on Communications [1].

http://dx.doi.org/10.3837/tiis.2016.04.010 **ISSN : 1976-7277**

1. Introduction

Social networks have gained exponential popularity that made online social networks sites some of the most important and significant traffic sources on the web. By their traffic volume, nine of the twenty most popular sites on the web are social networking sites [2]. Furthermore, the most popular social network, Facebook [3], serves 1.3 billion unique users, with more than 1 billion unique visitors per month as of the end of 2015. This popularity of social networks has motivated a wide spectrum of new technologies. Algorithms, protocols, and systems have been widely proposed on top of social networks, including applications to random-walk based routing [4]-[8], shortest-path based routing [9]-[11], social gossip [12]-[14], Sybil defenses [15]-[17], anonymous communication systems [18], [19], and information sharing [20], [21], among other technologies. Open-source software package of routing on top of social networks is available in [22], and is based on an ealier work that appeared in [23]. In its simplest form, a routing algorithm implemented on top of social networks aims to deliver a message between two nodes in the social graph, as highlighted in [24]. A survey of routing schemes is given in [25]. Applications of such routing algorithms include social experiments (outreach measurements, social influence, etc.), as well as other meaningful applications (such as communication during disasters) [5], [26]. In all of those applications, social ties are used to faciliate connectivity in a virtual network and to connect physically separated end-points. While these systems serve different purposes and follow different operational models, all of these schemes strike a balance among their algorithmic properties, connectivity, trust, and collaboration within the underlying social networks, all of which are utilized for bootstrapping such systems.

Collaboration is an essential feature of social networks. However, assumptions underlying collaboration are usually made to support end-results: all nodes are assumed to be collaborative in a categorical manner which is an unrealistic assumption about the behavior of typical participants in social systems. Using the forwarding application highlighted in [24], collaboration (or the lack of it thereof) indicates whether a user would like to engage in relaying messages on behalf of a source to a destination. To address this issue, we explore characterizing collaboration in social network based applications. We study the impact of classifying users into collaborative and rational (sometimes non-collaborative) users on such applications. Information routing applications built on top of social networks are the foci of this work, although we can extending findings easily to other applications, including social network-based Sybil defenses, and anonymous communication systems, among others.

To this end, the main contributions of this work are as follows.

- We propose a classification of users in social network-based systems to better understand how these systems are affected by social collaboration. Our classification divides users into two groups: collaborative and non-collaborative (rational). We explore the intuition of this classification from the social system viewpoint.
- We study the impact of our classification on routing in social systems. Thus, we consider several routing algorithms; random walks, shortest path, and breadth-first search (BFS) algorithms, and highlight how collaboration affects these algorithms.
- We suggest recommendations to incentivize collaboration in social systems, especially where collaboration of a few nodes greatly affects an entire social system.

By experimenting with real-world social network traces, our study unveils interesting results.

First, regardless of the level of collaboration of nodes, some social networks provide a poor performance. We find that route length in those social networks, on average, is large.

In our analysis, we study factors that impact the performance of routing application in social networks. We demonstrate that such performance does not only depend on the percentage of rational nodes, but rather on their topological properties. Such topological properties are essential and critical to the design and performance of social systems.

In our analysis, we use both online social networks and other orthognal networks, such as co-authorships, which bring social aspects to the communication network. Applications of those networks for routing may include social peer-to-peer networks in which nodes in social network participate in a peer-to-peer network, as accepted in the literature [15]-[17].

The rest of this paper is organized as follows. Section 2 introduces terminologies and preliminaries contained and referenced throughout the paper. Section 3 introduces the model for classifying users in the network based on their collaboration. Section 4 introduces our results for random walk-based routing on real-world social network traces. Related work is discussed in section 5 and section 6 concludes the paper.

2. Preliminaries

In this section we outline preliminaries used in the rest of this paper. In particular, we outline the network model, and breifly review various routing protocols.

2.1 Network Model

We represent the social network as an undirected and unweighted graph $G = (V, E)$, where $V=\{v_1,\ldots,v_n\}$ is the set of vertices, representing nodes in the social graph. $E=\{e_{ij}\}$ (where $1 \le i \le n$ and $1 \le j \le n$) is the set of edges connecting those vertices. $|V|=n$ denotes the size of G and $|E|=m$ denotes the number of edges in G. $A=[a_{ij}]_{n\times n}$ represents the adjacency matrix of G, where a_{ij} is defined as follows:

$$a_{ij} = \begin{cases} 1 & v_i \sim v_j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

In the rest of this paper, social network, network, and graph are used interchangeably to refer to both the physical network and the underlying social graph. A node is a user or a page in the social network and an edge is a relationship between a pair of users or pages.

Notice that the above model of undirected graph, which is best suited for routing in social systems. However, extending our results to directed social graphs is possible, although formalizing that theoretically would be a future direction.

2.2 Random Walk-based Routing

Random walk theory provides a framework for routing on networks. While it is nondeterministic on the length of a walk (cost) to an arbitrary destination, this technique can be effective in settings where nodes have only local knowledge of the topology or when the topology changes so frequently. In its simplest form, random walk based routing uses transition matrix P associated with G to randomly select forwarders at each node until the destination is reached. Recall A defined above, then $P=[p_{ij}]_{n\times n}$ is defined as follows:

$$p_{ij}=\begin{cases}\dfrac{1}{\deg(v_i)} & v_i \sim v_j \\ 0 & \text{otherwise}\end{cases} \tag{2}$$

Let v_s be the source and v_d be the destination to which a packet is intended. v_s uniformly at random selects one of her neighbors and forwards the packet towards her. At each time slot, the intermediate node on the path between the source and the estination checks if the destination is among its neighbors. If so, the intermediate node directly forwards the packet to the destination. Otherwise, the intermediate node performs the same random procedure by uniformly selecting one a neighbor and forward the walk towards it. This process is performed at each intermediate node until the condition $v_i = v_d$.

2.3 Shortest Path Based Routing

The shortest path based routing uses the shortest directed distance between two nodes. Let $w:E\rightarrow R$ be a weight function that assigns real-valued weights to edges in G. The weight of path p= $\langle v_1, v_2, \ldots, v_l\rangle$ is

$$w(p)=\sum_{r=1}^{\ell} w(v_{r-1}, v_r). \tag{3}$$

Furthermore, the shortest path between nodes v_i and v_j is defined as:

$$\delta(v_i, v_j)=\begin{cases}\min\{w(p): v_i \overset{p}{\rightarrow} v_j\} & \text{path from } v_i \text{ to } v_j \\ \infty & \text{otherwise}\end{cases} \tag{4}$$

where the set of nodes connecting two nodes v_i and v_j is denoted by $v_i \overset{p}{\rightarrow} v_j$. Since it may not be unique, a shortest path between v_i and v_j is any path with weight $w(p)=\delta(v_i,v_j)$. In its simplest form, computing the shortest path requires a global knowledge of the topology .

Many routing systems, including the OSPF (open shortest path first) algorithm, use Dijkstra's algorithm, which we also use in our paper. Unlike the BFS algorithm, Dijkstra's algorithm finds the shortest-path between nodes v_i and v_j in the case of non-negative weights assigned to the edges of G. As such, the shortest path between nodes v_i and v_j is defined as

$$\delta(v_i, v_j)=\min\{w(p): v_i \overset{p}{\rightarrow} v_j\} \tag{5}$$

if there is a path from *vi* to *vj*, or $\delta(v_i,v_j) = \infty$ otherwise [27].

In this context, and only for experiments, the weights of the edges are calculated using the Jaccard similarity coefficient, where the weight of the edge is the similarity between its vertices. The main motivation of using such weights is our interest in measuring how the topological structure of the graph influences its behavior. Any other metric, such as the reputation-based weights or cost-based weights, can be used for assigning weights.

2.4 Breadth-first Search Routing

Breadth-first search (BFS) algorithm is an archetype for many graph algorithms such as Dijkstra's shortest path algorithm. We use the textbook description of BFS found in [27], and refer the interested reader to the textbook for completeness.

3. Collaboration in Social Networks

In this section, we classify users in the social graphs into two categories and study the impact of this classification on the performance of routing algorithms within social networks. First, we classify users in the social network into collaborative users and probabilistically collaborative users. We then study the impact of this classification on the routing protocol.

3.1 Users Classification

3.1.1 Collaborative Users

In many distributed systems, such as peer-to-peer systems with file-sharing applications, the performance of the system depends on altruistic behavior of users. While the system is governed by a generic economical principle (e.g., tit-for-tat), a few users - the high percentile of users in terms of resources such as bandwidth, processing, and memory - contribute the majority of the resources required for the operation of such networks. Altruistic users, in our settings as well as in the general settings of any distributed system, participate in the system and serve others without expecting the same treatment from other users, i.e., in many cases, violating the tit for tat principle. In this work, collaborative users are altruistic and follow the routing protocol by dedicating all resources to forward messages on behalf of others. Vc denotes such nodes and the level of altruism is denoted by γ.

3.1.2 Probabilistically Collaborative Users

These users, who are otherwise called rational users, act less altruistically than collaborative users. In particular, while the altruism (i.e., parameterized by γ) of collaborative users is close to one, and hence the selfishness characterized by $\alpha = 0$, the altruism of rational users is characterized by α where $0 < \alpha < 1$. Typically, a rational node may participate in the routing protocol with probability $1-\alpha$ and deviate from it by dropping incoming packets or not collaborating with probability α. V_p denotes such nodes in G and its size (as a fraction of the size of the network) is β. Note that $\alpha + \beta = 1$. With that in mind, we make the following observations, assumptions, and rules:

- We note that each group of nodes in the graph is mutually exclusive, meaning that a node in the graph can only belong to one of the two categories above - hence V_p and V_c are exclusive subsets of V where $V=V_p \cup V_c$.
- We assume that a node may not change its behavior over the run time of the protocol, to simplify the analysis concerning the behavior of every node in the graph and how the behavior of a particular node would impact the overall performance. This is, through this assumption we try to marginalize the impact of the randomness generated by nodes altering their behavior. By doing so, we try to understand what underlying qualities of the graphs affect the operation of these routing algorithms.
- Last, although we use the same α for all nodes in the graph, this value may be adjusted per node to express a more realistic characterization of behavior, as we will see later. We omit the details of extending this parameter to the general case, and leave that as a future work.

3.1.3 Collaboration Impacts Routing

The behavior of the different nodes in the graph is shown as a state diagram in **Fig. 1**. The routing protocol is then described as a *biased* random walk where the event of not collaborating in the routing protocol is denoted by a loop from each node to the originator of the route. For simplicity, in the figure, we remove collaborative nodes, which can be seen as nodes across the route with α=0. As for the shortest-path and BFS-based routing, lack of collaboration results in shortest-path search failure. Accordingly, similar to above, we consider the percent of shortest-path (and BFS-based search) routing trials that fail among all possible trials in the graph among possible source-destination pairs.

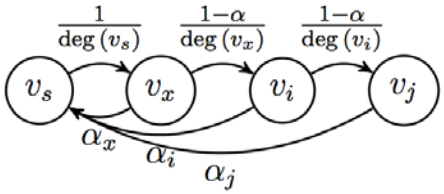

Fig. 1. Random-walk based routing on graph in mixed settings

3.2 Theoretical Formulation of Routing Cost

To characterize the routing cost (whether it is a random walk length, shortest walk length, or length number of rounds until a gossiping algorithm converges to the entire network), we provide the following theoretical analysis. First, we assume that the initial cost, when there are no selfish nodes deployed in the network, is w, and derive the cost in terms of that original cost. In the following, we consider two models for characterizing cost due to collaboration: cost under retransmission and cost under restarts.

3.2.1 Costs with Retransmissions

We consider a simpler model: a model where the selfishness of nodes is characterized as a self-loop, where nodes don't drop messages at the path to destination, but collaboration is considered as an additional cost. For that, we compute the expected cost $E[w']$ due to transmission failures because of selfishness on the path. Recall the probability of having exactly k selfish nodes on a path of length w (assuming they are independent) is given as

$$P(k) = \binom{w}{k} p^k q^{w-k} \tag{6}$$

where $p=\alpha$ and $q=1-\alpha$. Using the law of expectation, we have $E[w'] = \sum_{\forall i} P(i) w_i$. This is,

$$\begin{aligned} E[w'] = \sum_{i=0}^{w} P(i) w_i &= \sum_{i=0}^{w} \binom{w}{i} \alpha^i (1-\alpha)^{w-i} w_i \\ &= \sum_{i=0}^{w} \binom{w}{i} \alpha^i (1-\alpha)^{w-i} (w+i). \end{aligned} \tag{7}$$

Notice that w_i and w' denote the length of an arbitrary path and the cost of transmission failure due to users' selfishness on the path, respectively. The last equality happens because a single failure at the i-th step of the random walk would require only one additional step as a cost regardless of the location of the node on the path to the destination. Accordingly, k of such

failures at any different locations in the path would result in *k* additional steps. This can be simply added into the sum to compute the expectation.

3.2.2 Costs with Restarts

Now we address the general case where a failure due to selfishness of nodes results in rejection and restart of the walk. Similar to the case above, we have the general form of

$$E[w'] = \sum_{i=0}^{w} P(i)w_i = \sum_{i=0}^{w} \binom{w}{i} \alpha^i (1-\alpha)^{w-i} w_i. \tag{8}$$

However, the characterization of w_i in this case is different from that explained above. In particular, if a random walk fails only at one node, the length of the total walk (accumulates for the cost due to the failure) would depend on the location of the failure. This is, if the failure happens at the *i*-th location in the path to the destination, the total number of steps, assuming success in the second round, would be *w*+*i*. Accordingly, we compute the expected length of the added steps due to the failure, by floating the location of the failure at any *i* where $0<i\leq w$. For that, let's denote the expected cost due to one failure and retransmission by $\overline{w_i^1}$, where the failure happens in any of the *i* possible locations of the failure (in this case, the number of failures is *w* but we extend the notation to *i* so that to cover more than one failure, in which case *i* can be up to $i \times w$). Accordingly,

$$\overline{w_i^1} = (\sum_{i=0}^{w} w + i) \,/\, w = \frac{3w+1}{2}.$$

For two failures, the location of both failures determines the cost, where the expected length is computed as

$$\overline{w_i^2} = (\sum_{i=0}^{2w} w + i) \,/\, 2w = \frac{4w+1}{2}.$$

For the ultimate case of $i=w$, we have

$$\overline{w_i^2} = (\sum_{i=0}^{w^2} w + i) \,/\, w^2 = \frac{(w+1)^2}{2}.$$

By summing all cases above, and plugging the result into (4), we have

$$\begin{aligned} E[w'] = \sum_{i=0}^{w} P(i)w_i &= \sum_{i=0}^{w} \frac{\sum_{j=1}^{i\times w}(w+j)}{i \times w} \binom{w}{i} \alpha^i (1-\alpha)^{w-i} \\ &= \sum_{i=0}^{w} \frac{w(i+2)+1}{2} \binom{w}{i} \alpha^i (1-\alpha)^{w-i}. \end{aligned} \tag{9}$$

Finally, to compute the total cost, we consider the conditional probability of the following events. Let A_i denote the most recent event of transmission which is considered successful (no failure is triggered at any node on the path to the destination), and let $A_{i-1}... A_1$ denote the set of events that resulted in failure prior to A_i. We want to compute the conditional probability $P(A_i = 1 \mid A_{i-1} = 0, A_{i-2} = 0, \ldots, A_1 = 0)$.

Fig. 2 shows a comparison of the two models above, for varying values of α and *w*. From

those plots we notice that the overall cost is linear in α, and that the restart model is orders of magnitude more costly than that of the retransmission model. Also notice that we do not enforce any of the above models, by proposing them as two different models with different design settings: if the retransmission model is to be used, we need to maintain states in the network for operating the routing algorithm; these states get updated upon confirmed and verified collaboration or trigger retransmission (perhaps via a different forwarding node) if collaboration fails.

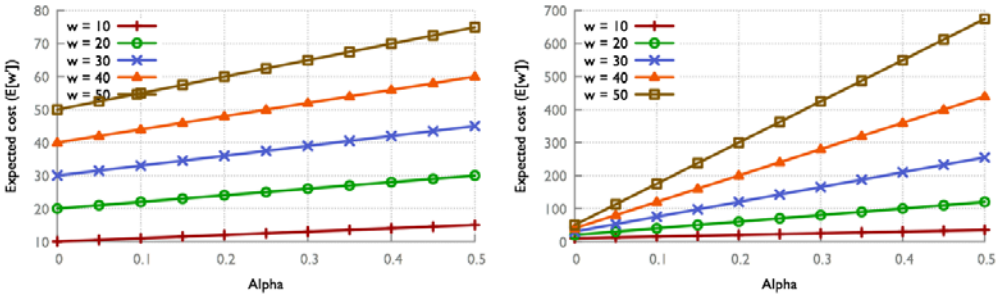

Fig. 2. The cost for varying α values and initial costs. Notice that the cost of running the given routing algorithm, with respect to the initial cost, is linear in w and α. Also notice that in relation with the cost under the retransmission model, the cost with the restart model is orders of magnitude larger.

4. Results and Discussion

We introduce the results of this study and elaborate on the findings. The social graphs used in this study are shown in **Table 1**, and their degree distributions are shown in **Figure 3**.

Table 1. Social graphs used in this study. Physics 1-3 are relativity, high-energy, and high-energy theory co-authorship [28]. **D** stands for the diameter and **R** stands for the radius of each graph.

Social network	Nodes	Edges	D	R
Physics 1 [28]	4,158	13,428	17	9
Physics 2 [28]	11,204	117,649	13	7
Physics 3 [28]	8,638	24,827	18	10
Wiki-vote [29]	7,066	100,736	7	4
Enron [28]	10,000	108,373	4	2
DBLP [30]	10,000	20,684	8	4
Facebook [31]	10,000	81,460	4	2
Youtube [32]	10,000	58,362	4	2

Some of these graphs are sampled from larger graphs using the breadth-first search [33]. As an indicator of the topological structure of the different graphs, we compute both the diameter and radius of each graph. By defining the eccentricity as the set of maximal shortest paths from each and every source to other destinations in the graph, the diameter is defined as the maximal eccentricity and the radius is defined as the minimal eccentricity. We compute diameters and radii to provide insight on the structure of the graphs. We observe that these parameters differ greatly from a graph to another, which implies different graph structures.

Furthermore, we refer the reader to [33] where it is shown that the structure of the graph, using the mixing time measure, is different and varies in each of these graphs.

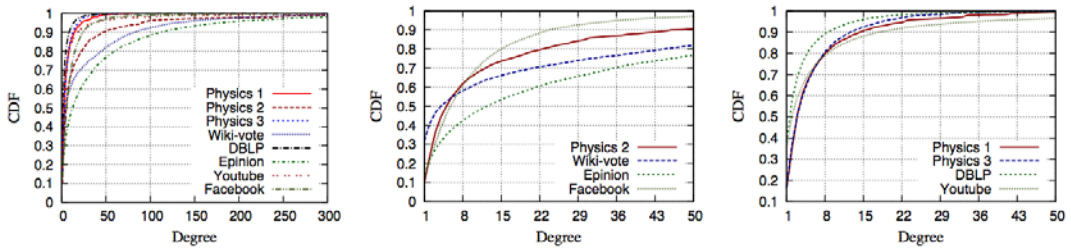

Fig. 3. Degree distribution of the graphs used in this study: (a) shows all data sets on long range of the degrees, (b) shows data sets with large degrees, and (c) shows data sets with small degrees.

4.1 Evaluation Metric

To evaluate the various schemes in similar settings, we use a unified evaluation metric: the cost of operating the different routing schemes on a given social graph. This is, the evaluation metric of the different routing schemes is the normalized expected number of transmissions per single message delivery operation between a source and a destination. The definition depends on the used algorithm, and is defined as

$$E[cost] = \frac{1}{S}\sum_{i=1}^{S} cost_i \tag{10}$$

where S is the sample size and $cost_i$ is the cost for a given pair of source and destination indexed by i. We define the normalized cost depending on the algorithms used as follows:

- Random walk-based routing: Normalized by the number of nodes in G as $E[cost]/n$
- Shortest path-based routing: Normalized by the diameter of the graph d; i.e., $E[cost]/d$.

4.2 Performance in Ideal Settings

In this Subsection, we study the performance of routing on social graphs in ideal settings, without considering collaboration as a constraint.

4.2.1 Random Walk-based Routing on Social Graphs

Considering the different graphs in **Table 1**, we first measure the performance of the simple random walk-based routing explained in section 2. We define the cost of routing over graphs as the normalized expected number of transmissions, which is the average number of times that a node in the graph transmits a packet for a single routing session from a given source to a given destination. To avoid the random behavior and bias in the measurements, we consider the case of routing a single packet from a given source (selected uniformly at random from the graph) to 1,000 arbitrary destinations, which are also selected uniformly at random from the graph. To reduce the bias in the measurements, we perform the same experiment, for the same source-destination pair, for 1,000 times, totaling 1,000,000 routing trials per data set. In addition, we take the average number of hops, which is then normalized by the network size, to

give the expected number of transmissions. The results of these measurements are shown in **Fig. 4**.

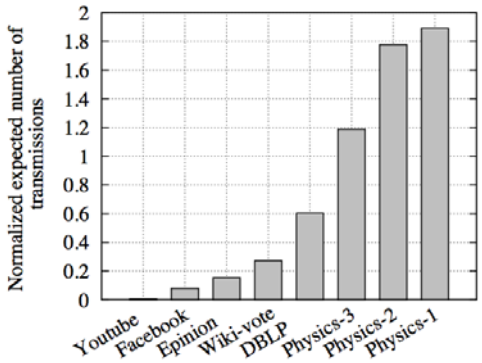

Fig. 4. Random walk based routing on graph in normal fully collaborative. The various social graphs exhibit different structures with varying qualities of the performance.

While they are close in size, we observe that the cost of routing over the different graphs is different, and correlated to the underlying graph structure. In principle, the performance of the graphs can be classified into two categories: well-performing graphs (Epinion, Youtube, Wiki-vote, and Facebook) and poorly performing graphs (Physics-1 to 3 and DBLP). We observe that the poor-performing graphs exhibit a strong community structure, as evidenced by their high modularity - a measure of the community structure in social networks; for more details on the term and its measurements in some graphs including those used in this paper see [34]. On the other hand, well-performing graphs have less clear community structure evidenced by their small modularity. Furthermore, the poor performance is associated with larger radius and diameter of the graph contrary to well-performing ones (shown in **Table 1**), which are associated with smaller radii and diameter values.

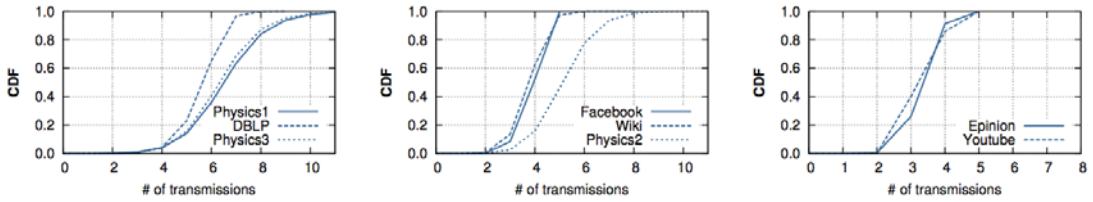

Fig. 5. The CDF of the number of transmissions per node in different social graphs using BFS routing.

4.2.2 Shortest Path Based Routing on Social Graphs

In this section we evaluate the BFS and Dijkstra routing algorithms. Mainly, we evaluate the expected number of transmissions between two randomly selected source and destination nodes. As indicated in the Random walk-based routing scenario, we define the cost of routing over graphs as the normalized expected number of transmissions, which is the average number of times that a node in the graph transmits a packet for a single routing session from a given source to a given destination. To evaluate the routing performance in the different graphs of **Table 1**, the experiment is repeated for 100,000 times for a random pair of source and destination in each graph. In **Figs. 5** and **6**, we show the CDFs of the number of transmissions per node in several social graphs using BFS and Dijkstra routing. These graphs illustrate the relation between the number of transmissions, the structure, and connectivity of the graph.

These results coincide with our finding regarding to the random walk-based routing in social graphs, where the cost of routing over the different graphs is strongly related to the structure and connectivity of the underlying graph and loosely affected by the applied routing algorithm.

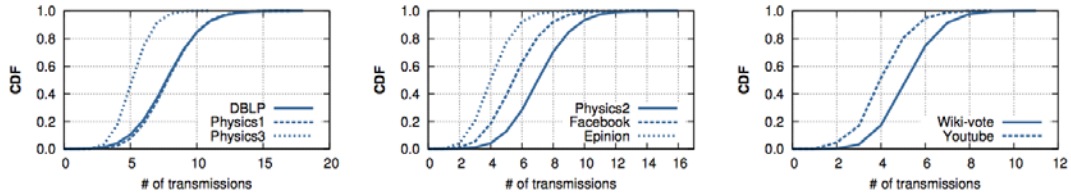

Fig. 6. The CDF of the number of transmissions per node when using Dijkstra routing algorithm.

Based on **Figs. 7** and **8**, the studied social graphs can be classified in the case of shortest path based routing into two categories: well-performing graphs (Epinion, Youtube, Wiki-vote, and Facebook) and poorly performing graphs (Physics-1 to 3 and DBLP). However, we observe that random walk-based routing outperforms the shortest path based routing in the number of transmissions. This is due to the deterministic nature of the underlying shortest path based routing.

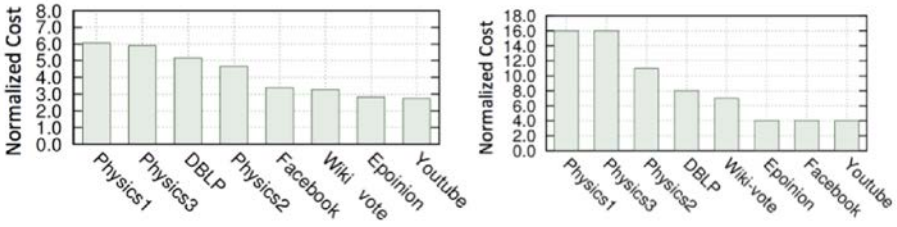

Fig. 7. The normalized expected number and maximum number of transmissions per node by using BFS routing on top of different social graphs.

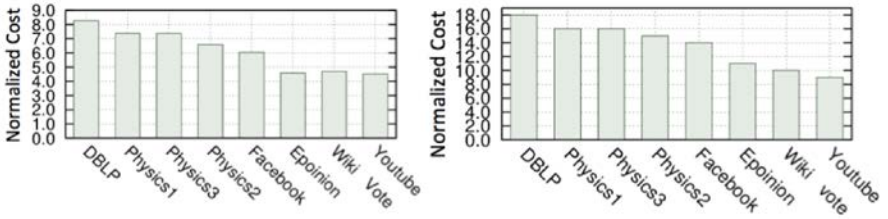

Fig. 8. The normalized expected number and maximum number of transmissions per node by using shortest path routing.

4.3 Performance with Social Collaboration

Using the collaboration model highlighted in section 3, we measure the cost of routing using both random walk-based and shortest path based techniques. Results provided here empirically complement the theoretical measures provided in section 3.2.

4.3.1 Collaboration in Random Walk-based Routing

We measure the performance of the routing protocol, with the same settings as above, when considering probabilistically collaborative nodes in the graph. We uniformly at random

sample subsets of the nodes $V_p \subset V$ as the set of probabilistically collaborative users, with the remaining nodes in the graph as totally collaborative (altruistic). As explained earlier, each probabilistic node v_i follows the protocol with probability α_i, which is uniformly selected in the range of 0.1 to 1, or drop the routing request with probability (1-α_i) (we trim the distribution from 0 to 0.1 for that the existence of very adversarial nodes may block traffic entirely due to the lack of multi-path). The results of the performance of the protocol on the different social graphs are shown in **Fig. 9**. By considering β as the percent of rational users, we consider different values of β (i.e., from 0 to 0.8 with 0.2 steps) where β=0 in each graph represents the performance of the corresponding social graph in **Fig. 10**. In brief, we make the following observations on the different experiments.

While the performance of the different social graphs initially differs greatly, as evidenced by the first experiment, the impact of the increasing percent of rational nodes is not linear but rather depends on the underlying social graph. For example, social graphs with strong community structure have rather fairly regular behavior (**Fig. 10(a)** to **Fig. 10(d)**). However, we observe that even with these social graphs, relatively large β dramatically increases the cost of routing. As shown in **Fig. 10(a)**, while the cost increment when moving from β=0.4 to β=0.6 is about 26%, the increment for moving from β=0.6 to β=0.8 is more than 100%. To understand this behavior, we list the rational nodes, and map them to their degrees. We observe that, while in the first case - where less impact is made on the performance in random walk-based routing due to lack of collaborative of some nodes - the degree is fairly distributed, in the second case - where higher impact is observed - some high degree nodes, with small α blocks flows between communities, and thus dramatically increases the cost of routing. In such graphs, the cost is exponential to β.

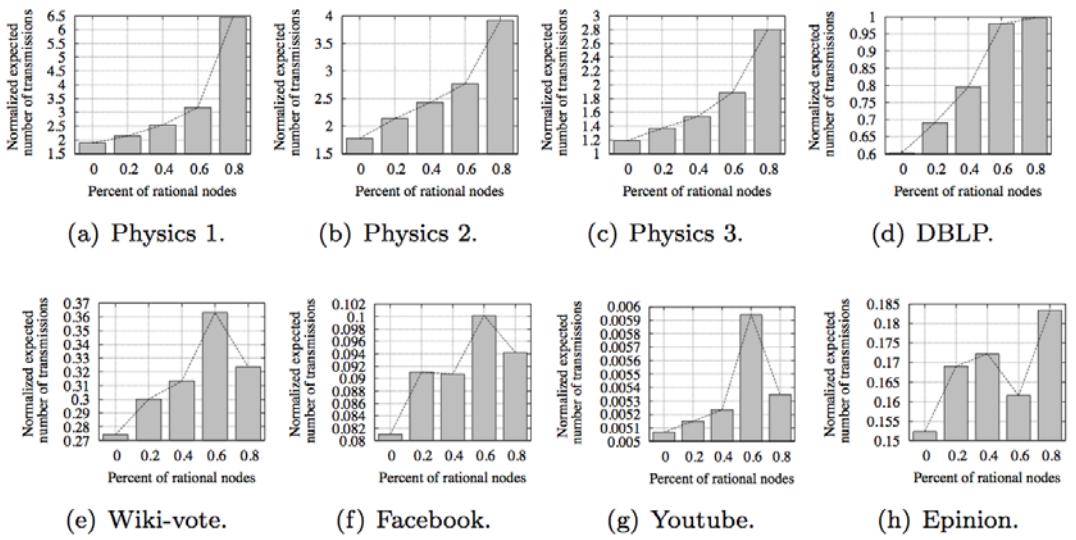

Fig. 9. The normalized expected number of transmissions using random walk-based routing.

On the other hand, the behavior of the initially well-performing graphs is in part harder to anticipate. In general, the routing protocol performs well, even when considering larger values of β in these graphs, though, like the previous case of poorly performing graphs it may have some odd behavior when high degree, intra-communities' nodes behave rationally, with small

β. This behavior happens to be the case in **Fig. 10(e)** at β=0.6, in **Fig. 10(f)** at β=0.2, 0.4, 0.6, in **Fig. 10(g)** at β=0.6, and in **Fig. 10(h)** at β=0.2, 0.4, 0.8.)

While the well-performing graphs are more sensitive where any node can be of importance to the random routing on such graphs, as evidenced by the existence of many high degree nodes (see **Fig. 4(b)**), the performance is still reasonable and within the theoretically acceptable bounds [35].

4.3.2 Collaboration in Shortest Path-based Routing

Considering the existence of probabilistically collaborative nodes in the graphs, we measured the performance of the shortest path based routing when a random sample subset of the nodes $V_p \subset V$ are probabilistically collaborative and the remaining are collaborative. As explained earlier, each probabilistic node v_i follows the protocol with probability α_i, which is uniformly selected in the range of 0.1 to 1, or drops the routing request with probability (1-α_i). As in the case of random walk-based routing, the results of the performance of the protocol with underlying shortest path based routing algorithm on the different social graphs are shown in **Figs. 10** and **11**. However, due to the deterministic nature of the shortest path based routing, we consider the failure rather than the cost of routing as the evaluation criterion, for different β values. The results show that, in terms of packet delivery rate, well-performing graphs (Epinion, Youtube, Wiki-vote, and Facebook) are less sensitive to selfish users compared to poorly performing graphs (Physics-1 to 3 and DBLP), which are dramatically affected by the rational users.

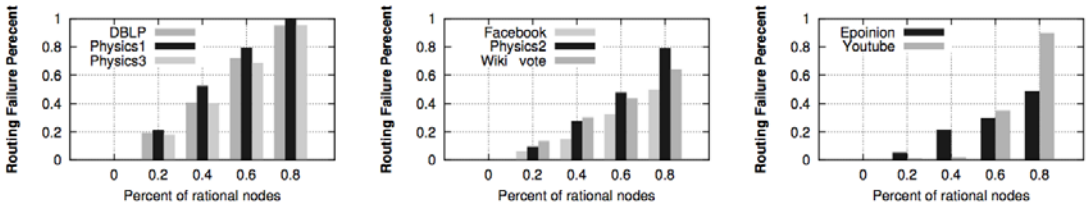

Fig. 10. BFS failure rate, expressed as the routing failure percent per β value (rational users) in each social graph.

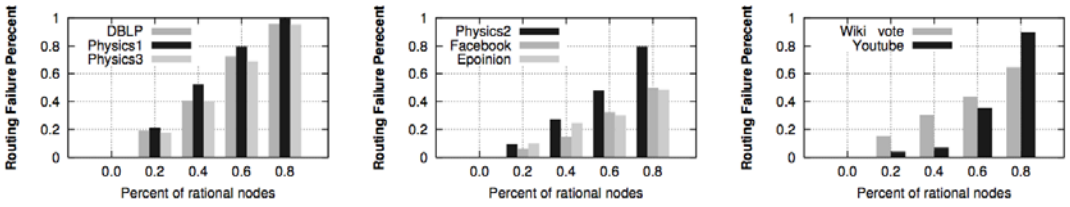

Fig. 11. BFS failure rate, expressed as the routing failure percent per β value (rational users) in each social graph.

One implication of these findings on applications is that, since the collaboration of high-degree nodes is more important to the overall performance of the network, investigating providing incentives for such nodes to be always collaborative to improve the performance is an interesting issue. For example, in delay tolerant networks (DTN) routing built on social networks [5], [36], [37]; it is assumed that all nodes are collaborative. Since it is not always the case, it will be interesting to deploy these observations and build incentives for collaboration, especially for those critical nodes, in that context.

We notice that the characterization of users into malicious and benign in away corresponds to assigning two extreme values to their behavior: 0 or 1. This is, a malicious user will always not forward on behalf of other users, whereas the benign user will always forward on behalf of other users. Modeling the behavior of users as probability also captures malicious users by adjusting the correct parameter set: 0 and 1. Furthermore, in the context of this study, and for the purpose of the applications discussed in the paper for routing, a malicious user would drop all messages (by adjusting collaboration parameters to 0), thus evaluating such scenario is trivial, and we do not consider it. However, one should keep in mind that such case is explicitly included in our modeling of user behavior.

5. Related Work

There has been a number of papers on the use of social networks for building communication and security systems, studying the performance of such designs on top of social networks, and analyzing the assumptions used in these designs as well. The closest to this study is the work in [38], where nodes are basically assumed to have some selfish behavior in each and every one of them, which follows some distribution (e.g., uniform, normal, or geometric). The major difference between our work and the work in [38] is actually twofold. First, while [38] considers traces of encounter-like wireless networks, we consider traces of static social graphs. While in the general sense both types of traces, static and encounter-based, could be of potential use to routing, we believe that static traces are more favorable to the assumption of trust which most routing protocols weigh a big value on to demonstrate effectiveness [39]. Second, and more important, the conclusions in this paper are at contradiction with the findings in [38]. In particular, whereas it is shown in [38] that selfishness does not much affect the behavior of the routing algorithm due to the multi-path characteristics of the underlying connections and links among nodes in the graph, we show strong evidence that the lack of collaboration by a few nodes with a particular characteristic (e.g., degree distribution) in static social graphs could greatly affect the effectiveness of the routing protocol built on top of the social network (see details in section 4). In total, while our work brings conclusions contradicting with the prior work in [38], it can be considered as an effort in the same direction to understanding collaboration in settings where social networks are used for improving routing in networked systems.

Systems built on top of social networks include Sybil defenses, such as the work in [15], [16], [40], [41]. Most of these defenses weigh trust in the social graph, and an algorithmic property that makes the operation of the defense on top of social network at reasonable cost. Routing and information sharing using social networks is explored in [4], [5], [6], [8], [20], [21] for different settings, where it is shown that connectivity in social graphs can be of benefit in disconnected networks. However, in all of those works collaboration is assumed to be categorical, and there were no attempts in the literature to characterize collaboration and understand how it impacts the cost of operating such routing and sharing algorithms and applications on top of social networks.

In line with our work to understand the underlying assumptions about applications built on top of social networks, there has been several attempts to characterize properties of such networks. The properties used for building Sybil defenses are studied in [33] where it has been shown that widely accepted properties of social networks, such as the mixing time, do not hold in a wide variety of social networks (results are confirmed in [42] and [43]), and trust is challenged in [39] where it has been shown that the guarantees of operating such defenses on top of social networks can be greatly improved by incorporating differential trust. It is worth noting that both the context and findings in these papers and the current paper, are different. In particular, while we show that a small number of nodes could thwart the utility of random

walk-based routing in social networks, the results in the former papers demonstrate that some social graphs are fast mixing whereas some are slower mixing, and this affects the systems built on top of these social graphs.

Similar to our work in essence is the work in [44] and [26]. In [44], Feng et al. studied social incentivies for enabling cooperation in spectrum sensing in distributed cognitive radio networks. In [26], Wang et al. studied the performance of information propagation in delay tolerant networks that utilize an underlying mobile social network fabric. In [45], [46] and [47], Le *et al*. studied how social behavior affects caching in DTN. In [48], Chung et al. demonestrated how social behavior can be used to improve communication and node discovery in mobile ad hoc networks. In [49], Wei et al. designed an efficient Sybil defense in general networks utilizing the trust fabric in social networks, and in [50] Liu et al demonestrated how to use temporal dynamics for attacking social network-based Sybil defenses.

5. Conclusions and Future Work

In this paper, by classifying nodes in social graphs into collaborative and probabilistically collaborative, we studied the impact of collaboration in social networks on the performance of information routing techniques, including random walk based routing, shortest-path based routings, all of which are built on top of social networks. Even without the classification part, we experimentally demonstrated that the cost of such protocols on top of some of real-world graphs is large while it is reasonable on others. We further show that some networks are very well-performing and meet the potential of such applications, whereas other networks are quite sensitive to the behavior of users: the lack of collaboration being small fraction of nodes in the graph would influence the cost associated with such algorithms when initiated from the majority of nodes in the graph. This observation is hypothesized to be related to the underlying structure and the characteristic of the non-collaborating node. The previous section identified three clusters for QoS classes and features to build up classification rules through unsupervised learning. In this section, the accuracy of the classification rules is evaluated experimentally. For classification, we chose the K-nearest neighbor (KNN) algorithm. Experimental results are compared with the minimum mean distance (MMD) classifier. Exploring theoretical models to characterize the performance of routing algorithms under behavior as parameters would be the work to be considered in the near future. This will benefit from rich literature, e.g., [51], [52].

References

[1] A. Mohaisen, T. AbuHmed, T. Zhu and M. Mohaisen, “Collaboration in social networks based contents dissemination,” in *Proc. of IEEE ICC*, pp. 1-6, 2012. Article (CrossRef Link).
[2] Ebizmba, www.ebizmba.com, 2009.
[3] Facebook, www.facebook.com, 2009.
[4] G. Bigwood and T. Henderson, “Social dtn routing,” in *Proc. of the ACM CoNEXT Conference*, ACM, pp. 1-2, 2008. Article (CrossRef Link).
[5] E. M. Daly and M. Haahr, “Social network analysis for routing in disconnected delay-tolerant MANETs,” in *Proc. of the 8th ACM International Symposium on Mobile ad hoc networking and computing*, New York, NY, USA, pp. 32-40, 2007. Article (CrossRef Link).
[6] J. Davitz, J. Yu, S. Basu, D. Gutelius and A. Harris, “ilink: Search and routing in social networks,” in *Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, pp. 931-940, 2007. Article (CrossRef Link).

[7] I. Mabrouki, X. Lagrange and G. Froc, "Random walk based routing protocol for wireless sensor networks," in *Proc. of ValueTools, ICST*, pp. 1-10, 2007. Article (CrossRef Link).
[8] S. Marti, P. Ganesan and H. Garcia-Molina, "Dht routing using social links," in *G. M. Voelker, S. Shenker (Eds.), IPTPS, Vol. 3279 of Lecture Notes in Computer Science, Springer*, pp. 100-111, 2004. Article (CrossRef Link).
[9] S. Lai and B. Ravindran, "On distributed time-dependent shortest paths over duty-cycled wireless sensor networks," in *Proc. of the 29th conference on Information communications*, Piscataway, NJ, USA, pp. 1685-1693, 2010. Article (CrossRef Link).
[10] S. Kwon and N. B. Shroff, "Analysis of shortest path routing for large multi-hop wireless networks," *IEEE/ACM Trans. Netw.*, vol. 17, no. 3, pp. 857-869, 2009. Article (CrossRef Link).
[11] I. Parris and T. Henderson, "Privacy-enhanced social-network routing," *Computer Communications*, vol. 35, no. 1, pp. 62-74, 2012. Article (CrossRef Link).
[12] S. M. A. Abbas, J. A. Pouwelse, D. H. J. Epema and H. J. Sips, "A gossip-based distributed social networking system," in *Proc. of the 18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises*, Washington, DC, USA, pp. 93-98, 2009. Article (CrossRef Link).
[13] Y. Fernandess and D. Malkhi, "On spreading recommendations via social gossip", in *Proc. of the 20th ACM Annual Symposium on Parallelism in Algorithms and Architectures*, New York, NY, USA, pp. 91-97, 2008. Article (CrossRef Link).
[14] A. Chaintreau, P. Fraigniaud and E. Lebhar, "Opportunistic spatial gossip over mobile social networks," in *Proc. of the First Workshop on Online Social Networks*, New York, NY, USA, pp. 73-78, 2008. Article (CrossRef Link)
[15] H. Yu, P. B. Gibbons and M. Kaminsky, "Toward an optimal social network defense against Sybil attacks," in *Proc. of ACM PODC*, pp. 376-377, 2007. Article (CrossRef Link).
[16] H. Yu, M. Kaminsky, P. B. Gibbons and A. Flaxman, "Sybilguard: defending against Sybil attacks via social networks," in *Proc. of SIGCOMM*, pp. 267-278, 2006. Article (CrossRef Link).
[17] G. Danezis, C. Lesniewski-laas, M. F. Kaashoek and R. Anderson, "Sybil-resistant dht routing," in *Proc. of ESORICS*, pp. 305-318, 2005. Article (CrossRef Link).
[18] S. Nagaraja, "Anonymity in the wild: Mixes on unstructured networks," *N. Borisov, P. Golle (Eds.), Privacy Enhancing Technologies, Vol. 4776 of Lecture Notes in Computer Science, Springer*, pp. 254-271, 2007. Article (CrossRef Link).
[19] A. Mohaisen and Y. Kim, "Dynamix: Anonymity on dynamic social structures," in *Proc. of ACM ASIACCS*, pp. 1-6, 2013. Article (CrossRef Link).
[20] X. Liang, M. Barua, R. Lu, X. Lin and X. S. Shen, "Healthshare: Achieving secure and privacy-preserving health information sharing through health social networks," *Computer Communications*, vol. 35, no. 15, pp. 1910-1920, 2012. Article (CrossRef Link).
[21] A. Passarella, R. I. M. Dunbar, M. Conti and F. Pezzoni, "Ego network models for future internet social networking environments," *Computer Communications*, vol. 35, no. 18, pp. 2201-2217, 2012. Article (CrossRef Link).
[22] M. Kwiatkowska, G. Norman and D. Parker, "Analysis of a gossip protocol in PRISM," *ACM SIGMETRICS Performance Evaluation Review*, vol. 36, no. 3, pp. 17-22, 2008. Article (CrossRef Link).
[23] Gossip Protocol, http://www.prismmodelchecker.org/casestudies/gossip.php.
[24] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins, "Geographic routing in social networks" *PNAS*, vol. 102, no. 33, pp. 11623-11628, 2005. Article (CrossRef Link).
[25] G. Sikander *et al.*, "A survey of cluster-based routing schemes for wireless sensor networks," *Smart Computing Review*, vol. 3, no. 4, pp. 261-275, August, 2013. Article (CrossRef Link).
[26] Z. Wang, S. Deng, H. Huang and Y. Wu, "Performance analysis of information propagation in DTN-like scale-free mobile social network," *KSII Transactions on Internet and Information Systems*, vol. 8, no. 11, pp. 3984-3996, November, 2014. Article (CrossRef Link).
[27] T. H. Cormen, C. Stein, R. L. Rivest and C. E. Leiserson, *Introduction to Algorithms*, 2nd Edition, McGraw-Hill Higher Education, 2001.

[28] J. Leskovec, J. Kleinberg and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proc. of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, New York, NY, USA, pp. 177-187, 2005. Article (CrossRef Link).
[29] J. Leskovec, D. P. Huttenlocher and J. M. Kleinberg, "Predicting positive and negative links in online social networks," in *Proc. of WWW*, pp. 641-650, 2010. Article (CrossRef Link).
[30] M. Ley, "The DBLP computer science bibliography: Evolution, research issues, perspectives," in *Proc. of String Processing and Information Retrieval, Springer*, pp. 481-486, 2009. Article (CrossRef Link).
[31] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy and B. Y. Zhao, "User interactions in social networks and their implications," in *Proc. of the 4th ACM European conference on Computer systems*, pp. 205- 218, 2009. Article (CrossRef Link).
[32] A. Mislove, M. Marcon, P. K. Gummadi, P. Druschel and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. of Internet Measurement Conference*, pp. 29-42, 2007. Article (CrossRef Link).
[33] A. Mohaisen, A. Yun, Y. Kim, "Measuring the mixing time of social graphs," in *Proc. of the ACM IMC*, pp. 383-389, 2010. Article (CrossRef Link).
[34] B. Viswanath, A. Post, K. P. Gummadi and A. Mislove, "An analysis of social network-based Sybil defenses," in *Proc. of SIGCOMM*, pp. 1-12, 2010. Article (CrossRef Link).
[35] Y. Li, D. Zou, Y. Liu, Z. Zhou and Y. Li, "On the efficiency of random walk routing in multihop wireless network," in *Proc. of GLOBECOM*, pp. 1-6, 2009. Article (CrossRef Link).
[36] Q. Li, S. Zhu and G. Cao, "Routing in socially selfish delay tolerant networks," in *Proc. of the 29th conference on Information communications*, Piscataway, NJ, USA, pp. 857-865, 2010. Article (CrossRef Link).
[37] R. J. Clark, E. Zasoski, J. Olson, M. H. Ammar and E. W. Zegura, "D-book: A mobile social networking application for delay tolerant networks," in *Proc. of Challenged Networks*, pp. 113-116, 2008. Article (CrossRef Link).
[38] P. Hui, K. Xu, V. Li, J. Crowcroft, V. Latora and P. Lio, "Selfishness, altruism and message spreading in mobile social networks," in *Proc. of First IEEE International Workshop on Network Science for Communication Networks*, pp. 1-7, 2009. Article (CrossRef Link).
[39] A. Mohaisen, N. Hopper, Y. Kim, "Keep your friends close: Incorporating trust into social network-based Sybil defenses," in *Proc. of INFOCOM*, Piscataway, NJ, USA, pp. 336-340, 2011. Article (CrossRef Link).
[40] G. Danezis and P. Mittal, "Sybilinfer: Detecting Sybil nodes using social networks," in *Proc. of NDSS*, pp. 1-16, 2009.
[41] C. Lesniewski-Laas, "A Sybil-proof one-hop DHT," in *Proc. of the 1st workshop on Social network systems*, pp. 19-24, 2008. Article (CrossRef Link).
[42] L. Alvisi, U. Austin, A. Clement, A. Epasto, U. Sapienza, S. Lattanzi and A. Panconesi, "Sok: The evolution of Sybil defense via social networks," in *Proc. of Oakland*, pp. 1-14, 2013. Article (CrossRef Link).
[43] A. M. Kakhki, C. Kliman-Silver and A. Mislove, "Iolaus: Securing online content rating systems," in *Proc. of the Twenty-Second International World Wide Web Conference*, Rio de Janeiro, Brazil, pp. 1-10, 2013. Article (CrossRef Link).
[44] J. Feng, G. Lu and X. Min, "Social incentives for cooperative spectrum sensing in distributed cognitive radio networks," *KSII Transactions on Internet and Information Systems*, vol. 8, no. 2, pp. 355-370, February, 2014. Article (CrossRef Link).
[45] T. Le, Y. Lu and M. Gerla, "Social caching and content retrieval in disruption tolerant networks (DTNs)," in *Proc. of ICNC*, pp. 905-910, 2015. Article (CrossRef Link).
[46] T. Le, H. Kalantarian and M. Gerla, "A novel social contact graph based routing strategy for delay tolerant networks," in *Proc. of IWCMC*, pp. 13-18, 2015. Article (CrossRef Link).
[47] T. Le, H. Kalantarian and M. Gerla, "Socially-aware content retrieval using random walks in disruption tolerant networks," in *Proc. of WOWMOM*, pp. 1-6, 2015. Article (CrossRef Link).

[48] E. Chung, J. Joy and M. Gerla, "DiscoverFriends: Secure social network communication in mobile ad hoc networks," in *Proc. of IWCMC*, pp. 7-12, 2015. Article (CrossRef Link).
[49] W. Wei, F. Xu, C. Tan Qun Li, "SybilDefender: A defense mechanism for sybil attacks in large social networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 12, pp. 2492-2502, 2013. Article (CrossRef Link).
[50] C. Liu, P. Gao, M. K. Wright and P. Mittal, "Exploiting temporal dynamics in Sybil defenses," in *Proc. of ACM Conference on Computer and Communications Security*, pp. 805-816, 2015. Article (CrossRef Link).
[51] R. Beraldi, "Random walk with long jumps for wireless ad hoc networks," *Ad Hoc Networks*, vol. 7, no. 2, pp.294-306, 2009. Article (CrossRef Link).
[52] G. Froc, I. Mabrouki and X. Lagrange, "Design and performance of wireless data gathering networks based on unicast random walk routing," *IEEE/ACM Trans. Networks*, vol. 17, no. 4, pp. 1214-1227, February, 2009. Article (CrossRef Link).

Manar Mohaisen received his M.S. degree in communications and signal processing from the University of Nice-Sophia Antiplois, France, in 2005 and Ph.D. from Inha University, Korea, in 2010, both in communications engineering. From 2001 to 2004, he was with the Palestinian Telecom. Co., where he was a cell planning engineer. Since Sept. 2010, he is with the Department of EEC Engineering, KoreaTech, Korea, where he is an assistant professor. His research interests include 3GPP LTE/-A systems, MIMO detection and precoding and social networks.

Aziz Mohaisen obtained his M.S. and Ph.D. degrees in Computer Science from the University of Minnesota, both in 2012. He is currently an Assistant Professor at the Computer Science and Engineering Department of the State University of New York at Buffalo. From 2012 to 2015, he was a Senior Research Scientist at Verisign Labs. Previously he was a Member of Engineering Staff at the Electronics and Telecommunication Research Institute, a large research and development institute in South Korea. His research interests are in the areas of networked systems, systems security, data privacy, and measurements. He published more than 80 peer-reviewed publications, and is a Senior Member of IEEE and a member of ACM.