This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2010 proceedings

Locating Experts via Online Social Networks

Kuang Xu, Jing Xie, Victor O.K. Li
Department of Electrical and Electronic Engineering
The University of Hong Kong, Pokfulam, Hong Kong, China

***Abstract*—Online social networking systems provide indirect access to a large number of people connected by multi-step chains of acquaintances, and plays an important role in the referrals for human information flow. In this paper, from a networking point of view, we study the problem of locating experts for relevant information via online social networks. We model the action of forwarding a question with random walk, adjusted by a node's awareness of the potential expertise of his immediate neighbours. Using the model we derive analytical expressions of the performance metrics of a referral session in terms of the nodes' awareness level of their neighbours and the percentage of nodes that may have answers to the posed question. We also utilize several real online social networks to study the modeled question-forwarding strategy, and find that the simulation results validate our analyses.**

I. Introduction

Recent years have seen the flourishing of numerous online social networks (OSN), spawning a surge of innovations and opportunities. Systems such as Facebook and MySpace, immense cyber-communities built around friendships in the off-line world, have emerged as top 10 sites globally in terms of traffic. Their success can be ascribed in part to the notable phenomena – *small world* [1] [2] – that maintains people in the real world are sparsely connected while the degree of separation of personal relationships [3] is relatively small. Another important factor is that the amount and kinds of information a person knows or is able to obtain from the public are limited. Compared with seeking a piece of information (e.g. travel tips, gift idea) directly through Web search engines which may not index the most relevant answers, friends or human experts often give more specific recommendations. In addition, a person tends to value answers from people he trusts or people trusted by people he trusts [4], rather than those of complete strangers. Thus one can utilize the underlying social network structure of OSNs to find relevant information [5] via his friends, his friends' friends, and so forth. In this paper, from a networking point of view, we study the problem of locating the right persons who have answers to specific questions via OSN. OSN-based information search has received attention in both research [6] and actual applications [7] [8] [9] [10], and we refer to it as online social search (OSS). Here, before jumping into the technical details of this work, we introduce the general operations of OSS, which serves as application scenarios for our study.

An OSN-based referral system maintains all its users' profiles (which could be updated from Facebook, including contacts, interest, expertise, etc.) and each user has a backend agent which processes the information queries. When a user poses a question (via mobile phone or Web portal), the system routes the question to his (selected) contacts (through text, email, or instant message). An agent receiving a question decides whether it suits its user according to the user's profile and, if not, forwards the question to the next-hop contacts' agents. If a potential match is found, the agent alerts its user to respond to that question, and if the user does not know the answer, the agent continues forwarding the question in the same way. The question is passed on in the underlying social network until it either gets an answer, or exceeds a hop-limit specified by its owner. Finally, the questioner may be presented with a great number of potential respondents.

Based on the above user scenario, we ask a question, "How does a referral session perform when a person looks for advice to a specific question?" A referral session refers to the process from when a question is injected into the system until it finds an answer or becomes obsolete, and the performance here concerns two aspects, namely, *success rate* and *referral cost*. The former represents the possibility of a question being forwarded to at least one person that has an answer to it, while the latter is the number of persons this question has visited upon termination of a referral session. Referral cost is important since it is related not only to the network resource consumption but also to the questioner's privacy. In other words, the higher the referral cost, the more people knows about the questioner's query.

Forwarding a question to an appropriate expert is non-trivial, as one is confronted with the trade-off between forwarding the question to as many neighbours as possible (e.g. flooding) at each step, and hence straining the willingness of possible responders [11], and forwarding them to a more compact set of neighbours, thus missing an appropriate expert. Considering the above trade-off, in this paper, instead of having a referral agent flood questions to all its neighbours or send them to a predetermined set of neighbours, we utilize *random walk* to model the question forwarding strategy through OSN (Section II). Since, as the user scenario depicts, an agent maintains its user's profile and personal social network, we equip the nodes in our model with the intelligence of awareness, assuming every node is aware of the potential expertise of his neighbours. Consequently, the action of forwarding a question is affected by this social context. Based on the model we derive analytical expressions of the performance metrics of a referral session (i.e. success rate and referral cost) in terms of nodes' awareness level (r) of their neighbours and the percentage of nodes (e) that may have appropriate answers to the issued question, with homogenous settings of the number

978-1-4244-6404-3/10/$26.00 ©2010 IEEE

of a node's neighbours (Section III). The analytical derivations are verified by simulation. Based on the analytical result, appropriate random walk parameters can be chosen to achieve a user's expectation on a referral session. We also apply the modeled question-forwarding strategy to the crawled data of a set of real OSNs [12] with various settings (Section IV). The simulation result shows that the performance of locating experts improves as r or e increases while the improvement is less significant when r or e becomes higher, and the underlying network connectivity has a positive relationship with the system performance. Finally, we conclude our study with suggestions for future work (Section V).

II. Modeling and Analysis

In this section, we model a node's action of forwarding a question, followed by analyzing the performance of a referral session.

A. Strategy modeling

We consider an OSN as an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes (OSN users) and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges (social ties) in the network. Each edge means one-hop question-forwarding is possible between the pair of nodes. Let $n = |\mathcal{V}|$ be the number of users in the system. We also denote by $\mathcal{N}_u \subseteq \mathcal{V}$ the set of neighbors of Node u, and $d_u = |\mathcal{N}_u|$, the number of users in this set. Since a node maintains its local social network, we equip the nodes in our model with the intelligence of awareness, assuming every node is aware of the potential expertise of his neighbors. We denote by $S_u(v, i)$ a node u's awareness of its neighbor v's expertise, with respect to Question i. $S_u(v, i)$ takes one of three possible values $\{-1, 0, 1\}$, such that

$$S_u(v,i) \triangleq \begin{cases} 1, & u \text{ knows } v \text{ is an expert on Question } i, \\ -1, & u \text{ knows } v \text{ is not an expert on Question } i, \\ 0, & \text{otherwise.} \end{cases}$$

The above assumption studies the simplest case that classifies nodes' awareness into three types. In other words, for every question i, we divide a node's neighbors into three possible sets. Nodes in the first set are regarded as experts on Question i, and those in the second set are considered as not holding relevant answers to Question i, and the rest of the neighbors are those that the node is uncertain about whether they are expert on i.

DEFINITION 1. *The **awareness level** of neighbors' expertise in Question i is defined as*

$$r_i \triangleq \frac{\delta_i}{\sum_{u \in \mathcal{V}} d_u}, \quad \text{where}$$

$\delta_i = |\{e_{uv} | v \in \mathcal{N}_u \text{ and } S_u(v,i) \neq 0, \text{ for all } u \in \mathcal{V}\}|$.

e_{uv} in the above definition refers to the directed edge from Node u to its neighbor v. We consider directed edge since, unlike a social tie between two people which represents the fact that the two people know each other, awareness here describes to what extent a person unilaterally feels whether he knows the expertise of another person on a particular question, and may not be symmetric. Thus r_i refers to the percentage of the directed edges among nodes that satisfy $S_u(v, i) \neq 0$. In other words, r_i is the probability that Node u knows the potential expertise of Node v, either "v is an expert on Question i" or "v is a layman (not an expert) on Question i", for each $v \in \mathcal{N}_u$.

DEFINITION 2. *The **expert density** on Question i is defined as*

$$e_i \triangleq \frac{l_i}{n},$$

where $l_i > 0$ refers to the number of people that have relevant answers to Question i in the network. In reality, a person posing a question may receive answers from multiple experts on this question. However, different experts on the same question may have different opinions (though all of them may be reasonable under different circumstances), and their answers may lead the questioner to different choices. Consequently, a person is often faced with information overload [13] and he can not effectively filter out the pieces of information that are most appropriate for him. Since the purpose of this paper is to study the performance of a referral session (i.e. success rate and referral cost of forwarding questions) from a networking point of view, we consider the homogeneous case that assumes advice from experts on the same question have the same effect on the questioner (all of which are considered successful), and we leave the study of the dynamics of information filtering as future work.

Algorithm 1 summarizes the question forwarding strategy from a node's perspective. If a node that receives Question i has expertise in i, it responds to the node that poses i with an answer, and this referral session is considered successful. If not, it checks to see whether some of its neighbors are potential experts on i and, if so, forwards i to a randomly selected expert neighbor. If there are no expert neighbors on i, it forwards i randomly to one of its neighbors excluding those who are are not experts on i, and if there are no neighbors of this category, the referral session is considered failed. Each question is also posed with a hop limit, and considered failed if this hop limit is exceeded.

B. Analysis

We first analyse the success rate of a referral session at a single step. We denote by $q_{i,u}$ the probability that a referral to an expert on Question i is satisfied at one step from Node u. To simplify the notation, we use q_i and d instead of $q_{i,u}$ and d_u, respectively, in this section.

LEMMA 1.

$$q_i = 1 - (1 - e_i)(1 - r_i e_i)^d \tag{1}$$

PROOF. There are two cases:

Case A: There is at least one node $v \in \mathcal{N}_u$ that satisfies $S_u(v, i) = 1$. It means that Node u knows that at least one of his neighbors has expertise in Question i. In this case, $q_i = 1$.

Case B: For all $v \in \mathcal{N}_u$, $S_u(v, i) = -1$ *or* 0. We let $\tau = |\{w | S_u(w, i) = -1, w \in \mathcal{N}_u\}|$. It means Node u knows τ of his neighbors do not have appropriate answers to

Algorithm 1 Question forwarding strategy (Question i is posed by Node u)

if Node v is an expert on i **then**
  respond u with answer, return with success
else
  check $\mathcal{N}_v$
  if $|\{w|S_v(w,i)=1\}|>0$ **then**
    forward i to $w \in \{w|S_v(w,i)=1\}$ with probability $\frac{1}{|\{w|S_v(w,i)=1\}|}$
  else if $|\{w|S_v(w,i)=0\}|>0$ **then**
    forward i to $w \in \{w|S_v(w,i)=0\}$ with probability $\frac{1}{|\{w|S_v(w,i)=0\}|}$
  else
    return with fail
  end if
end if

Question i. Accordingly, Node u randomly selects a neighbor $v \in \{w|S_u(w,i)=0\}$. In this case, the likelihood that the randomly selected node has an appropriate answer to Question i is thus $q_i = e_i$.[1]

Denote by $p(A)$ the probability of *Scenario A* taking place. The probability of the occurrence of *Scenario B*, $p(B)$, is $1-p(A)$. Since $|\mathcal{N}_u| = d$, thus

$$p(A) = 1-(1-P(S_u(v,i)=1))^d$$

where $P(S_u(v,i)=1)$ denotes the probability that $S_u(v,i)=1$, and we have made the assumption that the expertise of the d neighbors are independent. Note that $S_u(v,i)=1$ occurs if and only if both of the following two conditions are satisfied:

1) $S_u(v,i) \neq 0$ (with probability r_i);
2) Node v is an expert on Question i (with probability e_i).

Hence, the probability that $S_u(v,i)=1$ is

$$P(S_u(v,i)=1) = r_i \cdot e_i.$$

Therefore,

$$p(A) = 1-(1-r_i \cdot e_i)^d,$$

and,

$$\begin{aligned} q_i &= p(A)\cdot 1 + p(B)\cdot e_i \\ &= 1-(1-e_i)(1-r_i\cdot e_i)^d. \end{aligned}$$

We plot the single step success rate q_i against a range of r_i and e_i according to Eqn. (1) in Fig. 1. We identify several properties:

- *Property 1:* $\forall i,\ q_i = e_i$, if $r_i = 0$. $r_i = 0$ means there is no information of neighbors' expertises for Question i, and the corresponding performance is the same as a standard random walk.

[1]Strictly speaking, $q_i = \frac{l_i}{n-\tau}$, since Node u clearly knows that τ nodes are not experts on i and excludes them from the random selection process. However, it is easy to see that $\tau \ll n$ in OSN; thus we can obtain an approximation $q_i \approx \frac{l_i}{n} = e_i$.

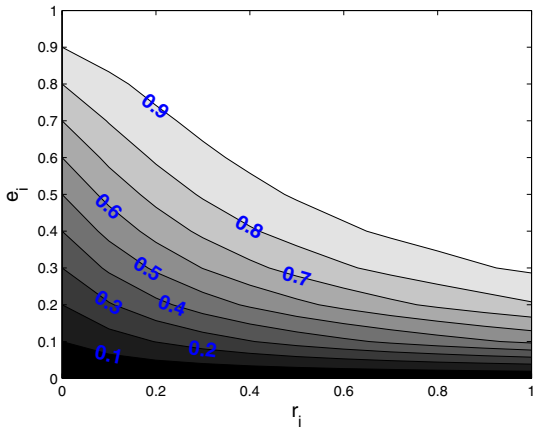


Fig. 1. Variation of q_i against r_i and e_i.

- *Property 2:* $\forall i,\ q_i$ is an increasing function of r_i and e_i, respectively. That is, $\frac{\partial q_i}{\partial r_i} > 0$ and $\frac{\partial q_i}{\partial e_i} > 0$. It coincides with the intuition that the per step success likelihood of a referral session tends to be higher for a question with more relevant experts, or when more information about neighbors is available.
- *Property 3:* The growing speed of q_i decreases as the parameters increase. That is, $\frac{\partial^2 q_i}{\partial {r_i}^2} < 0$ and $\frac{\partial^2 q_i}{\partial {e_i}^2} < 0$. This says that enhancing q_i becomes less effective as r_i and e_i grow.

Statistical analogy between random walk and uniform sampling is constructed through two perspectives. The study in [14] demonstrates that random walk and uniform sampling solve the coupon collection problem in the same order of magnitude. In addition, they arrive at the analogy through the illustration of a Chernoff bound on a sequence of Bernoulli trials. The comparisons also reveal that the precision of approximating random walk with uniform sampling depends on the underlying network connectivity. Since the typical OSN topologies exhibit properties which guarantee a large average degree as revealed in [12], it is appropriate to approximate the random walk with uniform sampling in our context. We consider the success rate of a referral session as the probability that an expert on the posed question is found within T steps of forwarding. Denote by j the j^{th} node visited by a question, and p_j the single step success rate from Node j. With probability $1-p_j$, Node j has to forward the question to a next node. Thus the probability that the referral session is unsuccessful is given by

$$\prod_{j=1}^{T}(1-p_j).$$

Hence, the success rate of a referral session with a T-hop limit of question forwarding is

$$P_s(T) = 1-\prod_{j=1}^{T}(1-p_j). \quad (2)$$

III. Performance Metrics

We consider a system in which a node looking for answers to a question issues k replicas of the question at the beginning.

The forwarding of each replica adopts the same strategy and is independent. In the system, a heterogeneous setting of d (the number of a node's neighbors) causes heterogenous success rate at each step, leading to complex computations. To simplify the analyses, we first consider the homogeneous setting of d and approach the heterogenous case based on several real OSNs in Section IV. Without loss of generality, in this section the subscripts for the question indices are omitted for the r, e, and q in Section II.

1) Success rate: We consider the success rate for a referral session as the probability that at least one expert on a question is found for one of the k replicas of the question within T hops of forwarding. Denote by q the single step success rate. From Eqn. (2), we obtain the success rate of a referral session, as

$$\begin{aligned} P_s(k,T) &= 1-(1-q)^{kT} \\ &= 1-(1-e)^{k\cdot T}(1-r\cdot e)^{dkT}. \end{aligned} \quad (3)$$

The above result can be utilized to customize the control setting of an OSN user's agent. Suppose a person that looks for a piece of information has a desired success rate of finding a relevant expert, formulated as $P_s(k,T) > \gamma_p$. By Eqn. (3), we can obtain the following requirement for $k \cdot T$:

$$k \cdot T > \frac{\log(1-\gamma_p)}{\log \Delta}, \quad (4)$$

where $\Delta = (1-e)(1-re)^d$. Accordingly, the questioner's agent can pick k and T based on (4).

2) Referral cost: We consider the referral cost as, upon termination of a referral session, the total number of nodes[2] that the replicas of a question have visited. Referral cost is an important metric since it reflects the degree that a questioner's privacy is exposed to others. From a networking point of view, it also measures the network resource consumption.

Let C be the number of nodes visited by a certain question. $C = j$ indicates that no expert on the question is found in the first $j-1$ steps and an expert is found in the j^{th} step. According to the definition of the single step success rate, q, that $C = j$ equals $q(1-q)^{j-1}$. A special case occurs when $C = T$, which indicates that no expert on the question is found in the previous $T-1$ steps. To summarize,

$$P[C=j] = \begin{cases} q(1-q)^{j-1} & 0 < j < T \\ (1-q)^{T-1} & j = T \\ 0 & \text{otherwise} \end{cases}$$

Therefore, we can compute the expected value of C by

$$E[C] = \sum_{j=1}^{T} j \cdot P[C=j]. \quad (5)$$

Substituting q given by (1) into (5) yields

$$E[C] = \frac{1-\mathcal{A}^{(T-1)}}{1-\mathcal{A}} + \mathcal{A}^{(T-1)}, \quad (6)$$

where $\mathcal{A} = 1-q = (1-e)(1-re)^d$.

[2]Here, we do not consider whether a node receiving a question has been visited by this question before.

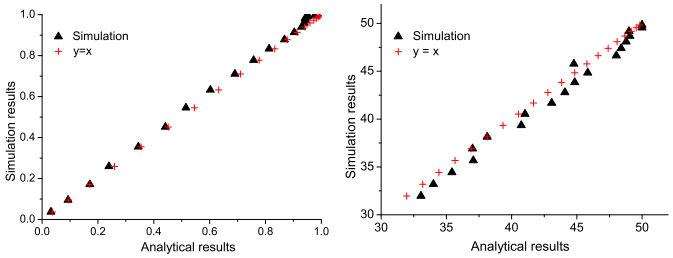
OSN	Orkut	LiveJournal	Flickr
Number of nodes	3,072,441	5,284,457	1,846,198
Estimated crawled fraction	11.3%	95.4%	26.9%
Number of links	223,534,301	77,402,652	22,613,981
Av. no. of friends per node	106.1	16.97	12.24
Fraction of symmetric links	100.0%	73.5%	62.0%

TABLE I
STATISTICS OF THE ONLINE SOCIAL NETWORK DATASETS [12].

For a referral session in which k replicas of a question is issued, the expected referral cost is given by

$$\begin{aligned} E[C[k]] &= k \cdot E[C] \\ &= k \cdot \{\frac{1-\mathcal{A}^{(T-1)}}{1-\mathcal{A}} + \mathcal{A}^{(T-1)}\} \end{aligned} \quad (7)$$

The requirement on the referral cost can not be explicitly expressed as a function of k and T. However, based on Eqn. (7), an agent can still numerically work out a proper pair of (k,T) according to its user's expectation on privacy exposure.



(a) Success rate (b) Referral cost

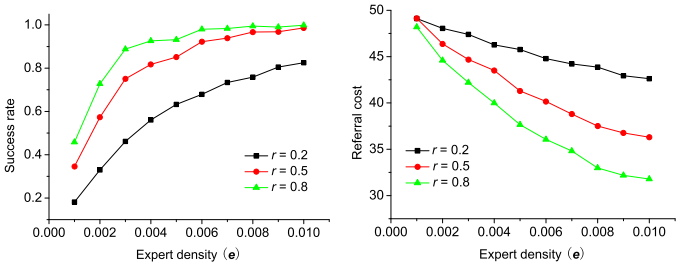
Fig. 2. Comparison between analytical results and simulation results.

We generate a regular graph where every node has a degree of $d = 10$, and we conduct 20 sets of simulations with different pairs of awareness level and expert density (r,e), with $r = 0.05, 0.1, ..., 1$ (20 values with increment of 0.05), and the corresponding $e = 0.0005, 0.001, ..., 0.01$ (20 values with increment of 0.0005). We set $k = 5$ and $T = 10$. Fig. 2 compares the analytical results with the simulation results. We observe that the analytical results match the simulation results well.

IV. EVALUATION

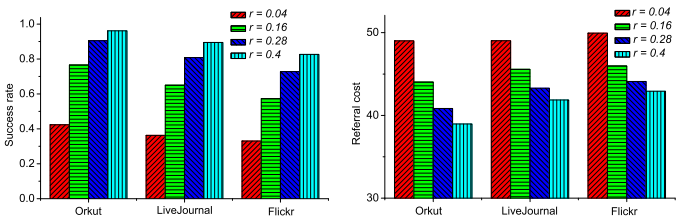
In this section, we empirically study the performance of a referral session based on our modeled question-forwarding strategy. We utilize the connectivity data of a set of OSNs, namely Orkut, LiveJournal, Flickr, collected by Mislove et al. [12]. Orkut is a website of explicitly defined social network to help users meet new friends and maintain existing relationships. LiveJournal is an online social network of bloggers. Flickr is a photo hosting and sharing website and online community platform. The major statistics of these datasets are summarized in Table I. We believe it is more realistic to evaluate the system on these real social network data. Since the networks are too big for evaluation, we sample several different portions from each network with Snowball sampling [15], whereby we randomly choose a single node and include all of its neighbors. Then all the nodes neighboring with those chosen in the previous step are picked. This process is repeated until the desired number of nodes are sampled.

Snowball sampling tends to pick hubs (high degree nodes) and under-sample low-degree nodes. Nevertheless, since the purpose of this work is to study the system on well connected networks with large-scale user participation (see Section II-B), the sampling method satisfies our requirement, and whether the beginning node is a hub does not make notable distinctions in statistical properties of the sampled networks. We set the size (number of nodes) for the sampled networks to $10,000$. $k = 5$ and $T = 10$.



(a) Success rate (b) Referral cost

Fig. 3. Performance against expert density e (Flickr).



(a) Success rate (b) Referral cost

Fig. 4. Performance against awareness level r.

Fig. 3 presents the simulation results with different awareness level $r = 0.2, 0.5, 0.8$ on the Flickr dataset. It illustrates the performance improvement as the expert density e increases. We also observe that the improvement speed decreases as e increases. Fig. 4 presents the performance over three OSN datasets. We set $e = 0.005$, and $r = 0.04, 0.16, 0.28, 0.4$ (4 values with increment of 0.12). We note that the performance improves as the awareness level r increases. The figure also reveals that the performance improvement is less significant when r is high. We also observe that the simulation results from the three different datasets are slightly different. In particular, the results from Orkut are better than those from the other two datasets, with respect to both the success rate and the referral cost. By investigating the sampled networks, we find that the Orkut dataset has higher average degree than that of LiveJournal or Flickr. This leads us to an empirical conclusion: the network connectivity, characterized by the average degree, is positively related to the system performance.

V. Conclusion

In conclusion, we study the problem of locating experts for relevant information via OSN from a networking point of view. We utilize random walk to model a node's action of forwarding a question, and adjust the forwarding probability by the node's awareness of the potential expertise of its neighbors. We derive analytical expressions of the performance metrics of a referral session (i.e. success rate and referral cost) in terms of the nodes' awareness level of their neighbors and the percentage of nodes that may have appropriate answers to the posed question. The analytical result can be utilized to customize the control setting of an OSN user's agent according to his expectation on a referral session. We also evaluate the system on the datasets of several real OSNs, and the simulation results validate our analyses. In this paper, we assume the homogeneous setting that advices from different experts on the same question have the same effect on the questioner, and it would be interesting to study the information filtering considering the system's trust [13] and reputation [16] mechanism. Our study also assumes every node is willing to forward questions for others. In the future, we would like to study the performance of a system when the incentive from a requesting node [17] and the altruism of a requested node [18] are considered.

Acknowledgement

This research is supported in part by the University of Hong Kong Strategic Research Theme of Information Technology.

References

[1] S. Milgram, "The small world problem," *Psychology Today*, pp. 61–67, May 1967.
[2] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun 1998.
[3] J. Kleinberg, "The small-world phenomenon: An algorithm perspective," in *Proceedings of the 32th Annual ACM Symposium on Theory of Computing*. 2000, pp. 163–170, ACM.
[4] T. Tassier and F. Menczer, "Emerging small-world referral networks in evolutionary labor markets," *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 5, pp. 482–492, Oct 2001.
[5] D. J. Watts, P. S. Dodds, and M. E. J. Newman, "Identity and search in social networks," *Science*, vol. 296, pp. 1302–1305, May 2002.
[6] J. Zhang and M. V. Alstyne, "SWIM: fostering social network based information search," in *CHI '04*. 2004, ACM.
[7] Aardvark, "http://vark.com/," .
[8] Google social search, "http://www.google.com/support/websearch/bin/answer.py?hl=en&answer=165228," .
[9] Hunch, "http://hunch.com/," .
[10] Yahoo!Answers, "http://answers.yahoo.com/," .
[11] H. Kautz, B. Selman, and M. Shah, "Referral Web: combining social networks and collaborative filtering," *Communications of the ACM*, vol. 40, no. 3, pp. 63–65, 1997.
[12] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. 2007, pp. 29–42, ACM.
[13] F. E. Walter, S. Battiston, and F. Schweitzer, "A model of a trust-based recommendation system on a social network," *Journal of Autonomous Agents and Multi-Agent Systems*, vol. 16, no. 1, pp. 57–74, Feb 2008.
[14] C. Gkantsidis, M. Mihail, and A. Saberi, "Random walks in peer-to-peer networks," in *Proceedings of INFOCOM '04*. 2004, IEEE.
[15] S. K. Thompson, *Sampling*, John Wiley & Sons Inc., 2 edition, April 2002.
[16] L. Mui, M. Mohtashemi, and A. Halberstadt, "A computational model of trust and reputation for e-businesses," in *Proceedings of HICSS '02*. 2002, p. 188, IEEE.
[17] J. Kleinberg and P. Raghavan, "Query incentive networks," in *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*. 2005, pp. 132–141, IEEE.
[18] K. Xu, P. Hui, V. O.K. Li, J. Crowcroft, V. Latora, and P. Lio, "Impact of altruism on opportunistic communications," in *Proceedings of ICUFN '09*. 2009, pp. 153–158, IEEE.