Algorithm to Trade Off between Utility and Privacy Cost of Online Social Search

Yan Li, Zhiyi Lu, Victor O. K. Li
Department of Electrical and Electronic Engineering
The University of Hong Kong
Email: {liyanhku, zylu, vli}@eee.hku.hk

***Abstract*—Online social search such as Quora and Zhihu brings new ways to obtain answers to questions in social networks. However, personal and other sensitive information may be exposed to others when the question spreads via the social network. People obtain utility when they obtain answers to their questions but may suffer privacy cost when their personal information is known by others. Researchers are seeking methods and tools to help users get utility and protect their privacy at the same time. In this paper, we study this problem by proposing a framework to quantitatively evaluate the utility and privacy cost in online social search. Besides, we design an algorithm under our framework to help users trade off their utility and privacy cost. Simulations are performed to illustrate the concepts in our framework and demonstrate the advantages of our algorithm.**

***Index Terms*—Trade-off, Utility, Privacy cost, Online Social Search**

I. Introduction

With the rapid development of online social networks (OSNs), we can get information much more easily and quickly than before. We can communicate with our friends more frequently and conveniently. By putting our resume on the social network, we could access more career opportunities. Specifically, we could seek answers from experts by using the question-and-answer social network website. We refer to the OSN-based question-and-answer mechanism as online social search (OSS). In a typical OSS system, a user looking for information can post a question and send it to his friends or persons recommended by the system. Users who get this question may answer it or forward it to others. Thus OSS can take advantage of OSNs to look for experts [1]. By using OSS, the question may finally reach the experts and get a great number of responses. In this way, the user who posts or who forwards the question may get utility.

However, privacy cost has become a major concern. This problem is exacerbated as we upload more of our personal information. Since the sharing of personal data is a popular activity over OSNs, sensitive personal information is frequently leaked. In an OSS system, people usually upload their real personal information to make themselves recognized by their friends and other users, or to be judged as experts by the system. However, when a question is asked and passed around to other users along friendship links, the questioner's personal information may also be exposed. The more the number of people who have received this question, the higher the questioner's privacy exposure.

It is desired to take advantage of online social search and protect our personal information at the same time. Several recommendation methods were designed based on the social network structure [2] and other personal information [3] in order to help people distinguish experts from others in the network. However, the privacy problem is seldom addressed. Current research discusses privacy protection in social networks from several aspects: the privacy settings [4], [5], the privacy problem according to people's behavior [6], [7], and how to build good protocols to protect people's privacy [8]. The effect of information diffusion was seldom considered even though it may greatly influence the utility and privacy cost. Furthermore, methods for quantitatively evaluating the utility and privacy cost are rare.

To help the user take advantage of social search, [9] studied the trade-off between the privacy risk and social benefit. To quantitatively evaluate the utility and privacy cost, information diffusion should also be taken into consideration [10]. In this paper, we propose a framework for quantitatively evaluating users' utility and privacy cost in OSS. The framework measures the utility and privacy cost, and several information diffusion models can be used in the framework.

In a real social search system, users can build his own profile showing their expertise and interests. They can also build friendships through the OSS system and can view questions their friends followed and answered, thus propagating the questions through friendship links. When a user wishes to ask a question in the system, he may post the question to some persons (selected by the user, or recommended by the system) in the beginning. Then the question spreads through the social network. The utility and privacy cost may change with the spread of the question. Suppose the OSS system has information on the expertise of the user and the social network structure. How do we choose the right people to ask the question in the beginning so we could find more experts to answer our question at the end while not suffering too much privacy exposure? We build an algorithm to solve this problem and use the framework to quantitatively evaluate the utility and privacy cost and test our algorithm.

The remainder of this paper is as follows. The framework to trade off the utility and privacy cost is presented in Section II,

978-1-4799-6664-6/16/$31.00 ©2016 IEEE

followed by the analysis of the online social search problem and our proposed Utility and Privacy Cost algorithm in Section III. We apply our framework to the real data of OSN with various of settings in Section IV. Finally, we conclude our study with suggestions for future work in Section V.

II. Framework of the trade-off between utility and privacy cost in Online Social Search

In this section, we use our framework to evaluate the trade-off between utility and privacy cost in online social search.

A. Framework of the utility and privacy cost of online social search

When people use social networks, they usually upload their personal information and share their information with others. The system has some regulations for profile uploading and information diffusion. For example, Facebook has 27 settings for personal information and four options (oneself, friends, friends of friends, everyone) to choose the people who can view your information. Users also have additional information like photos, questions, etc. When a user wants to spread something, a piece of information is formed. The information is a mixture of some personal information and additional information. The user may incur privacy cost and receive utility when information is spread to others.

Our framework can be divided into three parts. The first part is the measurement of the utility people may get from the social network. As there are many different kinds of utility, different scoring systems should be built. The scoring system should take some factors into consideration, like people's relationship, the quality of the answer people may get in a social search problem and so on. Based on this scoring system, we can measure how much utility may be obtained from an activity in the social network. On the other hand, we also need to measure the privacy cost, which is the second part of the framework. Part of the scoring system for the privacy cost is built based on the profile settings [11], [12], and this can be used to evaluate the privacy cost. However, people's activity like sharing behavior may spread personal information. Thus, the measurement of privacy cost may not only be influenced by the sensitivity of the information, but also by the transmission process. The third part of our framework is the transmission model. Several models in influence maximization can be used in our framework. In [13], the authors proposed two models: *Independent Cascade* (IC) model and *Linear Threshold* (LT) model, which are the basis of many other variant models.

Based on our framework of trade-off between utility and privacy cost, several strategies can be designed to help users get more utility and protect their privacy.

B. Assumption and definition in Online Social Search problem

The social search problem [14] is one in which a user wants to find the answer for a question in the social network where some experts exist. As mentioned before, the questioner needs to make a trade-off between the utility and the privacy cost. The more people he send the question to, the higher the probability he may get the answer, but the higher the privacy cost he suffers.

We consider an OSN as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{L})$ where $\mathcal{V}$ is the set of nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges of the social network. Let $n = |\mathcal{V}|$ be the number of users in the system. The edges represent the friendship or other relationship in the social network, and the question may pass through the edges in the online social search system. For every edge $(u, v) \in \mathcal{E}$, $p(u, v)$ denotes the probability of influence from u to v. We adopt the Independent Cascade (IC) model [13] here, which will be described in more details later. $L = \{l_1, l_2, ..., l_k\}$ is the set of labels to indicate expertise in various fields. For example, $L = \{computerscience, economics, geogrophy, ...\}$. Each node $u \in V$ has a set of labels $LB(u) \subseteq L$.

We assume that the experts needed in a certain field will give the questioner utility if he answers the question, and each person who view the question may cause privacy cost. Thus, we can use the label set of a person to measure how much utility a questioner may get when his question is passed to that person. When we ask a question in the social network, we may have a set of labels $L_e \subseteq L$ representing the expertise required. For example, we need to find people who have labels $\{computerscience, economics\}$ when we want to ask a computational economics question. Then L_e would be $\{computerscience, economics\}$. We set the utility value p_{l_i} for each label l_i. For a label in L_e, the utility will be positive, otherwise, it will be zero. The utility that one may get from a user u with label set $LB(u)$ who answers the question will be $p_{LB(u)} = \Sigma_{l_i \in LB(u)} p_{l_i}$. Different from the problem mentioned in [15], which uses a label for target marketing, we use labels to measure the utility we may get when asking a question. Besides, we also consider the privacy cost people may suffer when they ask a question in the social network.

People may share their personal information when using the social network, and the personal information may spread when the question spreads to others. In addition, the question itself may also contain some personal information. Thus, people may incur privacy cost when their personal information become known by others. We simplified the scoring system of the privacy settings. We assume one kind of personal information like age, sex, bank account may have a certain privacy cost and neglect the relationship of them. We use $PI = \{pi_1, pi_2, ...pi_m\}$ to denote all kinds of personal information for one person in the system. For each kind of personal information pi_i, we have privacy cost $c(pi_i)$ when it is known by others. The cost of a kind of personal information is determined by the user. The measurement of the privacy cost may vary with different users. When a person asks a question, the question or other personal information may incur privacy cost when spread. Denote private information spreaded in the OSN as a set of personal information $PIS \subseteq PI$. Under our assumption that each kind of information may cause privacy cost separately, the privacy cost incurred when the personal information PIS is known by one person is $C = \sum_{pi_i \in PIS} c(pi_i)$.

When the question spreads, people's utility and privacy cost may change. We assume that the online social search system does not have access control in order to let more users get the answer if they have interest in that question. The questioner starts by selecting an initial set of nodes to ask the question. We use the IC model [13] to represent the diffusion of the information. In the IC model, the user may choose a set of seed nodes $S \subseteq V$. Let S_t be the node set newly activated at time t, with $S_0 = S$ and $S_t \cap S_{t-1} = \emptyset$. At time t+1, every node $u \in S_t$ tries to activate its neighbors $v \in V \setminus \cup_{0 \leq i \leq t} S_i$ independently with probability $p(u, v)$. We use $A(S)$ to denote the set of nodes eventually activated by the seed node set S. $\sigma(S)$ is the expected value of $|A(S)|$. $\bar{E} = \{i | i \in A(S), LB(i) \cap L_e \neq \emptyset\}$ is the set of experts activated by the seed set S. Then the utility the questioner may get by choosing the seed set S will be $U_{L_e}(S) = \Sigma_{u \in \bar{E}} p_{LB(u)}$. The cost will be $C \times \sigma(S) = C\sigma(S)$.

III. Utility and Privacy Cost Algorithm

Our goal is to identify the seed node set to spread the question in the beginning to have a better trade-off of utility and privacy cost. In particular, we want to maximize the utility per privacy cost, that is, to maximize

$$U_{L_e}(S)/C\sigma(S)$$

Kempe et al. [13] proved two properties of the $\sigma(\cdot)$ function. One is submodularity, i.e. $\sigma(S \cup \{v\}) - \sigma(S) \geq \sigma(T \cup \{v\}) - \sigma(T)$ for all $v \in V$ and all subsets S and T with $S \subseteq T \subseteq V$; and the other is monotonicity, i.e. $\sigma(S) \leq \sigma(T)$ for all set $S \leq T$. For any non-negative function $\sigma(\cdot)$ that is both submodular and monotone, it can be proved that the simple greedy algorithm can provide $1 - 1/e$ approximation for maximizing $\sigma(S)$ among all sets S of size k.

We can also show that U_{L_e} is a non-negative, monotone and submodular function. However, $U_{L_e}(S)/C\sigma(S)$ is not submodular, then we could not use the simple greedy algorithm as the approximation for maximizing. If we only consider how to maximize the utility, then several algorithms like Greedy Algorithm, Degree Discount Algorithm [16] could be modified to be used here. The Labeled Degree Discount [15] heuristic which is used for target marketing can be used here to efficiently find the effective seed nodes to give user more utility. In order to protect our personal information at the same time, we propose our Utility Privacy Cost Discount Algorithm to efficiently find the seed nodes which could give us more utility under the same privacy cost.

The Degree Discount Algorithm assumes that the probability of the influence in IC model is small. So it is efficient to only consider one-hop neighbor nodes and select nodes with high degree values which may have more potential to influence others to be the seed nodes. The Degree Discount Algorithm need to be run k times if one want to select k seed nodes. The main idea of the Degree Discount Algorithm is that after selecting a node w as the seed node, then we will recalculate the expectation of the influence of w's neighbor v as v's expectation of the influence may decrease due to the selection of seed w.

Let $N(v)$ be the neighborhood of the node v, d_v be the degree of the node v. t_v denotes the number of neighbors of node v that are selected as seeds. The probability parameter in the IC model is p. Then the probability that node v is not influenced by its neighbor seeds is $(1-p)^{t_v}$, and the expected number of additional nodes influenced by v is $(1-p)^{t_v} \cdot (1 + (d_v - t_v) \cdot p)$.

We revised the Labeled Degree Discount Algorithm to our Utility Degree Discount Algorithm to identify seed nodes to maximize our utility. Algorithm 1 describes the Utility Degree Discount Algorithm. Here we consider the situation that the question is related to only one field. We suppose that all the experts in one field are the same in their expertise and will surely answer the question upon receiving the question. Thus it means that the more experts receiving the question, the more utility the questioner will get. The label of the experts would be L_e and we suppose the utility that each expert gives the questioner is 1.

We make two modifications compared to the Degree Discount Algorithm. Firstly, only the experts would give us utility, the degree of node v is modified to be the number of neighbors who are experts in a certain field. Then the new degree of node v is $d_v = \{u | (u, v) \in \mathcal{E} \& LB(u) = L_e\}$. We use t_v to denote the number of neighbors of v who are both seeds and experts, and use s_v to denote the number of neighbors of v are seeds. Secondly, whether v is an expert will influence the degree discount. If v is an expert, then $LB(v) = L_e$, $dd_v = (1-p)^{s_v}[1 + (d_v - t_v)]$, else $dd_v = (1-p)^{s_v}(d_v - t_v)$.

Thus, the Utility Degree Discount Algorithm is shown in Algorithm 1, in which S is the seed node set.

Even though the Utility Degree Discount Algorithm could identify the experts efficiently, it incurs privacy cost. The more people receiving our question, the more privacy cost incurred. Here we assume that each user who gets to know the question would give the questioner the same privacy cost.

We could not directly use Degree Discount Algorithm to get the approximation of the maximum of a non-submodular function. In order to consider the utility and privacy cost at the same time, we make a modification on the Utility Degree Discount Algorithm. As we suppose each user who knows the question would cause privacy cost, then the privacy cost may be related to the degree of the user. The Utility Cost Ratio is used instead of the degree to select the seed nodes.

dg_v represents the degree of the node v, other symbols are the same with the Utility Degree Discount Algorithm. We change dd_v to represent the utility and privacy cost ratio, and it will also be recalculated after each seed is selected. $dd_v = d_v/dg_v$ in the beginning. After seeds are selected, we recalculate dd_v. If v is an expert, then $LB(v) = L_e$, $dd_v = (1-p)^{s_v}[1 + (d_v - t_v)]/(dg_v - s_v)$, else $dd_v = (1-p)^{s_v}(d_v - t_v)/(dg_v - s_v)$.

Algorithm 1 Utility Degree Discount Algorithm

```
Initialize $S = \emptyset$
for each node $v \in \mathcal{V}$ do
    compute its degree $d_v$
    $dd_v = d_v$
    Initialize $|t_v| = 0$,$|s_v| = 0$
    for $i = 1$ to $k$ do
        Select $u = argmax_{v \in \mathcal{V} \setminus S}\{dd_v\}$
        $S = S \cup \{v\}$
        for each neighbor $v$ of $u$ and $v \in \mathcal{V} \setminus S$ do
            $s_v = s_v + 1$
            if $LB(u) = L_e$ then
                $t_v = t_v + 1$
            end if
            if $LB(v) = L_e$ then
                $dd_v = (1-p)^{s_v}[1 + (d_v - t_v)]$
            else
                $dd_v = (1-p)^{s_v}(d_v - t_v)$
            end if
        end for
    end for
end for
return $S$
```

Algorithm 2 Utility Privacy Cost Ratio Discount Algorithm

```
Initialize $S = \emptyset$
for each node $v \in \mathcal{V}$ do
    compute its degree $d_v$
    $dd_v = d_v / dg_v$
    Initialize $|t_v| = 0$,$|s_v| = 0$
    for $i = 1$ to $k$ do
        Select $u = argmax_{v \in \mathcal{V} \setminus S}\{dd_v\}$
        $S = S \cup \{v\}$
        for each neighbor $v$ of $u$ and $v \in \mathcal{V} \setminus S$ do
            $s_v = s_v + 1$
            if $LB(u) = L_e$ then
                $t_v = t_v + 1$
            end if
            if $LB(v) = L_e$ then
                $dd_v = (1-p)^{s_v}[1 + (d_v - t_v)]/(dg_v - s_v)$
            else
                $dd_v = (1-p)^{s_v}(d_v - t_v)/(dg_v - s_v)$
            end if
        end for
    end for
end for
return $S$
```

IV. Evaluation

In this section, we study empirically the trade-off of the utility and privacy cost based on our framework. We use the connectivity data of Facebook [17]. The dataset of the social network consists of 4039 nodes and 88234 edges. As the dataset does not have the expertise information, we choose an alternative solution. Since experts or people with the same interests usually form into groups or communities, we could use community detection method to detect communities and treat each community as an experts community in one field.

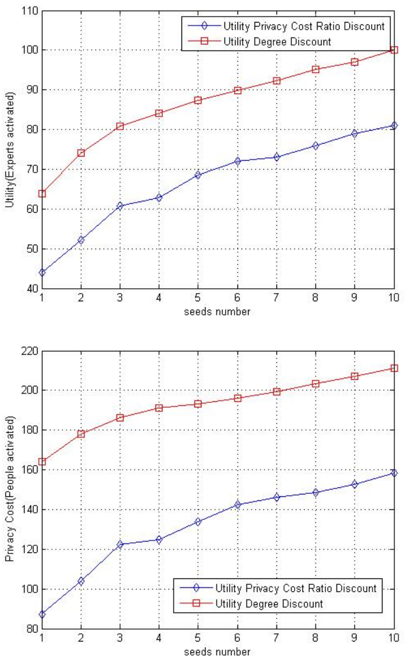

Fig. 1. Utility and Privacy Cost of question 1

Louvain Method [18] is a community detection method which has a better performance and less computation time than other methods. We use it to divide the graph into 101 communities. The sizes of the communities range from 426 nodes to 2 nodes. We treat each community as an experts community from one field when we conduct experiments to evaluate the effectiveness of our algorithm.

We select three communities with different sizes and treat them as experts communities in three areas to do the simulation. The size of community 1 is 426 nodes, of community 2 is 400 nodes, and of community 20 is 41 nodes. We suppose that the influence probability in the IC model is 0.02. We suppose that three questions are asked in the system and the experts for each question are from the three communities. We use the Utility Degree Discount Algorithm and the Utility Privacy Cost Ratio Discount Algorithm to determine the seed nodes and analyze how the Utility and Privacy Cost change with the selection of seed nodes.

Fig. 1 and Fig. 2 present the Utility and Privacy Cost with different number of seed nodes. The results are obtained as averages of 10000 runs. They illustrate how Utility and Privacy Cost change when the number of seed nodes increases. We can see that both the utility and privacy cost may increase as we choose more seed nodes. After more seeds are selected, the speed of increase would slow down. We also observe that the utility and privacy cost may change greatly with the expert community structure (c.f. community 1 and community 2.)

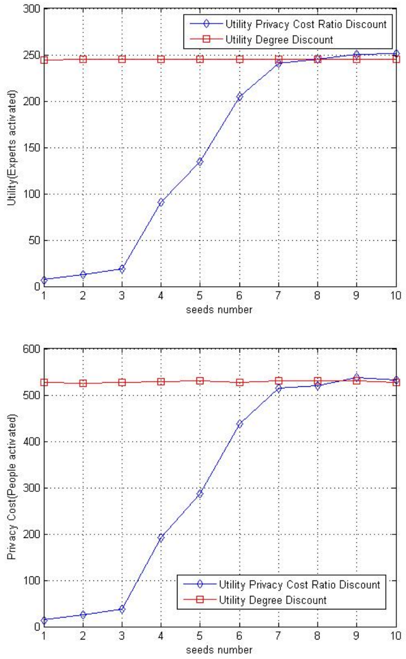

Fig. 2. Utility and Privacy Cost of question 2

Besides, we find that both algorithms would find most of the experts even though we choose only a small number of seed nodes. Compared to the Utility Privacy Cost Ratio Discount Algorithm, the Utility Degree Discount Algorithm could give the user more utility with the same number of seeds. However, it would also incur more privacy cost.

In order to judge which algorithm have a better performance on the trade-off of the utility and privacy cost, we analyze how the utility and privacy cost ratio changes with the number of seed nodes. As we simplified the utility of the expert to be 1 unit and the privacy cost incurred when one person knows the question to be 1 unit, we use the ratio of the number of experts and the number of people who know the question as the ratio of the utility and privacy cost. We analyze and compare the ratio of the utility and privacy cost of the two algorithms.

Fig. 3 presents the simulation results of the utility and privacy cost ratio for the two algorithms and for the three questions. When we do our simulation on community 3, we change the probability of the IC model to 0.05 in order to find if the performance of our algorithm will be influenced by the probability. We run 10000 times and get the average of the ratio. We also get the 95 percent confidence interval. We observe that the Utility Privacy Cost Ratio Discount Algorithm has a better performance in protecting privacy than the Utility Degree Discount Algorithm when we choose less nodes. When more nodes are selected, then the performance differential would decrease since the two algorithms may choose similar group of nodes when more seeds are needed. Besides, the performance of the Utility Privacy Cost Ratio Discount algorithm will still be better than the Utility Degree Discount Algorithm when the probability of the IC model changes.

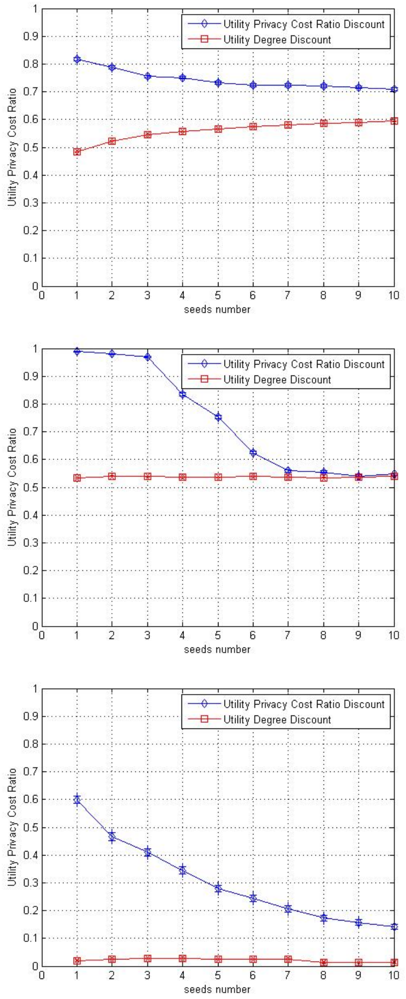

Fig. 3. Utility and Privacy Cost ratio of 3 questions

V. Conclusion and Future work

In this paper, we study the trade-off between utility and privacy cost in OSS. We propose a framework which quantitatively evaluates the utility and privacy cost, and accounts for the information diffusion model. As current OSS systems do not have access control mechanism of the question, we specially try to find a strategy to help the questioner seek seed nodes to get a better trade-off of the utility and privacy cost. We propose our Utility Privacy Cost Ratio Discount Algorithm to solve this problem. Our algorithm is a first step to analytically study the trade-off problem in the OSS. From our study, we also find that the privacy cost may be highly related to the experts distribution and the information diffusion control of the system.

Besides, we assume that all expertise information is already know by the system. As many online social search websites have some expertise mechanisms to find expert, we would like to improve our algorithm by incorporating such mechanisms in the future. In addition, we hope to use our framework to design a better information diffusion control mechanism for the system, which would help us make a better trade-off.

References

[1] D. J. Watts, P. S. Dodds, and M. E. Newman, “Identity and search in social networks,” *Science*, vol. 296, no. 5571, pp. 1302–1305, 2002.

[2] J. Zhang, M. S. Ackerman, and L. Adamic, “Expertise networks in online communities: structure and algorithms,” in *Proceedings of the 16th International Conference on World Wide Web*, ACM, 2007, pp. 221–230.

[3] M. Fazel-Zarandi, H. J. Devlin, Y. Huang, and N. Contractor, “Expert recommendation based on social drivers, social network analysis, and semantic data representation,” in *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, ACM, 2011, pp. 41–48.

[4] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove, “Analyzing facebook privacy settings: user expectations vs. reality,” in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement*, 2011, pp. 61–70.

[5] K. Lewis, J. Kaufman, and N. Christakis, “The taste for privacy: An analysis of college student privacy settings in an online social network,” *Journal of Computer-Mediated Communication*, vol. 14, no. 1, pp. 79–100, 2008.

[6] B. Krishnamurthy and C. E. Wills, “Characterizing privacy in online social networks,” in *Proceedings of the First workshop on Online Social Networks*, ACM, 2008, pp. 37–42.

[7] S. Ahern, D. Eckles, N. S. Good, S. King, M. Naaman, and R. Nair, “Over-exposed?: privacy patterns and considerations in online and mobile photo sharing,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2007, pp. 357–366.

[8] R. Zhang, J. Zhang, Y. Zhang, J. Sun, and G. Yan, “Privacy-preserving profile matching for proximity-based mobile social networking,” *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 656–668, 2013.

[9] M. Yang, Y. Yu, A. K. Bandara, and B. Nuseibeh, “Adaptive sharing for online social networks: a trade-off between privacy risk and social benefit,” in *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2014, pp. 45–52.

[10] Y. Zeng, Y. Sun, L. Xing, and V. Vokkarane, “Trust-aware privacy evaluation in online social networks,” in *2014 IEEE International Conference on Communications (ICC)*, 2014, pp. 932–938.

[11] K. Liu and E. Terzi, “A framework for computing the privacy scores of users in online social networks,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, no. 1, p. 6, 2010.

[12] Y. Wang, R. K. Nepali, and J. Nikolai, “Social network privacy measurement and simulation,” in *2014 International Conference on Computing, Networking and Communications (ICNC)*, IEEE, 2014, pp. 802–806.

[13] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” in *Proceedings of the Ninth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, ACM, 2003, pp. 137–146.

[14] K. Xu and V. O. Li, “Privacy exposure of online social search.” in *IEEE GLOBECOM*, 2010, pp. 1–5.

[15] F.-H. Li, C.-T. Li, and M.-K. Shan, “Labeled influence maximization in social networks for target marketing,” in *2011 IEEE Third Inernational Conference on Social Computing (SocialCom) and 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT)*, 2011, pp. 560–563.

[16] 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*, ACM, 2009, pp. 199–208.

[17] J. Leskovec and J. J. Mcauley, “Learning to discover social circles in ego networks,” in *Advances in Neural Information Processing Systems*, 2012, pp. 539–547.

[18] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.