

Personal Social Screen – A Dynamic Privacy Assignment System for Social Sharing in Complex Social Object Networks

Lei Li

School of Computer Science
Florida International University
Miami, FL 33199
Email: lli003@cs.fiu.edu

Tong Sun

Xerox Innovation Group
Xerox Corporation
Webster, NY 14580
Email: Tong.Sun@xerox.com

Tao Li

School of Computer Science
Florida International University
Miami, FL 33199
Email: taoli@cs.fiu.edu

Abstract—Online social networks allow millions of individuals to create online profiles and share information with vast networks of friends, and often, unknown strangers. Privacy within social networking sites is often undefined, which might render potential privacy risks. In this paper, we present a dynamic trust-based privacy assignment system to help people select the privacy preference on-the-fly to the piece of content he/she is sharing, where trust information is derived from social network structure and user interactions. Our proposed system, Personal Social Screen (*PerCial*), first automatically detects a two-level topic-sensitive community hierarchy using the available resources, and then assigns privacy preference for users based on their personalized trust networks. Preliminary results on a social object network dataset collected from *Flickr* demonstrate the efficacy and effectiveness of our proposed system.

I. INTRODUCTION

Online information sharing (e.g., photos, posts, check-in locations and status updates) in popular social media web sites, such as *Twitter*, *Facebook*, *Flickr* etc., has grown exponentially in past several years. However, this phenomenon causes two major concerns: one is online users' privacy (i.e., some content you shared online is related to your private information); another is noisy data (i.e., flooding irrelevant content in your friends' "content stream"). The privacy in social networking sites is often poorly defined and non-transparent, which leaves people unaware of the consequences or implications of their actions since their personal information might be shared beyond their trusted friends/contacts [1], [2], and makes it extremely easy for third parties to create digital dossiers of their behavior.

In general, people have the preference on information sharing, i.e., sharing identifiable information with their trusted individuals, e.g., friends, family, colleagues, and people who have similar interests. Therefore, it is imperative to identify potential communities within social networks and to provide users a way of explicitly assigning their privacy preference on different communities that they are interested in. To this end, we investigate the correlations between social trust and personal privacy in the setting of social networks, and then propose a trust-based privacy assignment system to help users

easily setup their privacy preferences on specific social objects, e.g., posts, photos and music.

Specifically in our proposed system, a two-level community hierarchy is initially constructed based on the trust relations among online users in a social network, where the trustiness is populated from the social network structure (e.g., friendship) and the user-user interactions (e.g., conversations between users or common activities among users). To help users assign privacy preference on different communities, *PerCial* automatically analyzes the intrinsic properties of the trust correlation between the user and the communities, as well as the communities themselves, and recommends the user with a list of privacy-assignable communities for each level of the community hierarchy. User can drill down or roll up along different communities and set up their privacy preferences on the favorite communities when sharing potential private information or topic-sensitive objects.

To the best of our knowledge, our proposed system is the first journey towards helping online users dynamically setup their privacy preferences to a social sharing activity at different granularities of communities that are topically relevant. The contribution of our paper is three-fold:

- Our system formalizes the topic-sensitive trust relations among users by analyzing available resources in multi-modal social networks.
- We propose to loosely separate the entire social network and organize it as a two-level overlapping community hierarchy to facilitate users' navigation and exploration on communities.
- We propose to recommend communities to users for privacy preference assignment by modeling the community selection problem as a budgeted maximum coverage problem.

II. RELATED WORK

A. Privacy in Social Networks

The study proposed in [3] showed that, more than half (57%) of adult internet users say they have used a search

engine to look up their name and see what information was available about them online. Young adults, far from being indifferent about their digital footprints, are the most active online users in several dimensions. For example, more than two-thirds (71%) of social networking users ages 18-29 have changed the privacy settings on their profile of social networking services to limit what they share with others online.

Most social media sites (including Facebook) use a pre-defined list of privacy settings to govern all shares. Recently, Google uses the concept of “social circle” in upcoming Google+ social network to allow people explicitly manage their friend lists into “static” circles, which gives people choices of which friend circles he/she would like to share the content with. Although it facilitates more intimate sharing and connections with these circles, Google’s “social circle” remains as a primary contact management approach and it does not address the issues of selective sharing in user’s extended social network (e.g., sharing within a community, or with friends-of-friends). Meanwhile, the social conversations are dynamically evolving and at-the-moment sharing is usually topic-sensitive rather than confined by static “social circles”.

B. Community Detection

Community detection in networks has been studied for years in different areas, e.g., physics and computer science. Multiple approaches have been published to address the problem of detecting communities [4], [5]. Due to the special characteristics of social networks, finding overlapping communities is attracting more and more attention in recent years.

In general, the social network is often represented as a node-centric view [6], [7], [8] or a edge-centric view [9], and then analysis on such views is conducted to obtain potential communities. However, previous methods aim at obtaining a flat partition of the social network, but cannot obtain a community hierarchy to better capture the intrinsic relations among different communities. Recently, several techniques for overlapping hierarchical community detection are proposed. For example, [10] applies typical hierarchical clustering to link graph. Tang et al. [11] propose to detect a community structure purely based on the friendship in the social network, and then apply a regularization strategy on extracted communities for classifier construction. Different from prior approaches, our proposed framework first expands the social network by incorporating conversation and activity relations among users, and then hierarchically separates the enriched network.

III. SYSTEM FRAMEWORK

Our proposed system consists of two core components: 1) An offline module that automatically detects multi-resolution topic-sensitive overlapping communities in social networks using available resources, e.g., users’ friend relations, users’ conversations, users’ common activities, etc.; and 2) An online module to provide a trust-personalized community hierarchy based on a given user’s trust to communities with different granularities. Figure 1 depicts an overview of our proposed privacy assignment system.

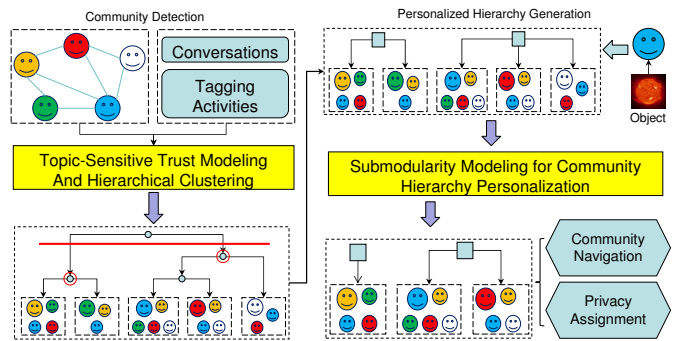


Fig. 1. System Overview of “PerCial”.

I. *Community Hierarchy Generation*: Our system initially constructs a social graph based on users’ relations, conversations and other available resources. The resulted graph is essentially a multi-modal graph, where the nodes represent online users, and the edges denote the interactions between users. Then we perform community detection on this graph, and finally obtain a two-level community hierarchy.

II. *Personalized Hierarchy Generation*: Based on the generated two-level community hierarchy, our system analyzes the trust correlations between a given user and the communities that the user belongs to at each level of the hierarchy. The personalization is achieved by modeling the community selection problem as a budgeted maximum coverage problem and then addressing it using a greedy approach.

IV. COMMUNITY DETECTION

Given a social network, along with heterogenous relations (e.g., friendships, conversations and activities) among online users, our goal is to automatically construct a two-level community structure of this social network. To this end, we first detect atomic communities based on multiple relations of users, and then perform a hierarchical agglomerative clustering on small communities of the network under the constraint of conversations and activities. To obtain the two-level community hierarchy, we perform dendrogram cut on the generated community structure in a dynamic way. Note that our algorithm allows users to belong to multiple communities.

A. Atomic Community Generation

In reality, online users tend to have more interactions with their social neighbors, i.e., friends and colleagues, or people who have similar interests with them. Based on this observation, we treat a user and his/her friends as an atomic community. For a social network with n users, we can have at most n atomic communities. Note that this allows one user to be engaged in multiple different communities simultaneously. For example, a user might be interested in photography and movies, and therefore he/she may belong to two different communities; Or a user who has preference on sports might participate in two different fine-grained communities, e.g., football and basketball.

Generally speaking, a user in social network may have hundreds or thousands of friends; however, it is common that this user might have instant interactions with only a small number of his/her friends, but not all of them. The more interactions between users, the more trust users might have. Besides the trust correlation, users who have similar interests might be in the same community as well. Therefore, to prune atomic communities described above, we employ three distinct correlations between the user and his/her friends, i.e., the conversations, the common activities (users might tag for the same object or comment on the same post), and the similarity of topics derived from posts. Formally, assume user u_i and u_j are friends, with the entire conversation set c_i and c_j (i.e., the messages that a user sends and receives), and the friend-oriented conversation set \hat{c}_i and \hat{c}_j (i.e., the messages that user u_i and u_j send to each other). Also, u_i and u_j tag or comment on social object set o_i and o_j , respectively (i.e., the co-occurrence of tagging activities between user u_i and u_j on some photos). The posts sets of these two users are denoted as p_i and p_j (i.e., the comments that a user posts on some objects). Then, we derive the trust score t_{ij} as

$$t_{ij} = \alpha \cdot \frac{|\hat{c}_i| + |\hat{c}_j|}{|c_j| + |c_i|} + \beta \cdot \frac{|o_i \cap o_j|}{|o_i \cup o_j|} + \gamma \cdot \text{sim}(p_i, p_j). \quad (1)$$

Here the first component represents the trust obtained from the historical conversations, the second component describes the trust oriented from the common activities, and the third one aims to capture the topical similarity between the posts of these two users¹. α , β and γ are weights for each component, determining how much our algorithm relies on the corresponding component. To dynamically prune atomic communities, we first compute the trust scores between the given user and his/her friends, and then sequentially select friends with the trust score larger than the median. By doing this, the atomic community becomes more compact and topic-sensitive, and therefore the subsequent hierarchical clustering procedure can be further sped up.

Discussion: In Eq(1), the parameters α , β and γ represent how important the corresponding component is when computing the pairwise user trust scores. These three parameters can help differentiate social networks with distinct properties.

- α : For social networks like *Facebook*, conversations account for a large proportion of user's activities. Users in such networks post and comment on other users' walls.
- β : For social networks like *Flickr*, people post and tag a gigantic amount of photos, and therefore the common activities, like tagging on the same images, would be predominant in such networks.
- γ : For social networks like *Twitter*, information transferring, i.e., topic diffusion, is the major activity over others. People would post a lot of tweets, and therefore we can place more weight on the component of the topical similarity when considering such networks.

¹Here the topical similarity can be calculated using the cosine similarity on two word vectors.

B. Community Hierarchy Generation

After obtaining all the atomic communities, we propose to perform hierarchical clustering on these atoms to generate a community hierarchy. Specifically, we recursively merge two communities with maximum group trust until all the communities are merged into one single community or certain criterion is satisfied, e.g., the community number reaches a predefined threshold. Here we define pairwise group trust as

$$\mathcal{T}_{G_k, G_l} = \sum_{u_i \in G_k, u_j \in G_l} t_{ij}, \quad (2)$$

where G_k , and G_l are two communities being merged. By performing dendrogram cut on the generated community structure under a predefined community number, or using a cluster evaluation measurement to automatically decide what is the best layer to cut, a two-level community hierarchy can be obtained, where the first level contains communities with general topics, and the second level involves fine-grained atomic communities. Note that the entire social network is loosely separated due to the overlapping property of atomic communities. To enrich the community representation, we apply topic detection (e.g., Latent Dirichlet Allocation (LDA) [12]) on the historical conversations of each community in the hierarchy, and select representative words as the topic to describe the major interest of the corresponding community.

In general, the social network evolves over time, i.e., new users might be engaged in the social network, a lot of conversations might happen, and common activities between users might be triggered everyday. To timely capture such changes in the social network, the procedure of community hierarchy generation is performed offline in a fixed frequency, e.g., one time per day.

V. PERSONALIZED HIERARCHY GENERATION

The community hierarchy provides us an elegant base for further assigning privacy preference. Based on this hierarchy, our system first automatically generates a personalized community sub-hierarchy of a given user, and then presents this sub-hierarchy to the user for interaction. Here the sub-hierarchy is obtained by considering the user's trust over the communities, and the topical correlations between the object being posted and the profile of communities. Specifically, we investigate the submodularity of the communities that the user belongs to, and then model community personalization as a budgeted maximum coverage problem, which can be solved in a greedy way.

A. Introduction to Submodularity

Let E be a finite set and f be a real valued nondecreasing function defined on the subsets of E that satisfies

$$f(T \cup \{\varsigma\}) - f(T) \leq f(S \cup \{\varsigma\}) - f(S), \quad (3)$$

where $S \subseteq T$, S and T are two subsets of E , and $\varsigma \in E \setminus T$. Such a function f is called a **submodular** function [13]. Intuitively, by adding one element to a larger set T , the value increment of f can never be larger than that by adding one

element to a smaller set S . This intuitive diminishing property exists in different areas, e.g., in social network, adding one new friend cannot increase more social influence for a more social group than for a less social group. Submodularity modeling has been employed into multiple research areas, e.g., document summarization [14] and news recommendation [15].

The budgeted maximum coverage problem is then described as: given a set of elements E where each element is associated with an influence and a cost defined over a domain of these elements and a budget B , the goal is to find out a subset of E which has the largest possible influence while the total cost does not exceed B . This problem is NP-hard [16]. However, [16] proposed a greedy algorithm which sequentially picks up the element that increases the largest possible influence within the cost limit and it guarantees the influence of the result subset is $(1 - 1/e)$ -approximation. Submodularity resides in each “pick up” step. A key observation is that submodular functions are closed under nonnegative linear combinations [17].

B. Submodularity Modeling for Hierarchy Personalization

Given a higher-level community that a user belongs to, the user’s trust over different sub-communities might vary significantly. To personalize the community selection, we need to consider the sub-communities that are more trusted by the user. In addition, the object being posted would have a piece of description with it, which can help us further filter the communities by emphasizing the similarity between the description and the community profile. Based on this intuition, our community personalization strategy can be described as follows (note that \mathcal{C} denotes a higher level community, \mathcal{S} represents the selected sub-communities, and ς is the sub-community being selected). After selecting ς ,

- The topics discussed in \mathcal{S} should be more relevant to the description of the object;
- \mathcal{S} should carry more trust for the user to assign privacy.

Per the above strategy, we define a quality function f to evaluate the current selected community set \mathcal{S} over the entire sub-community set \mathcal{N} as

$$f(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{c_1 \in \mathcal{S}} Trust(u, c_1) + Sim(o, \mathcal{S}), \quad (4)$$

where c_1 denotes a sub-community, u represents the given user, o is the social object being posted, $Trust(u, \cdot)$ describes how the user u trusts on the community and $Sim(\cdot, \cdot)$ represents the topical similarity between the community profile and the description of the object o . The resultant value of Eq.(4) ranges in $[0, 2]$. Notice that the description of a social object could be several sentences describing the details of the object. For example, assume you are uploading a photo of “Niagara Fall” to *Flickr*, you can specify the photo as “Niagara Fall is the most spectacular fall that I have ever seen”.

In Eq.(4), two components are involved, corresponding to the community selection strategy we list above. The former gives us the evidence that how much the selected sub-community set \mathcal{S} is trusted by the user, and the latter provides

a perspective on how similar that the selected community set is with the description of the object. $f(\mathcal{S})$ balances the contributions of different components. Note that both components are naturally submodular functions. Based on the non-negative linear invariability of the submodular function, $f(\mathcal{S})$ is also a submodular function. Suppose ς is the candidate sub-community, the quality increase is therefore represented as follows:

$$I(\varsigma) = f(\mathcal{S} \cup \{\varsigma\}) - f(\mathcal{S}). \quad (5)$$

The goal is to select a list of sub-communities which provide the largest possible quality increase within the budget². Hence, personalized community selection is transformed to the budgeted maximum coverage problem. For each sub-community set, a greedy algorithm is employed to solve the budgeted maximum coverage problem, by sequentially selecting the sub-communities with the largest quality increase based on the selected sub-community set until the budget is reached.

Discussion: The submodularity-based community selection strategy provides users with a diverse sub-community list within each general community in terms of topical aspect. Moreover, by taking into account the social trust among users, communities in the generated sub-community list might carry different privacy relevance (i.e., trust). In this way, we provide online users a flexible environment to assign privacy preference on different communities.

VI. PRELIMINARY EXPERIMENTAL EVALUATION

A. Real-World Dataset

For experiments, we crawled a set of social object network data from *Flickr*, including users, photos posted by users, comments and tags of photos and group information of users. Notice that in *Flickr*, there are some potential privacy issues if users do not set uploaded photos as private or visible to specific user groups. Third parties might be able to track online users’ activities, e.g., where the user goes. Therefore, we choose *Flickr* data for experimental analysis.

For evaluation purpose, only English comments are kept in the dataset. We also remove users without frequent activities, e.g., posting photos, commenting or tagging, and user groups whose group members are less than 5. After preprocessing, the *Flickr* dataset contains 4,219 users associated with the corresponding photo, comment and tag information. The number of groups in the dataset is 1,071.

B. Experimental Setup

The evaluation of our proposed privacy assignment system can be decomposed into two stages: (1) evaluating the performance of the community detection module; and (2) investigating the overall effectiveness of our proposed system. Our community detection algorithm is compared against four different methods, including two graph-based community detection algorithms MODULE [5] and SHRINK [18], and two overlapping clustering algorithms CONGA [6] and

²Here the budget can be predefined based on the structure of the original community hierarchy.

EAGLE [8]. Note that we compare the user clusters of the higher level in our community hierarchy against with the results of other methods. To evaluate the community hierarchy personalization and the overall effectiveness of our proposed method for social sharing privacy assignment, we report a user study on different experience indexes.

In our experiments, we employ two different measures, Normalized Mutual Information (NMI) and Jaccard Index (JI). NMI is an information-theoretic based measurement, aiming to evaluate the quality of communities generated by different methods. It is currently widely used in measuring the performance of network clustering algorithms [19]. Formally, the measurement metric NMI can be defined as

$$NMI = \frac{-2 \sum_{i,j} N_{ij} \cdot \log(\frac{N_{ij} \cdot N}{N_{i \cdot} \cdot N_{\cdot j}})}{\sum_i N_{i \cdot} \cdot \log(\frac{N_{i \cdot}}{N}) + \sum_j N_{\cdot j} \cdot \log(\frac{N_{\cdot j}}{N})}, \quad (6)$$

where N is the confusion matrix, N_{ij} is the number of nodes in both cluster X_i and Y_j , $N_{i \cdot}$ is the sum over row i of N and $N_{\cdot j}$ is the sum over column j of N . Note that the value of NMI ranges between 0.0 (total disagreement) and 1.0 (total agreement). JI [20] can be calculated as

$$JI(C_1, C_2) = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}, \quad (7)$$

in which n_{11} is the number of pairs of samples in the same cluster for both C_1 and C_2 , n_{01} and n_{10} are the number of samples' pairs belonging to the same cluster in one solution, but not in the other.

To facilitate online users' navigation and exploration on different levels of communities, we employ LDA on community profiles (i.e., the content of posts in a community) and provide concise summaries for each community using the result of LDA consisting of representative words that describe the major interest of community members. In the experiment, we set the number of topics for LDA as 50, and the maximum number of iterations as 100. Due to the space limit, we omit the detailed experimental procedure for tuning these parameters.

C. Effectiveness of Community Detection

Our community detection algorithm produces a list of clustering results and each clustering result corresponds to one layer of the generated dendrogram. To measure the quality of each clustering result, we use Dunn's Validity Index [21] as the metrics. This validity measure is based on the idea that high-quality clustering produces well-separated compact clusters. Generally, the larger Dunn's Index, the better the clustering.

After determining the best layer to cut the dendrogram, we can treat the number of clusters K in this layer as the optimal cluster number. To achieve this, we calculate Dunn's Index for each level of the community hierarchy, and then choose the one with the largest Dunn's Index. We intentionally filter the data by removing the user groups whose member are less than a pre-defined threshold μ , and then run different algorithms on the preprocessed dataset to evaluate the performance. In our experiment, μ is set to 10, 20, 30, 50, respectively.

Figure 2 shows the evaluation result on different levels of the community hierarchy.

From Figure 2, we can observe that: (i) when μ is small, a lot of small communities might be involved into the dataset, which might result in the activity sparsity, i.e., not too many activities happened in small communities, and therefore, the optimal Dunn's Index in more general community levels is relatively lower than the one with larger μ ; and (ii) they achieve the highest Dunn' Index when the hierarchy level is around 150. Therefore, we perform the dendrogram cut of the community hierarchy on the level 150, and obtain the top level of the community hierarchy. Here each general community is related to some general topics being discussed by community members. Note that not all these 150 communities are being represented to login users; instead, only the communities that a user belongs to would be selected as part of the personalized community structure.

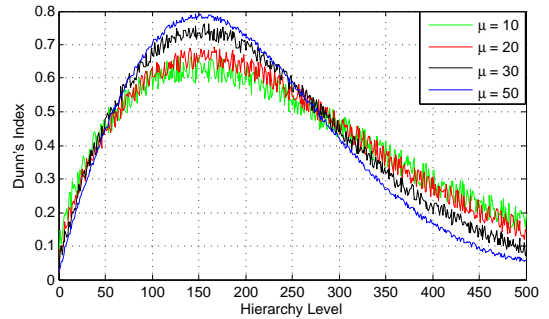


Fig. 2. Evaluation Result on Different Levels of the Community Hierarchy.

To measure the results of community detection in terms of community quality, we calculate NMI and JI of two different community lists: the result of the community detection algorithm, and the group information we have crawled from *Flickr*. Since our proposed system detects topic-sensitive communities based on the trust among different users, it is intuitive that our method is capable to detect communities with random cardinality. Other algorithms, such as MODULE [5], try to detect communities purely based on the social graph structure, and they prefer the communities with larger size. Further, it is difficult to incorporate additional available information being considered in our proposed method into these algorithms. To investigate the performance of our proposed community detection method, we use the experiment setting for deciding the best layer. The comparison result is reported in Table I.

TABLE I
EVALUATION ON DIFFERENT COMMUNITY DETECTION ALGORITHMS

Approaches	$\mu = 10$		$\mu = 20$		$\mu = 30$		$\mu = 50$	
	NMI	JI	NMI	JI	NMI	JI	NMI	JI
MODULE	0.5379	0.5230	0.5628	0.5804	0.6207	0.6332	0.7047	0.6938
SHRINK	0.6137	0.5821	0.6277	0.6035	0.6458	0.6217	0.6928	0.6736
CONGA	0.6534	0.6412	0.6803	0.6219	0.6980	0.6434	0.7121	0.7295
EAGLE	0.6607	0.6329	0.6947	0.6608	0.7233	0.6954	0.7332	0.7001
PerCial	0.7235	0.7120	0.7433	0.7395	0.7588	0.7602	0.7625	0.7743

The results in Table I provides us a couple of implications: (i) The modularity-based methods, such as MODULE, perform

poorly when small communities are dominant in the dataset, which results from the resolution limit of the modularity; (ii) The methods that aim to detect overlapping communities perform better than the hard partition-based ones on our *Flickr* dataset; and (iii) Our proposed community detection method – PerCial – is capable to deal with communities with different cardinalities since we not only consider the social graph structure, but also take into account the topic-sensitive information and activities.

D. A User Study

In order to verify the efficacy of the community hierarchy personalization, we implement a prototype system and conduct a user study on it. Specifically, 50 volunteers are hired to experience PerCial when they are trying to post new content or upload new photos. We recommend communities to these volunteers based on their profiles and the objects being posted. For evaluation purpose, we define several indexes to measure the satisfaction of user experience. Each experience index is rated by the volunteers in a range of 1 to 5, where 1 – “Execrable”, 2 – “Below Average”, 3 – “Average”, 4 – “Above Average”, and 5 – “Exceptional”. The experience indexes include: (i) The response time of recommending communities for privacy assignment; (ii) The relatedness of the recommended communities to their interest; and (iii) The diversity of the recommended community list.

We collect all the volunteers’ ratings on these three experience indexes. To analyze the investigation result, we plot the number of users with different ratings on the three indexes, and get a histogram as shown in Figure 3. From the result, we observe that all the three aspects of PerCial can satisfy the requirements of the majority users.

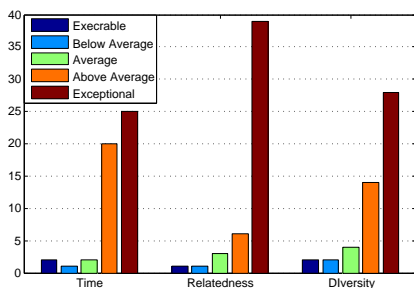


Fig. 3. User experience on different indexes.

VII. CONCLUSION

In this paper, we proposed a privacy assignment system for social sharing that utilizes the abundant resources in social object networks. Our system provides online users a flexible platform for social object sharing, and helps them assign privacy preference by automatically generating topic-sensitive communities that they might be interested in. For future work, we plan to explore trust relations between online users by taking into account the users’ social influence. In addition, how to maintain social sharing status for each user is also deserved to be exploited.

VIII. ACKNOWLEDGEMENT

The work is partially supported by Xerox University Affairs Committee (UAC) grant.

REFERENCES

- [1] C. Dwyer, S. Hiltz, and K. Passerini, “Trust and privacy concern within social networking sites: A comparison of facebook and myspace,” in *Proceedings of the 13th Americas Conference on Information Systems*, 2007.
- [2] R. Gross and A. Acquisti, “Information revelation and privacy in online social networks,” in *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*. ACM, 2005, pp. 71–80.
- [3] M. Madden and A. Smith, “Reputation management and social media: How people monitor their identity and search for others online,” *Pew Internet and American Life Project*, 2010.
- [4] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu, “Community detection in large-scale social networks,” in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM, 2007, pp. 16–25.
- [5] M. Newman, “Detecting community structure in networks,” *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 321–330, 2004.
- [6] S. Gregory, “An algorithm to find overlapping community structure in networks,” *Knowledge Discovery in Databases: PKDD 2007*, pp. 91–102, 2007.
- [7] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.
- [8] H. Shen, X. Cheng, K. Cai, and M. Hu, “Detect overlapping and hierarchical community structure in networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 8, pp. 1706–1712, 2009.
- [9] T. Evans and R. Lambiotte, “Line graphs, link partitions, and overlapping communities,” *Physical Review E*, vol. 80, no. 1, p. 016105, 2009.
- [10] Y. Ahn, J. Bagrow, and S. Lehmann, “Link communities reveal multi-scale complexity in networks,” *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.
- [11] L. Tang, X. Wang, H. Liu, and L. Wang, “A multi-resolution approach to learning with overlapping communities,” in *Proceedings of the First Workshop on Social Media Analytics*. ACM, 2010, pp. 14–22.
- [12] D. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation,” *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [13] G. Nemhauser, L. Wolsey, and M. Fisher, “An analysis of approximations for maximizing submodular set functions,” *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.
- [14] J. Li, L. Li, and T. Li, “Mssf: a multi-document summarization framework based on submodularity,” in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*. ACM, 2011, pp. 1247–1248.
- [15] L. Li, D. Wang, T. Li, D. Knox, and B. Padmanabhan, “Scene: a scalable two-stage personalized news recommendation system,” in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*. ACM, 2011, pp. 125–134.
- [16] S. Khuller, A. Moss, and J. Naor, “The budgeted maximum coverage problem,” *Information Processing Letters*, vol. 70, no. 1, pp. 39–45, 1999.
- [17] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, “Cost-effective outbreak detection in networks,” in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2007, pp. 420–429.
- [18] J. Huang, H. Sun, J. Han, H. Deng, Y. Sun, and Y. Liu, “Shrink: a structural clustering algorithm for detecting hierarchical communities in networks,” in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 219–228.
- [19] A. Lancichinetti, “Community detection algorithms: a comparative analysis,” *Physical Review E*, vol. 80, no. 5, p. 056117, 2009.
- [20] P. Tan, M. Steinbach, V. Kumar *et al.*, *Introduction to data mining*. Pearson Addison Wesley Boston, 2006.
- [21] J. Dunn, “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters,” *Cybernetics and Systems*, vol. 3, no. 3, pp. 32–57, 1973.