Interest-Matching Information Propagation in Multiple Online Social Networks

Yilin Shen, Thang N. Dinh, Huiyuan Zhang, My T. Thai
Department of Computer and Information Science and Engineering, University of Florida, USA
{yshen, tdinh, huiyuan, mythai}@cise.ufl.edu

ABSTRACT

Online social networks have become an imperative channel for extremely fast information propagation and influence. Thus, the problem of finding a minimum number of seed users who can eventually influence as many users in the network as possible has become one of the central research topics recently. Unfortunately, most of related works have only focused on the network topologies and largely ignored many other important factors such as the users' engagements and the negative or positive impacts between users. More challengingly, the behavior of information propagation across multiple networks simultaneously remains an untrodden area and becomes an urgent need. Our work is the first attempt to tackle the above problem in multiple networks, considering these lacking important factors. In order to capture the users' engagement, we propose to targeting the set of interest-matching users whose interests are similar to what we try to propagate. Then, we develop our Iterative Semi-Supervising Learning based approach to identify the minimum seed users. We validate the effectiveness of our solution by using real-world Twitter-Foursquare networks and academic collaboration multiple networks.

Categories and Subject Descriptors

G.4 [**Mathematical Software**]: Algorithm design and analysis

General Terms

Algorithms, Experimentation, Performance

Keywords

Multiple Online Social Networks, Information Propagation, Interest Prediction, Iterative Semi-Supervised Learning

1. INTRODUCTION

The rapid growth of Online Social Networks (OSNs), such as Facebook, Twitters, and LinkedIn, has made them become one of the most important channels for fast information propagation and influence. Many efficient solutions have been introduced in the literature for the information propagation problem, focusing on finding the smallest set of selected seed users with the biggest influence. However, these solutions heavily depend on the network topologies and ignore the crucial factors as discussed next.

First of all, the above studies assume that a person is influenced, i.e., adopts a product, if many of his friends believe in the product, thus he will continue influencing others. However, consider an advertisement of a sport video game to a person who has absolutely no interest in it, that is, he has no engagement into the advertisement and will not propagate it further. Thus, one of the key factors for an effective propagation is to targeting the interest-matching users, whose interests are similar to what we try to propagate. Secondly, the existing works also ignore the negative and positive impacts between users. One word from a very trusted friend is worth more than thousand words from others. The challenges of studying information propagation not only include the above lacking factors but also advance to a higher level of difficulty when information can be propagated across several networks simultaneously. Due to this feature, a user can post his message in multiple OSNs simultaneously and influence many friends on both networks at the same time. Unfortunately, the behavior of this type of information propagation still remains unexplored.

In this paper, we investigate the information propagation problem in multiple networks regarding the above mentioned lacking factors, called *Minimum Seed multi-Information Propagation* (MSIP), which seeks for the minimum number of seed users who can spread the information to as many as interest-matching users in multiple OSNs. This is the first attempt to study the information propagation problem in multiple networks considering other important properties: interest-matching users, user negative/positive relations, and message timeliness. We propose a graph-based model to couple the multiple networks together while still carry those properties intact. As one of the first work addressing the interest-matching users, we provide a semi-supervised learning based prediction approach which can predict a large amount of users' interests based on a very limited available information. After the prediction of interest-matching users, we propose two solutions based on a novel idea of iterative semi-supervised learning to the MSIP problem and its variants. The performance of our approaches is evaluated on two datasets, Twitter-Foursquare and academic collaboration multiple networks, in which we use online APIs to crawl their topologies and users' interests.

Organization: Section 2 presents the network model, propagation model, and problem definitions. The solution to the MSIP problem are proposed in Section 3. In Section 4, we

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

describe the datasets and evaluate the performance of our proposed approach on them.

Related Works: The k-top nodes problem in social networks was first studied by Domingos *et al.* [6]. Afterwards, many researches sprung up to either maximize the influence or minimize the seed nodes to propagate information with respect to social network topologies [9, 13, 15]. A lot of works are proposed afterwards to focus on the variations of this problems and the improved approaches [4, 5, 10, 13]. Taking the social relationships into account, Tang *et al.* [16] showed its impact on information propagation problem.

2. PRELIMINARIES

Social Network Model: Consider k multiple OSNs. Each network consists of users and their relations — links formed by people indicating their friendships. All links are weighted, which reflect the degree of user relations. The weight can be either positive or negative to express the degree of their agreements and disagreements. A user is referred to as a *crossing user* if he appears in more than one network and all other users are called *non-crossing users*. In addition, each user has a set of interest tags such as sports and music. As mentioned earlier, most users do not have their interest tags available online due to their concerns about privacy and security. Also, each propagation message is assigned some interest tags based on its content. Note that the interest tag of a message can be obtained using the unsupervised learning technique [8] based on its content. In this paper, we formulate the k multiple social networks as k weighted directed graphs $G_1 = (V_1, E_1, w_1), G_2 = (V_2, E_2, w_2), \ldots, G_k = (V_k, E_k, w_k)$ where each graph G_i has $|V_i| = n_i$ users (nodes), $|E_i| = m_i$ links (edges) and an adjacency matrix $\mathbf{A}_i$. A directed link (u, v) represents that user u is following user v, and thus u is called a *follower neighbor* of v and v is a *hub neighbor* of u. Each element w^i_{uv} in $\mathbf{A}_i$ reflects the weight of link (u, v) in G_i, which can be either positive or negative. There are a small number of users, each of whom has a set of corresponding interest tags $I_u \in \mathcal{I}$ available online, are called *interest-available users*, where $\mathcal{I}$ is the entire interest set. Also, each message m has its interest tag as $I_m \in \mathcal{I}$.

Linear Threshold Propagation Model: The linear threshold propagation (LTP) model [9] intuitively represents the latent tendencies of users to be influenced. In this widely-accepted propagation model, apart from the *seed users* who start to spread the information, a user is called *influenced* if he obtains the information from more than some threshold of his friends with the same opinions as his. Starting with an initial set of active nodes (seed users) A_0, the dynamics of information propagation unfold round by round as follows. The propagation process is deterministically in discrete rounds: in round t, all nodes that were active in round $t-1$ remain active, and we activate any node u for which the total weight of its active neighbors is at least θ_u fraction of all its neighbors, i.e., $\sum_{j \in N_i(u) \cap \Psi_i^{(t-1)}} w^i_{uj} \geq \theta_u \sum_{j \in N_i(u)} w^i_{uj}$, where $N_i(u)$ denotes the neighbors of u in network G_i, $\Psi_i^{(t-1)}$ is the set of influenced users after $t-1$ rounds in network G_i, and the threshold $\theta_u \in [0, 1]$. Using this LTP model, we note that the different opinions and foes will slow down the information propagation. The process goes on for a maximum number of d rounds and a vertex once becomes active will remain active until the end.

Problem Definition and Notations: To study the information propagation in multiple social networks, considering the user interests, their relations, and message timeliness, we formulate the following new problem, *Minimum Seed multi-Information Propagation* (MSIP).

PROBLEM 1 (MSIP). *Given k multiple social networks $G_1 = (V_1, E_1, w_1), \ldots, G_k = (V_k, E_k, w_k)$, a small number of interest-available users $U \subset \cup_{1 \leq i \leq k} V_k$, a message m on topic $I_m \in \mathcal{I}$ and k influence threshold vectors, i.e., $\mathbf{\Theta}_i$ for each network G_i, MSIP problem asks for a minimum set of seed users such that at least β fraction of all users can be influenced after d rounds according to the LTP model, i.e., $|\bigcup_i \Psi_i^{(d)}| \geq \beta |\bigcup_i V_i|$, and maximizes the influenced users interested in I_m.*

Solving the above problem requires us to address the following two main challenges: (1) How to predict the interest-matching users, who are interested in the message? (2) What role do the crossing users play in spreading information across multiple networks?

We introduce some notations which will be used in the rest of paper. During the diffusion of a message m, Table 1 lists the appearance vectors associated with each user u and the label vectors associated with each network G_i. For simplicity, we concatenate these label vectors in Table 1 for all k networks and further define $\mathbf{y}^{(t)} (= ((\mathbf{y}_1^{(t)})^T, \ldots, (\mathbf{y}_k^{(t)})^T)^T)$, $\mathbf{y}$, ℓ, π and $\imath$ of dimension $N \times 1$ and $N = \sum_i n_i$ as the corresponding label vectors for the multiple networks. In addition, the threshold vector for all networks is also defined as $\mathbf{\Theta} = (\mathbf{\Theta}_1^T, \ldots, \mathbf{\Theta}_k^T)^T$ where $\mathbf{\Theta}_i$ is composed of the threshold θ_u of each user u in network G_i.

3. TSNL APPROACH

Our proposed *Total Seed Nodes Learning* (TSNL) approach to the MSIP problem consists of three main steps to tackle the challenges mentioned above:

(1) *Network Coupling* is to construct a new network $\mathcal{G}_\mathcal{T}$ based on the k multiple networks, in which the impacts of crossing users can be correctly presented.

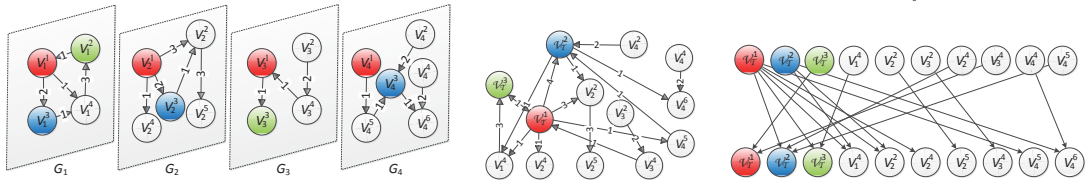
(2) *Interest-matching Users Prediction* with respect to message m can be realized by minimizing the penalty function $\mathcal{P}_\mathcal{T}$ on $\mathcal{G}_\mathcal{T}$ regarding incorrect predictions on the above coupled network. In order to take advantage of the interest-available users in the prediction, we design the penalty function $\mathcal{P}_\mathcal{T}$ based on the idea of graph-based *Semi-Supervised Learning* (SSL), which follows from the regularization framework. Thus, $\mathcal{P}_\mathcal{T}$ is composed of two items: on one hand, the predicted interest-matching users should be close to the interest-available users; on the other hand, the penalties of wrong predictions should be minimized. These two items are referred to as *loss function* and *regularizer* respectively.

(3) *Seed Users Identification* is based on a novel idea, *Iterative Semi-Supervised Learning* (ISSL), which imposes the predicted interest-matching users to be influenced in the end. The idea is to minimize the cost function composed of three items: *iteration function*, loss function, and regularizer (instead of using traditional two items). Specifically, the purposes of these three items are as follows: the loss function forces the influenced users close to the interest-matching users after d rounds; the regularizer aims to minimize the number of seed users; the iteration function guarantees that the identified seed users in previous iterations will not be replaced by other users. Starting from the empty set, the seed users can be iteratively identified by solving the cost function using the ISSL approach.

Networks Coupling: Intuitively, the goal of coupling multiple networks together is to reinforce the effects of crossing users. As an example illustrated in Fig. 1(b), the idea is as follows: in the coupled network, we merge each crossing node in all networks into one node. The new neigh-

	Name	Description
η_u^i	Appearance Vectors	each η_u^i has dimension n_i and the j^{th} entry in η_u^i is 1 if u appears as user j in network G_i
w_u	Appearance Times	the number of networks user u appears
$\mathbf{y}_i^{(t)}$	Influenced Label Vector after Round t	the j^{th} entry $y_{ij} = 1$ if user j is influenced after t rounds and $y_{ij} = -1$ otherwise
$\mathbf{y}_i$	Seed Label Vector	equivalent to $\mathbf{y}_i^{(0)}$ for simplicity
ℓ_i	Interest-Available Label Vector	the j^{th} entry $\ell_{ij} = 1$ if user j in G_i is known to be interested in I_m, $\ell_{ij} = -1$ if j does not, and the j^{th} entry $\ell_{ij} = 0$ when the interest tag of j is not available
π_i	Interest-Prediction Label Vector	user j is predicted to be more interested in I_m if the j^{th} entry ℓ_{ij} is more approaching to 1 (each element in π_i lies in $[-1, 1]$)
$\imath_i$	Interest-Matching Label Vector	the j^{th} entry $\imath_{ij} = 1$ if user j is predicted to be interest-matching and $\imath_{ij} = -1$ otherwise

Table 1: Notations associated with each user u and each network G_i



(a) An Instance of multiple networks (b) Coupled Graph $\mathcal{G}_\mathcal{T}$ (c) Bipartite Graph Γ_T

Figure 1: TSNL on an instance of 4 multiple networks (G_1, G_2, G_3, G_4) (All crossing nodes are colored)

bors of each crossing node, both incoming and outgoing, include its neighbors in all the k networks. If a link (between two crossing nodes) appears on more than one network, its weight in the coupled network is the sum of all its weights on these networks. We denote the coupled network as $\mathcal{G}_\mathcal{T} = (\mathcal{V}_\mathcal{T}, \mathcal{E}_\mathcal{T}, \mathcal{W}_\mathcal{T})$, in which each crossing node appears only once in $\mathcal{G}_\mathcal{T}$. In addition, $\mathcal{W}_\mathcal{T}$ represents two types of weights: node weight $\mathcal{W}^u$ and link weight $\mathcal{W}_{uv}$. Accordingly, the label vectors ℓ, π and $\imath$ are mapped to ℓ^φ, π^φ and $\imath^\varphi$ on $\mathcal{G}_\mathcal{T}$ respectively.

Interest-matching Users Prediction: In order to predict interest-matching users, we design a penalty function $\mathcal{P}_\mathcal{T}$ on a constructed bipartite graph Γ_T based on the coupled graph $\mathcal{G}_\mathcal{T}$. The incorrect predictions are brought to its knees by minimizing this penalty function $\mathcal{P}_\mathcal{T}$, which contains the penalties of incorrect predictions with respect to both hub users and follower users. Thus, the following bipartite graph is constructed in the first place to differentiate hub users and follower users.

Bipartite Graph Γ_T: Since our purpose is to differentiate hub users and follower users, we construct a bipartite graph $\Gamma_T = (\mathcal{H}_\mathcal{T}, \mathcal{F}_\mathcal{T}; \mathcal{L}_\mathcal{T})$ on the coupled network $\mathcal{G}_\mathcal{T}$, where $\mathcal{H}_\mathcal{T}$ and $\mathcal{L}_\mathcal{T}$ denote the set of *hub users* and *follower users* respectively. Roughly speaking, *hub users* are followed by users having the relevant interests and *follower users* are those following at least one hub user. The mutual relationship between hub and follower users is: a good hub user is followed by many good followers and a good follower follows many good hubs. As shown in Fig 1(c), we use a simple method [12] by selecting the follower users to be the users following at least one user and the hub nodes as the users having at least one follower. That is, by defining $d_T^+(u)$ and $d_T^-(u)$ as the in-degree (number of u's followers) and out-degree (number of users u follows) of node u on $\mathcal{G}_\mathcal{T}$, we have $\mathcal{H}_\mathcal{T} = \{u : d_T^+(u) > 0\}$ and $\mathcal{F}_\mathcal{T} = \{u : d_T^-(u) > 0\}$. Note that $\mathcal{H}_\mathcal{T} \cup \mathcal{F}_\mathcal{T} = \mathcal{V}_\mathcal{T}$, $\mathcal{L}_\mathcal{T} \subseteq \mathcal{H}_\mathcal{T} \times \mathcal{F}_\mathcal{T}$ and a user can be both a hub user and a follower user.

Penalty Function $\mathcal{P}_\mathcal{T}$: Recall that we aim to both impose prediction matching the interest available users and minimize the incorrect predictions for other users. Borrowing the idea of semi-supervised learning, we define a penalty function $\mathcal{P}_\mathcal{T}$ as $\mathcal{P}_\mathcal{T} = \Omega_P \|\pi^\varphi - \ell^\varphi\|^2 + P$ with some large constant Ω_P. The first item $\Omega_P \|\pi^\varphi - \ell^\varphi\|^2$, the loss function, forces the interest-matching users close to the interest-available users; The second item P, the regularizer, casts the penalties of the incorrect predictions. Clearly, after minimizing $\mathcal{P}_\mathcal{T}$, a user is more interested in I_m if its entry of the interest-prediction label vector π^φ is closer to 1 (less interested if closer to -1) according to the definition of the interest-available label vector ℓ^φ. In terms of the regularizer P, its formulation takes into account both follower user penalties P_f and hub user penalties P_h on the above bipartite graph Γ_T, i.e., $P = \lambda P_f + (1 - \lambda) P_h$, where some experience value as normalized parameter $\lambda \in (0, 1)$, which will be evaluated in the experiment.

The idea of follower user penalty P_f is to consider the sum of penalties for all pairs of follower users. For each pair of followers, the penalty between them is further measured by the sum of their similarity penalties with respect to each hub user. Particularly, given a hub user, the similarity penalty can be measured using the weights of links from the follower users to him. Therefore, the hub user penalty P_h is $P_f = \sum_{u,v \in \mathcal{F}_\mathcal{T}} \sum_{h \in \mathcal{H}_\mathcal{T}} p_{uv}^h$, where the penalty function p_{uv}^h for any two follower users u, v and some hub user h is threefold: (1) both u and v have the same interest as h's, but their interest-prediction labels π_u^φ and π_v^φ are different; (2) only one of u and v has the same interest as h's but they have the same interest-prediction labels, i.e., $\pi_u^\varphi = \pi_v^\varphi$. The underlying intuition is that the two follower users are more likely to have the same interest if they closely follow some hub users; (3) when neither u nor v has the same interest as h's, there is no penalty for any classification of u and v due to the difficulty to measure the penalty in this case.

Since the link weights represent the degree of user relations and the consistence of their opinions, we use $\mathcal{W}_{uh}\mathcal{W}_{vh}$ on the coupled graph $\mathcal{G}_\mathcal{T}$ as the interest similarity of u and v to h. Thus, the penalty p_{uv}^h can be written as

$$p_{uv}^h = \frac{\chi_{uv}^h \mathcal{W}_{uh}\mathcal{W}_{vh}}{\sum_{l \in \mathcal{F}_\mathcal{T}} \mathcal{W}_{lh}} \left(\frac{\pi_u^\varphi \sqrt{\mathcal{W}^u}}{\sqrt{\sum_{l \in \mathcal{H}_\mathcal{T}} \mathcal{W}_{ul}}} - \frac{\chi_{uv}^h \pi_v^\varphi \sqrt{\mathcal{W}^v}}{\sqrt{\sum_{l \in \mathcal{H}_\mathcal{T}} \mathcal{W}_{vl}}} \right)^2$$

where χ_{uv}^h is an indicator parameter representing the three cases mentioned above as

$$\chi_{uv}^h = \begin{cases} -1 & \mathcal{W}_{uh}\mathcal{W}_{vh} \leq 0 \\ 1 & \mathcal{W}_{uh} \geq 0, \mathcal{W}_{vh} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Consider the various interests of hub user h. We normalize the interest similarity by its relations to all hubs, i.e.,

$\sum_{l\in\mathcal{F}_\mathcal{T}} \mathcal{W}_{hl}$, due to the intuition that the penalty could be reduced when there are a great number of followers of h. Similarly, π_u^φ and π_v^φ are normalized with respect to their weights and their relations to all followers. The hub user penalties P_h can be derived similarly.

Seed Users Identification: According to the interest-prediction label vector $\pi^{\varphi *}$ (obtained by minimizing $\mathcal{P}_\mathcal{T}$), in order to propagate the information to as many as interest-matching users, we first determine the interest-matching users by selecting the first largest $\beta|\mathcal{V}_T|$ users in $\pi^{\varphi *}$ and setting their entries in $\imath^\varphi$ to be 1. Then, we propose a novel *Iterative Semi-Supervised Learning* (ISSL) approach finding the minimum number of seed users iteratively to influence these interest-matching users to the greatest extent. As introduced at the beginning of this section, the seed users can be identified by minimizing the cost function $\mathcal{C}_\mathcal{T}$ which consists of three items: (1) The *loss function* $\Omega_I\|\imath - \mathbf{y}^{(d)}\|^2$ forces the influenced users after d rounds to be the interest-matching users labeled by $\imath$ which is injected from $\imath^\varphi$; (2) The *regularizer* $(\mathbf{y}+\mathbf{e})^T diag(\mathbf{w})^{-1}(\mathbf{y}+\mathbf{e})$ denotes the number of seed users in which each crossing user is accounted only once since $\mathbf{w} = (w_1, \ldots, w_N)$ represents the appearance times of each user; (3) The *iteration function* $\mu\|\mathbf{y}^* - \mathbf{y}\|^2$ containing the current solution $\mathbf{y}^*$ after each iteration forces the identified seed users remaining as they are in next iteration. By defining $\delta = (\frac{1}{w_1}, \ldots, \frac{1}{w_N})$ with respect to the weight vector $\mathbf{w}$, the objective function can be simplified to be Linear Programming (LP) formulation after relaxing all elements in $\mathbf{y}^t$ to $[-1, 1]$, along with the four constraints.

$$
\begin{array}{ll}
\min & \mathcal{C}_\mathcal{T} = -\Omega_I\langle\imath, \mathbf{y}^{(d)}\rangle - \mu\langle\mathbf{y}^*, \mathbf{y}\rangle + \langle\delta, \mathbf{y}+\mathbf{e}\rangle \\
\text{s.t.} & diag(diag(\mathbf{\Theta})\mathcal{A}_\mathcal{T}\mathbf{e})\mathbf{y}^{(t)} \le \mathcal{A}_\mathcal{T}(\mathbf{y}^{(t-1)}+\mathbf{e}) - \mathbf{e} \\
& (\forall 1 \le t \le d) \\
& \mathbf{y}^{(t)} \ge \mathbf{y}^{(t-1)} (\forall 1 \le t \le d) \\
& \langle\delta, \mathbf{y}^{(d)}+\mathbf{e}\rangle \ge 2\beta\langle\delta, \mathbf{e}\rangle \\
& \langle\eta_u, \mathbf{y}^{(t)}\rangle \ge w_u\langle\eta_u^i, \mathbf{y}_i^{(t)}\rangle \\
& (\forall 1 \le t \le d, 1 \le i \le k, u \in V_i)
\end{array} \quad (1)
$$

where the three items in $\mathcal{C}_\mathcal{T}$ correspond to loss function, iteration function and regularizer respectively, and $\mathbf{e} = (1, \ldots, 1)^T$ is a uniform vector, μ and Ω_I are some large constants (we will further evaluate in the experiment). The $\mathcal{C}_\mathcal{T}$ in the above LP is derived from the fact that $\|\imath\|^2 = \|\mathbf{y}\|^2 = \|\mathbf{y}^{(d)}\|^2 = \|\mathbf{y}^*\|^2 = N$ (can be omitted) since all elements in these vectors are either 1 or -1. The first constraint ensures that the user is labeled as influenced in round t iff there are at least θ_u fraction of the total link weights of u's neighbors at $t-1$. The second one imposes that a node keeps its status in the following rounds once it is influenced. The third one guarantees that at least β fraction of users are influenced after d rounds. The last constraint is to make sure that a crossing user is influenced if he has been influenced in at least one network.

Based on this LP formulation, we identify the seed users as follows: Starting with an empty set, the seed users can be identified iteratively by minimizing the cost function $\mathcal{C}_\mathcal{T}$ in each iteration until the solution is feasible; that is, at least β fraction of users can be influenced after d rounds. Moreover, in each iteration, let $\mathbf{y}_s$ be the solution of the LP (1), we first set the previous identified seed users to be 1 in $\mathbf{y}_s$ and identify all positive elements in $\mathbf{y}_s$ to be the corresponding new seed users. If $\mathbf{y}_s$ is the same as the previous $\mathbf{y}^*$, we pick the largest non-one element in $\mathbf{y}_s$ to identify it as a new seed user by rounding the corresponding element in $\mathbf{y}_s$ to 1 and in the meanwhile round all other elements to -1.

Table 2: Multiple Networks Dataset

Network	Users	Links	Crossing Users
Twitter	48,092	16,304,712	4,446
Foursquare	44,896	1,664,402	
Condensed Matter	40,421	175,693	(CM-HET) 2,875
High-Energy Theory	8,361	15,751	(HET-NS) 89
Network Science	1,589	2,742	(NS-CM) 541

4. EXPERIMENTAL EVALUATION

In this section, we first describe the two real-world network datasets, *Twitter-Foursquare Multiple Networks* and *Academic Collaboration Multiple Networks*, and our crawling approaches. Then the effectiveness of both interest-matching predictions and seed identifications are evaluated on these two datasets for our proposed TSNL approach.

Dataset Description:

Twitter-Foursquare Multiple Networks (TFMN) is composed of two popular social networks, Twitter and Foursquare, in which the crossing users are those who connect their Twitter and Foursquare accounts. To obtain the Twitter network, we apply the unbiased sampling approach [7] to sample a portion of Twitter network from the complete Twitter network, which is provided by Kwak *et al.* [11]. Then, we link the corresponding users together as crossing users with the aid of the provided Twitter usernames in the their Foursquare accounts via Foursquare API v1. Finally, we further use this Foursquare API to obtain the users and links within two-hop neighbors of these crossing users. The weight of each link in Twitter is inferred from [17], 467 million tweets posted by 17 million users from June to December in 2009, using the frequency of tweets between each pair of users. In Foursquare, due to the lack of message dataset, we set the weights for each link to be 1. Furthermore, we associate users with various interests retrieved from the Klout API [1] (for those available in Klout), which measures the impact of users on various topics based on the content they created. Among all the interests available in Klout, we map more than 2500 of them, covering a wide range of topics. In our experiment, we evaluate our approaches using three popular tags, i.e., "iPhone", "travel" and "food" and report the average performance on these interest tags.

Academic Collaboration Multiple Networks (ACMN) is composed of three academic collaboration networks compiled by Newman [3]: (1) *Condensed Matter Collaboration Network (CM)*, (2) *High-Energy Theory Collaboration Network (HET)*, and (3) *Co-authorship Network in Network Science (NS)*. Due to the interdisciplinary nature of these research areas, many authors are identified in more than one of these networks, i.e., an author who has published papers in both fields of High Energy Theory and Network Science is a crossing user. To determine the research interests of users, we use Mendeley Web (for those available in Mendeley Web), an online research network on research papers and trends, to classify the research sub-areas. Consider the most popular 25 tags for documents in related areas via Mendeley API [2], such as community structure, solar cells, laser etc. The authors, according to their names, can be mapped to the detailed documents corresponding to these sub-areas. Our evaluation is averaged by the three above interests. The details of these two datasets are illustrated in Table 2.

Performance of TSNL Approach:

Interest-matching User Prediction: To measure the performance of interest-matching user prediction, given a set of interest-available users in each dataset, we select a subset of them as an input and assume that the rest of them don't have the interest profile available. Then, we apply our prediction approach and further compare our predicted result

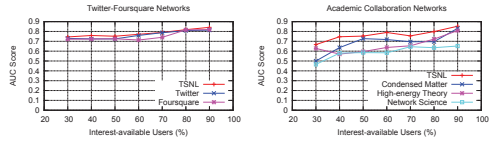


Figure 2: Interest-matching User Prediction

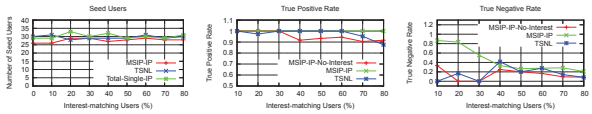


Figure 3: Small-scale Sampling Multiple Networks

back with the ground truth, i.e., the given set of interest available users, using the metric of AUC score. In penalty function P, we set the large constant Ω_P to be 10000 and the normalized parameter λ to be 0.7, based on the fraction of hub users among all users, by considering larger penalties of the incorrect prediction between followers.

Figure 2 reports the average AUC scores of interest-matching prediction in the above two datasets respectively. As can be seen, the AUC scores of prediction using our approaches arise with the increase of interest-available users. Even when there are only 30% interest-available users, the AUC score using TSNL reaches 0.7 in both datasets while the prediction on single network is a little bit less in Twitter-Foursquare networks and even less than 0.5 in academic collaboration networks. When given more than 60% interest-available users, our prediction is accurate since the AUC score is only a little bit smaller than 0.9. This reveals both the effectiveness of our approach and the importance of "crossing users" in multiple social networks. The underlying reason is that our approach (coupling and prediction method) takes into account that a crossing user can retweet or mention their messages on Twitter from Foursquare. That is, the interest similarity between such user and all neighbors in both networks are well reflected in the penalty function.

Seed Identification: Using the predicted interest-matching users as above, we evaluate the seed identifications on the small-scale sampling networks first in terms of the size of seed users and *true positive/negative influence rates* (TPR /TFR), which is defined as, after d rounds, the rate that interest-matching users are influenced and other users are not influenced respectively. Given different percentages of interest-matching users, we test on 100 users sampled from TFMN and ACMN using the unbiased sampling approach [7] and evaluate the average performance of 100 samplings. In our experiments, we compare our TSNL approach with the following other solutions: *MSIP-IP*: the integral/optimal solution of equation(1); *Total-Single-IP*: the sum of integral solution on all networks; and *MSIP-IP-No-Interest*: the optimal solution without considering interest-matching users. To solve the integer programming, we use a powerful matlab toolbox, YALMIP [14]. In addition, we set the large constant Ω_I and μ in cost functions $\mathcal{C}_\mathcal{T}$ and $\mathcal{C}_\mathcal{I}$ to be 10000. As revealed in Figure 4, most of the time our TSNL solution is very close to the optimal solutions of MSIP-IP (only at most 5% less than optimal) and even outperform the optimal Total-Single-IP. It implies that the existence of crossing users essentially helps to provide a more effective information diffusions by sharing their messages on multiple social networks. The interesting case is that they can even beat MSIP-IP sometimes, however, with the cost of lower TPR and TFR. The number of seed users and TPR/TNR are approaching optimal solution with the existence of more and more interest-matching users. Moreover, the integral solutions of MSIP-IP-No-Interest have comparatively lowest TPR and TNR although they can identify relatively smaller size of seed users. The true negative rate, as illustrated in Figure 4, drops sharply with the increase of interest-matching users since the users who are not interested in the message cannot be successfully avoided when more users are required to be influenced.

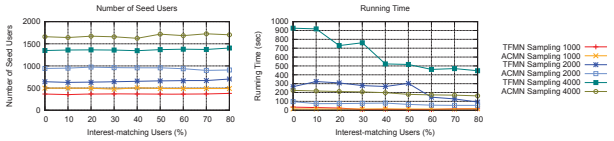


Figure 4: Case Studies on Multiple Social Networks

Figure 4 further reveals some interesting observations for different number of interest-matching users in larger sampling networks. The number of seed users almost keeps the same while the running time decreases rapidly and the TPR/TFR remain quite high after increasing the number of interest-matching users. This reveals the importance of interest-matching users again, i.e., the correct prediction of interest-matching users can not only target in the users to influence but also reduce the running time substantially.

Acknowledgement

This work is partially supported by NSF Career Award 0953284.

5. REFERENCES

[1] http://developer.klout.com/.
[2] http://dev.mendeley.com/.
[3] http://www-personal.umich.edu/ mejn/netdata/.
[4] O. Ben-Zwi, D. Hermelin, D. Lokshtanov, and I. Newman. An exact almost optimal algorithm for target set selection in social networks. In *EC '09*, pages 355–362, New York, NY, USA, 2009. ACM.
[5] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *KDD '09*, pages 199–208, New York, NY, USA, 2009. ACM.
[6] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD '01*, pages 57–66, NY, USA, 2001. ACM.
[7] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in Facebook: A case study of unbiased sampling of OSNs. In *Proceedings of IEEE INFOCOM 2010*, pages 1–9. IEEE, Mar. 2010.
[8] A. Gliozzo, C. Strapparava, and I. Dagan. Investigating unsupervised learning for text categorization bootstrapping. In *Proceedings of HLT '05*.
[9] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD'03*, pages 137–146. ACM, NY, USA, 2003.
[10] M. Kimura and K. Saito. Tractable models for information diffusion in social networks. In *Principles of Data Mining and Knowledge Discovery*, pages 259–271, 2006.
[11] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of WWW '10*, pages 591–600, New York, NY, USA, 2010. ACM.
[12] R. Lempel and S. Moran. Salsa: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160, 2001.
[13] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD '07*, pages 420–429, New York, NY, USA, 2007. ACM.
[14] J. Löfberg. Yalmip : A toolbox for modeling and optimization in MATLAB. In *Proceedings of CACSD*, Taiwan, 2004.
[15] Y. Shen, D. T. Nguyen, and M. T. Thai. On the hardness and inapproximability of optimization problems on power law graphs. In *COCOA '09*, pages 197–211, Berlin, Heidelberg, 2010. Springer-Verlag.
[16] S. Tang, J. Yuan, X. Mao, X.-Y. Li, W. Chen, and G. Dai. Relationship classification in large scale online social networks and its impact on information propagation. In *INFOCOM*, pages 2291–2299, 2011.
[17] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 177–186, New York, NY, USA, 2011. ACM.