Influence and Similarity on Heterogeneous Networks

Guan Wang † Qingbo Hu † Philip S. Yu †*
†University of Illinois at Chicago *King Abdulaziz University, Saudi Arabia
{gwang26, qhu5, psyu}@uic.edu

ABSTRACT

In the social network research, the studies on social influence maximization and entity similarity are two important and orthogonal tasks. On homogeneous networks, social influence maximization research tries to identify an initial influential set that maximizes the spread of the information, while similarity studies focus on designing meaningful ways to quantify entities' similarities. When heterogeneous networks are becoming ubiquitous and entities of different types are related to each other, we observe the possibility of merging the two directions together to improve the performance for both of them. In fact, we found that influence values among one type of nodes and similarity scores among the other type of nodes reinforce each other towards better and more meaningful results.

Therefore, we introduce a framework that computes social influence for one type of nodes and simultaneously measures similarity of the other type of nodes in a heterogeneous network. First, we decouple the target heterogeneous network (or we call it *Influence Similarity (IS)* network) into three different parts: *Influence network*, *Similarity network* and *information tunnels* (*IT*) between them. Through *IT*, we exchange the influence scores and the similarity scores to calculate more precise similarity and influence scores in order to improve both of their qualities. The experiment results on real world data shows that our framework enables influence maximization framework to identify more influential seeds in Influence network and similarity measures to produce more meaningful similarity scores in Similarity network simultaneously.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications–Data Mining

Keywords

Influence, Similarity, Social Network

1. INTRODUCTION

Two prominent techniques of social network mining are *social influence maximization* and *entity similarity analysis*. Given a user population and their relations in an information cascade scenario, social influence maximization is centered on constructing a meaningful influence network, identifying a set of most influential nodes in the network, and maximizing the spread of influence through such a social network. Similarity analysis, as another flourishing research direction, is proposing methods for measuring nodes similarities, based on the network structure and node features. Although influence and similarity analysis could provide tools for similar applications, including information retrieval, ranking, or recommendation, they are often considered as orthogonal techniques that are studied separately, besides they are often applied to homogeneous networks. When the heterogeneous network is becoming ubiquitous, we notice that it is not only necessary but also beneficial to combine the two techniques into one framework because we can use information from one side to calibrate the other side.

Given a set of users and their influential relations, the ultimate goal of social influence maximization research is to identify limited number of influential people as seeds from which the spreading of information entity would be maximized. An Influence network of the users is first constructed according to such relations. Algorithms that depict different information cascading rules are developed on the network to explain the real cascading phenomena.

Previously, majority of such research assume that the cascading network is given and so are activation probabilities among nodes [1, 4] which is either a fix number for every node or weighted value of a node's degree. Their major focuses are under different diffusion models, e.g., independent cascade [1, 4], linear threshold [4], how to design the best algorithm to identify the "influential" users. Furthermore, they compared the performance of these algorithms on the cascading coverage, i.e., the number of final activated people in the given network. To the best of our knowledge, there are two important phenomena that are constantly overlooked by the previous research. The first one is that they seldomly demonstrate the **seed qualities** in the real network that are selected by their algorithms. The second overlooked phenomenon is the definition of activation probability. Under which circumstance would a user activate another, i.e., pass over the information, is an important factor to the whole cascading process. However, majority of other work simply assume that the probabilities are given and randomly assign these probabilities in their experiment settings. We have only found 1 publication [3] that explicitly studies how to calculate such probabilities, and their work is in a totally different problem setting. Our approach addresses the two overlooked factors by introducing similarity measures. In a heterogeneous network, the activation probability between two nodes of the same type is often related to the other type of nodes they connect to. Considering such connections would offer us a more precise way of activation probability model and a better influence maximization result. We also perform

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

a detailed seed qualities comparison to demonstrate these qualities do get improved by our model.

The research community has proposed numerous similarity measures, symmetrical or asymmetrical, for nodes in social networks, which consider node features, link features, and other semantic features. Also, the similarity measures have been defined over either homogeneous or heterogeneous networks among the same type of nodes or different types of nodes. A simple clue for designing meaningful similarity measures is to customize the definition and consider more information according to application's scenario.

In our case, due to the heterogeneous nature of the network, when computing the similarity of one type of nodes, we should put the influence of the other type of nodes into the formulation. This leads to an asymmetric similarity formulation. This approach has never been studied before. As we analyzed in the above subsection, introducing similarity to influence maximization in a heterogeneous network could be beneficial to many key aspects on the influence maximization. Further more, similarity measure gets more customized information from influence maximization side, which should potentially be beneficial too. Our experiment results confirm this **mutual beneficial relation**.

Similarity and influence computation can benefit each other, which motivates us to study how to effectively combine them together in one framework. Our technique adopts reinforcement scheme on the top of heterogeneous network de-coupling. To be more specific, we first define a special bi-typed heterogeneous network as Influence Similarity (IS) network; We then de-couple its different types of nodes into two homogeneous networks based on their relations, and we do the maximization of social influence spreading on the Influence network and expectation of similarity measures on the Similarity network. However, the two networks are not totally separated. There is a *latent tunnel* connecting them for the sake of delivering information back and forth to improve the performance influence and similarity analysis.

We summarize our major contributions as follows.

- We propose a new angle of treating influence maximization and similarity computation together. To our best knowledge, our work is the first to explicitly explore how to make use of both techniques together to analyze a heterogeneous network in a more comprehensive way.
- We study the mutual improvements of influence maximization and similarity computation on each other. An iterative algorithm with optimization on decoupled heterogeneous networks is tailored for this reinforcement relationship.
- We demonstrate its effectiveness through real world social network analysis. Our method outperforms state of the art in influence maximization and similarity computation when they are performed separately.

2. INFLUENCE-SIMILARITY COMPUTATION FRAMEWORK

2.1 IS network and its de-coupling

IS network is a special type of heterogeneous network with edge features of different practical meanings for different edge types. We have observed that it is generic enough to capture important relations for different types of nodes and explore the hidden reinforcement between influence and similarity. First, we formally define the generic **I**nfluence **S**imilarity network. We then explain necessary concepts related to our model in the next subsection.

DEFINITION 2.1. *(Influence similarity Network) An IS network is a directed heterogeneous network $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{F})$ of two different types of nodes, four types of edges with associated edge features. For ease of presentation, let $\mathcal{V}_\mathcal{X}$ be the set of nodes we want to study influence on and $\mathcal{V}_\mathcal{Y}$ the set of the type of nodes for similarity research, where $\mathcal{V} = \mathcal{V}_\mathcal{X} \cup \mathcal{V}_\mathcal{Y}$. There are four types of edges E_{XX}, E_{XY}, E_{YX}, E_{YY} connecting different types of nodes, and $\mathcal{E} = E_{XX} \cup E_{XY} \cup E_{YX} \cup E_{YY}$. $\mathcal{F}$ is a feature vector associated with different types of edges. $\mathcal{F} = \mathcal{F}_\mathcal{X} \cup \mathcal{F}_\mathcal{Y}$. $\mathcal{F}_\mathcal{X} = \{f_X | \forall e_X \in E_{XX}\}$ is a vector of variables, each one of which describes the influence scores between two nodes of an edge e_X. Similarly, $\mathcal{F}_\mathcal{Y} = \{f_Y | \forall e_Y \in E_{YY}\}$ is another vector of variables for similarity scores on the other type of nodes.*

Given an IS network, it is worth noting is that it is **application dependent** on the categorization of $\mathcal{V}_\mathcal{X}$ and $\mathcal{V}_\mathcal{Y}$. On the abstract level, the goal of IS modeling is to pick up influential initial seeds to maximize the spreading of social influence on nodes $\mathcal{V}_\mathcal{X}$ and simultaneously compute the similarity of nodes $\mathcal{V}_\mathcal{Y}$, so that the results of the two tasks reinforce each other. However, before applying the model, one should fix the one type of nodes as type $\mathcal{V}_\mathcal{X}$ and another type of nodes as type $\mathcal{V}_\mathcal{Y}$, so that the categorization is meaningful to the specific application.

To fulfill our goal to obtain better results for both tasks by calibrating each other, we propose our framework in three steps. What we are introducing now is the first step, IS network decoupling. As the majority of previous research on either influence or similarity is on *homogeneous* networks, we want to first decouple IS network into two homogeneous ones and information tunnels between them.

DEFINITION 2.2. *(IS network decoupling) IS network decoupling is a mapping $\mathcal{L} : \mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{F}) \rightarrow G_X(V_X, E_X, F_X) \times G_Y(V_Y, E_Y, F_Y) \times G_{IT}(V_X \cup V_Y, E_{XY} \cup E_{YX})$, In the mapping, we have $V_X = \mathcal{V}_\mathcal{X}$, $V_Y = \mathcal{V}_\mathcal{Y}$, $E_X = \mathcal{E}_{\mathcal{X}\mathcal{X}}$, $E_Y = \mathcal{E}_{\mathcal{Y}\mathcal{Y}}$, $F_X = \mathcal{F}_\mathcal{X}$, $F_Y = \mathcal{F}_\mathcal{Y}$. $E_{XY} = \mathcal{E}_{\mathcal{X}\mathcal{Y}}$ and $E_{YX} = \mathcal{E}_{\mathcal{Y}\mathcal{X}}$.*

As seen in the above definition, in the decoupling process, we first preserve the nodes and edge structures for both influence network and similarity network. That is to separate the IS network into two by structurally removing edges $\mathcal{E}_{\mathcal{X}\mathcal{Y}}$ and $\mathcal{E}_{\mathcal{Y}\mathcal{X}}$. However, these edges are actually preserved as the information tunnels: $G_{IT}(V_X \cup V_Y, E_{XY} \cup E_{YX})$. We call it information tunnel, since the similarity and influence information can transit through these connections. One should be aware that $G_{IT}(V_X \cup V_Y, E_{XY} \cup E_{YX})$ is a bipartite graph, which only represents the connection between V_X and V_Y. Thus, its edges have no weights. Moreover, E_{XY} and E_{YX} are complimentary to each other. If an edge $a_X b_Y$ belongs to E_{XY}, $b_Y a_X$ is in E_{YX}. For example, in a paper-author network, that a paper is "written" by an author actually is the same as the author "writes" the paper. Figure 1 illustrates such decoupling process. In the following subsections, we will model the information passing in details in coordination with the influence similarity reinforcement.

2.2 Maximizing Influence Spreading on Influence Network

Similar to the state-of-the-art research on the social influence maximization, our task on the Influence Network is to identify k seed nodes such that the spreading of information can be maximized. The influence of z_X on w_X is often represented by the probability that z_X *activates* w_X, in another word, information is passed from z_X to w_X. Here we follow the classic independent cascade (IC) model to simulate the information diffusion. However, instead of having the activation probabilities on every pair

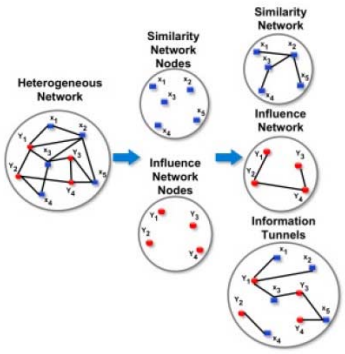

Figure 1: Edge Transformation in IS Decoupling

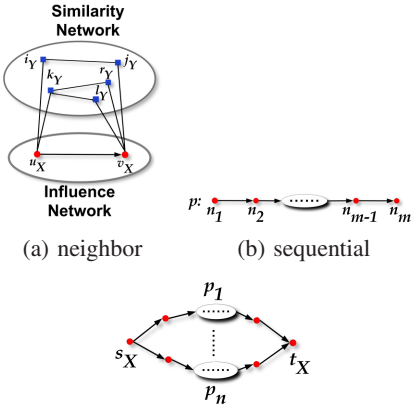

Figure 2: Activation Paths

of nodes are the same value drawn from uniform distribution, we come up with a more fine-grained activation probability definition. Therefore, our design is the same diffusion process with IC model with finer-grain direct neighbor activation probability.

Let $h(u_X, v_X)$ denote the probability that u_X can activate v_X in the independent cascade model. Figure 2 shows how we define $h(u_X, v_Y)$ step by step, where each step is explained as follows.

- When two nodes u_X and v_X are neighbors (Figure 2(a)) in the Influence network $G_X(V_X, E_X, F_X)$ (where nodes are circles), the activation probability from u_X to v_X is obtained by considering the similarity scores of their connected nodes in the similarity network (where nodes are squares). We suppose the similarity of i_Y and j_Y is $g(i_Y, j_Y)$ (formally defined later), the similarity of k_Y and r_Y is $g(k_Y, r_Y)$, and the similarity of k_Y and l_Y is $g(k_Y, l_Y)$. We consider u_X can influence v_X through $i_Y j_Y$, $k_Y r_Y$ and $k_Y l_Y$ independently. Thus $h_1(u_X, v_X) = 1-(1-g(i_Y, j_Y))(1-g(k_Y, r_Y))(1-g(k_Y, l_Y))$. More generally, if $u_X v_X \in E_X$, we define

$$h_1(u_X, v_X) = 1 - \prod_{\substack{i_Y j_Y \in E_Y \\ u_X i_Y, v_X j_Y \in E_{XY}}} (1 - g(i_Y, j_Y)) \quad (1)$$

- When two nodes n_1 and n_m are not direct neighbors but they are connected via a sequential path p (Figure 2(b)), then we can define

$$h_p = \prod_{i=1}^{m-1} h_1(n_i, n_{i+1}). \quad (2)$$

- At last, when we try to calculate $h(s_X, t_X)$, and if there exists n paths between s_X and s_Y, we also assume that all these paths are independent. Thus,

$$h(s_X, t_X) = 1 - \prod_{i=1}^{n} (1 - h_{p_i}) \quad (3)$$

The major difference between the above formulation and traditional activation probability in IC model is the activation probability between neighbors. We do not use uniform distribution to draw a number and fix it for every edge in the influence network. Other than that, we calibrate the activation probability for every edge based on their nodes' connections to the similarity network and how those connections interact in the similarity network. Most recently, a study from [3] has explored the way of tailoring edge-dependent activation probability for influence maximization. However, it has no consideration of similarity information from another network, and it has no intention to use the reinforcement of influence and similarity information.

Now we will discuss how we define similarity scores in the similarity network based on influence scores in the Influence network.

2.3 Similarity Measure on Similarity Network

We want to first clarify that the similarity in this paper is an **asymmetric** concept. The similarity score of two nodes could be symmetric and it has valid meanings in many applications. **Asymmetric similarity** is a complementary concept to symmetric similarity because there also exists many other practical circumstances where $g(u_Y, v_Y) \neq g(v_Y, u_Y)$. For example, in a paper citation network, paper A having certain similarity to one of its citations B does not necessarily mean B has the same level of similarity with A. In a social network, entities' similarity are also often asymmetric, e.g., the way a song being similar to a video in sharing events is often different from the way a video being similar to a song. Therefore, we model the similarity in an asymmetric way.

Given two nodes $u_Y \in V_Y$ and $v_Y \in V_Y$ in the Similarity network $G_Y(V_Y, E_Y, F_Y)$, let $g(u_Y, v_Y)$ be the similarity of u_Y and v_Y. When defining similarity score of two nodes, we explore a similar manner of the state-of-the-art link based on similarity research [6] by considering the interactions of common nodes connected to them. SimRank essentially considers common neighbors of two nodes as a starting point of their similarity measure. It goes through an iterative process to update such similarities based on the updated similarity values of other nodes in the network. PRank goes one step further by considering in-degree similarity and out-degree similarity in a directed network. PRank also utilizes an iterative process because two nodes' similarity depends on other nodes' similarities as well.

The major differences between our similarity computation and PRank's similarity computation lie in two folds. First, we let influence values from another type of nodes contribute in similarity computation, in addition to link-based analysis of PRank. Second, when considering the similarity of two nodes, for each node in one node's neighbor set, we take the most similar node of it in the other node's neighbor set, rather than computing pair-wise similarities as PRank does. Figure 3 gives an example that illustrates the whole similarity computation. A_1 and A_2 are sets of nodes of in-links of i_Y and j_Y respectively, while B_1 and B_2 are sets of nodes of their out-links. In our model, $g(i_Y, j_Y)$ is related to $g(a_1, a_2), \forall a_1 \in A_1, a_2 \in A_2$ and $g(b_1, b_2), \forall b_1 \in B_1, b_2 \in B_2$.

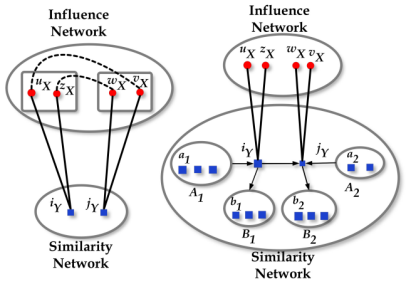

Figure 3: similarity Example

Furthermore, as we see i_Y and j_Y are connected to nodes in Influence network (represented in circle), the influence among u_X, v_X, z_X and w_X also contribute to $g(i_Y, j_Y)$. We use a weighted sum of two different parts to merge information from the similarity network and the Influence network to compute $g(i_Y, j_Y)$. The first part is another weighted sum of the similarity of i_Y, j_Y's in-links set and the similarity of their out-links set in the similarity network. The second part is the influence between the connected nodes of i_Y and connected nodes of j_Y in the Influence network.

Formally, the similarity between i_Y and j_Y is defined as follows.

$$g(i_Y, j_Y) = \sigma((\lambda \frac{1}{|I(i_Y)|} \sum_{\{k|(k,i_Y)\in I(i_Y)\}} max_{\{l|(l,j_Y)\in I(j_Y)\}} g(k,l)$$

$$(1-\lambda)\frac{1}{|O(i_Y)|} \sum_{\{k|(k,i_Y)\in O(i_Y)\}} max_{\{l|(l,j_Y)\in O(j_Y)\}} g(k,l))$$

$$+(1-\sigma) max_{\substack{z_X i_Y \in E_{XY} \\ w_X j_Y \in E_{XY}}} h(z_X, w_X) \quad (4)$$

Here $h(z_X, w_X)$ is the influence of z_X on w_X, and $I(i_Y)$ and $O(i_Y)$ are i_Y's in-degree and out-degree neighbors.

2.4 Iterative Algorithm for IS Computation

Before diving into detail computations, we first briefly review our ultimate goal. Out of many practical demands we reduce a heterogeneous network to an IS network where we want to compute influence maximization on Influence network and similarity analysis on the similarity network. We have observed the benefit of passing information between these two tasks so that the results' qualities of both of them can be improved. We have also formulated new activation probabilities for influence maximization purpose and new similarity measures for similarity analysis. Both formulations originate from state-of-the-art works in these two fields.

From Eq. $1 \sim 3$, we know that there is no close form solutions for each individual activation probability h and similarity score g, because they are nonlinear, non convex, and mutually dependent. Due to the fact that h and g are intractable, we design an iterative procedure to approximate their values. We also design a pruning and dampening michanism to accelerate the computation. The whole algorithm is as the following.

Algorithm 1 IS algorithm

Input: Influence Network $G_X(V_X, E_X, F_X)$; Similarity Network $G_Y(V_Y, E_Y, F_Y)$; Information Tunnel $G_{IT}(V_X \cup V_Y, E_{XY} \cup E_{YX})$; Simulation times R
Output: $g(u,v)\ if\ uv \in E_Y, h(x,y)\ if\ x \in V_X, y \in V_X$

// Initialization step
Compute the first round's result of PRank of u, v as $g_0(u,v)$
// Start to compute g and h
Set $i = 1$
while g or h is not converged **do**
  // Compute h_{i-1}
  Enumerate each $g(u,v)$ to get $p(x \to y)$ for $xu \in E_{XY}, yv \in E_{XY}$
  First prune: remove all $p(x \to y)$ which is below average of all non-zero $p(x \to y)$
  Set all $t(x,y) = 0$
  for j=1 to R **do**
    Simulate each edge's activation
    if x can reach y through a path of activated edges **then**
      $t(x,y)$++
    end if
  end for$h(x,y) = t(x,y)/R$
  Second prune: remove all $h(x,y)$ which is below $\frac{\sum_j h(x,j)}{\#non-zero\ h(x,j)}$
  // Compute g_i
  Compute each $g_i(u,v)$ using Eq. 4
  i++
end while

3. EVALUATION

3.1 Dataset Description

We use paper collection of ACM digital library [5] as an instance of IR network. We treat the paper citation network as similarity network. If paper A cites paper B, we know that the authors of A have influence on authors of B. Therefore, we also construct an Influence network of author relations. In total, the similarity network has $217,335$ nodes and $632,751$ edges, while the Influence network has $250,566$ nodes and $1,486,909$ edges. The number of edges between similarity network and Influence network is $518,358$. In the experiments, we compare our model with the state of the art social influence maximization algorithm [1, 4] on influence part and similarity computation algorithm [6] on the similarity part, respectively. On the activation simulation we set the simulation times $R = 1000$ for our method and both baselines (described below). We choose λ to be 0.5 and σ to be 0.8 for the IR model. In practice, our method converges on g and h values after 10 iterations. The experiment system is implemented in Java with JDK1.6, Eclipse and conducted on machines with Quad Core CPU with 2.2 GHz and 4GB RAM.

3.2 Comparing Seed Quality in Social Influence Maximization with State-Of-The-Art IC Model

3.2.1 Baseline Description

We compared with two baseline methods. The first one is the classic way to assign activation probability in original IC model, which is uniformly drawing a probability depending on the number of edges between two nodes. Despite the simplicity, it does not differentiate any edge. We will show that, with the help of the similarity network information, we may assign more reasonable probability to each of these edges in order to pick up more reasonable "influential seeds". Therefore, in our method, activation probabilities are not uniformly distributed. To make a fair comparison, we control the median of the uniform distribution to be the same as the median of our distribution.

We also see that the activation routes change a lot from the first baseline to our method due to our non-uniformed activation probability distribution. By comparing with the first baseline, we will show that this new activation path structure is more reasonable. Furthermore, we want to show that each assignment of the activation probability on each path is also reasonable. Therefore, we design a second baseline as follows. We first obtain a distribution of our activation probabilities. Secondly, by following that distri-

bution, we generate a random number as the activation probability for every edge. Therefore, this generated network has the same path structure and activation probability distribution, but different assignment of each edge. We use it as our second baseline. Since this baseline has the same activation path structures as IS's, but randomness for each edge, its performance should be in between of IS and original IC and closer to IS because it is IS's variation.

3.2.2 *Seeds Quality Comparison*

In majority of social influence maximization research, seeds selection is the origins of their motivation. However, few works actually discussed what kinds of people those seeds are. We can demonstrate the IS computation generates more reasonable seeds.

Figure 3.2.2 shows the average G-index of our seed list and the two baselines. In addition to the whole area Influence Network, we extracted author relations within three sub-areas separately, including Database and Data Mining (DBDM) area , Information Retrieval, AI, and Machine Learning (IRAIML) area, and Computer Architecture and Hardware (CAHW) area. Our approach achieves better performance in overall network and sub-area networks. We also have similar result for H-index, which we choose to not show here to save some space.

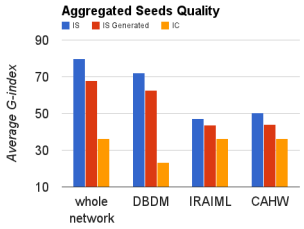

Figure 4: Seeds Quality Comparison

3.3 Comparing Similarity Computation with PRank model

We use k-$medoids$ to cluster nodes in similarity network. Since we have similarity scores of nodes, we plug those scores as similarity measure into the clustering method. The goal is to see which set of scores produce higher quality clusters. We use the compactness of the result clusters as such quality measures. Here we present the comparison result obtained from three sub-areas similarity network in Figure 5. We only show one subnetwork as others have similar performances. As one can see, the similarity scores produced by IS scheme performs consistently better than PRank [6]. The compactness is defined by Davies-Bouldin index.

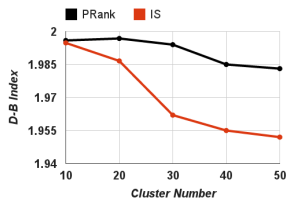

Figure 5: Compactness Comparison in similarity Networks

4. RELATED WORK

What we present in this paper starts from an unique observation that combining social influence and similarity analysis could benefit each other through information exchange. Influence maximization problem is an important research branch on social networks. The task of influence maximization is to choose some seed users to spread certain information as wide as possible, given a certain social network [1,4]. However, majority of them do not pay attention to how to get activation probabilities, which is one part of the problem we are solving in this paper. Amit et.al [3] did the first work to attack this open problem. Our work is different from theirs in two ways. First, we do not use any guidance as action logs. Second and more importantly, their work does not consider heterogeneous network and does not use similarity measure to calibrate activation probability results. Similarity analysis on social network is usually based on nodes' common neighbors or link properties, e.g., [6] Different than most of them, we apply more information from social influence side to improve the similarity measure, which is not considered in previous similarity research.

We have also noticed that there is another work studied a total different relation of social influence and similarity together [2]. They studied how people's influence and their similarities affect each other. In another word, they consider the same type of nodes' similarity and influence in a homogeneous network. We consider influence of one type of nodes, with the information of similarity of another type of nodes, and vise versa, in a heterogeneous network.

5. CONCLUSION

We observe the benefits of modeling influence and similarity together for ubiquitous heterogeneous IS network. We design a method for such modeling and demonstrate the lifts for both sides using a large scale real world data. We believe that analysis on IS network has a promising future because social influence and similarity studies are two building blocks for many research interests, such as clustering, classification and recommendation.

6. ACKNOWLEDGEMENT

This work is supported in part by NSF through grants IIS-0905215, CNS-1115234, IIS-0914934, DBI-0960443, and OISE-1129076, US Department of Army through grant W911NF-12-1-0066, and Google Mobile 2014 Program.

7. REFERENCES

[1] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *KDD*, 2009.

[2] David Crandall, Jon Kleinberg, Dan Cosley, Siddharth Suri, and Daniel Huttenlocher. Feedback effects between similarity and social influence in online communities, 2008.

[3] A. Goyal, F. Bonchi, and L.V.S. Lakshmanan. Learning influence probabilities in social networks. In *WSDM'10*.

[4] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[5] J. Tang and J. Zhang et. al. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, 2008.

[6] P. Zhao, J. Han, and Y. Sun. P-rank: a comprehensive structural similarity measure over information networks. In *CIKM*, 2009.