2013 IEEE 13th International Conference on Data Mining

Least Cost Influence in Multiplex Social Networks: Model Representation and Analysis

Dung T. Nguyen, Huiyuan Zhang, Soham Das, My T. Thai
Department of Computer & Information Science & Engineering
University of Florida
Gainesville, Florida, USA
Email: {dtnguyen, huiyuan, sdas, mythai}@cise.ufl.edu

Thang N. Dinh
Department of Computer Science
Virginia Commonwealth University
Richmond, Virginia, USA
Email: {tndinh}@vcu.edu

***Abstract*—The least cost influence (LCI) problem, which asks to identify a minimum number of seed users who can eventually influence a large number of users, has become one of the central research topics recently in online social networks (OSNs). However, existing works mostly focused on a single network while users nowadays often join several OSNs. Thus, it is crucial to investigate the influence in multiplex networks, i.e. the influence is diffused across a set of networks via shared users, in order to obtain the best set of seed users.**

In this paper, we propose a unified framework to represent and analyze the influence diffusion in multiplex networks. More specifically, we tackle the LCI problem in multiplex OSNs by reducing multiplex networks to a single network via various coupling schemes while preserving the most influence propagation properties. Besides the coupling schemes to represent the diffusion process, the framework also includes the influence relay, a new metric to measure the flow of influence inside and between networks. The experiments on both real and synthesized datasets validate the effectiveness of the coupling schemes as well as provide some interesting insights into the process of influence propagation in multiplex networks.

I. Introduction

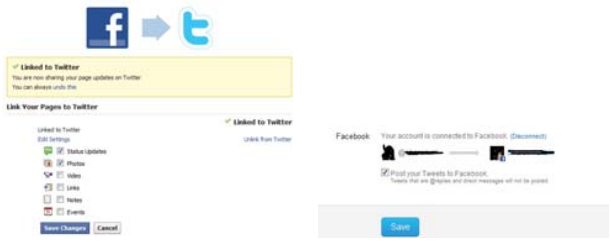
In the recent decade the popularity of online social networks, such as Facebook, Google+, Myspace and Twitter etc., has created a new major communication medium and formed a promising landscape for information sharing and discovery. On average, Facebook users spend 7 hours and 45 minutes per person per month [2]; 3.2 billion likes and comments are posted every day on Facebook [1]; 340 million tweets are sent out everyday on Twitter [2]. Such engagement of online users fertilizes the land for information propagation to a degree never achieved before in mass media. More importantly, OSNs also inherit one of the major properties of real social networks – the word-of-mouth or peer-pressure effect in which an individual's opinion or decision is influenced by his friends and colleagues. Due to the considerable impact of this effect on the popularity of new products [4], [10], OSNs have rapidly become one of the most attractive choices for raising the awareness of new products or brands as well as to reinforce the connection between customers and companies. The crucial problem is how to find the least advertising cost set of influencers who can influence a massive number of users.

There is a considerable number of overlapping users among major OSNs which creates a huge effect on the diffusion of information in these networks. When a user joins different networks, s/he can relay the information from one network to another. Let us consider the following typical scenario to illustrate this phenomenon. Jack, a user of both Twitter and Facebook, logs in Twitter and knows about an excellent product from his friend. He right away falls in love with the new product and eagerly shares the information by tweeting it. Moreover, he configured his Twitter and Facebook accounts as illustrated in Fig. 1 that allows him to automatically post on his Facebook's wall whenever he has a new tweet and vice versa. As the consequence, the product information is exposed to his friends in both networks and the information further spreads out on both networks. If we only consider the information propagation in one network, the reach of the information is estimated incorrectly and thus the influence of users in these networks. In this case, the influence of Jack should be the combination of his influence in both networks. As shown in Fig. 2, the fraction of overlapping users is considerable, therefore studying the influence only in one network fails to identify most influent users. This motivates us to study the problem in multiplex networks where the influence of users is evaluated based on OSNs in which they participate.

Related works. Nearly all the existing works studied different variants of the least cost influence problem on a single network. Kempe et al. [12] first formulated the influence maximization problem which asks to find a set of k users who can maximize the influence. The influence is propagated based on a stochastic process called Independent Cascade Model (IC) in which a user will influence his friends with probability proportional to the strength of their friendship. The author proved that the problem is NP-hard and proposed a greedy algorithm with approximation ratio of $(1-1/e)$. After that, a considerable number of works studied and design new algorithms for the problem variants on the same or extended models such as [7], [13]. There are also works on the linear threshold (LT) model for influence propagation in which a user will adopt the new product when the total influence of his friends surpass some threshold. Feng et al. [20] showed NP-completeness for the problem and Dinh et al [8] proved the inapproximability as well as proposed efficient algorithms for this problem on a special case of LT model. In their model, the influence between users is uniform and a user is influenced if a certain fraction ρ of his friends are active.

Recently, researchers have started to explore multiplex net-

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.24



(a) Auto post from Facebook to Twitter (b) Auto post from Twitter to Facebook

Fig. 1. Auto update across social networks

users of these sites...

... are also users of these sites	LinkedIn	Facebook	Twitter	MySpace
LinkedIn	-	12%	21%	6%
Facebook	82%	-	91%	57%
Twitter	31%	20%	-	17%
MySpace	36%	49%	70%	-

Fig. 2. The number of shared users between major OSNs in 2009 [3]

works with works of Yagan et al. [19] and Liu et al. [14] which studied the connection between offline and online networks. The first work investigated the outbreak of information using the SIR model on random networks. The second one analyzed networks formed by online interactions and offline events. The authors focused on understanding the flow of information and network clustering but not solving the least cost influence problem. Additionally, these works did not study any specific optimization problem of viral marketing. Shen et al. [18] explored the information propagation in multiplex online social networks taking into account the interest and engagement of users. The authors combined all networks into one network by representing an overlapping user as a super node. This method cannot preserve the individual networks' properties. Nguyen et al. [17] studied the influence maximization problem which has different constraints with ours.

In this paper, we study the Least Cost Influence problem which asks for a set of users with minimum cardinality to influence a certain fraction of users in multiplex networks. Due to the complex diffusion process in multiplex networks, it is difficult develop the solution by extending previous solutions in a single network. Additionally, studying the problem in multiplex networks introduces several new interesting concerns: how to evaluate the influence of overlapping users in multiplex networks? in which network, a user is easier to be influenced? which network propagates the influence better? To answer these intricate questions, we first introduce a model representation to illustrate how information propagate in multiplex networks via coupling schemes. Coupling schemes combine multiple networks into one network while retaining the influential properties of the original networks partially or fully. After coupling the networks, we can exploit existing solutions on the single network to solve the problem. This is a powerful and comprehensive procedure to study LCI. Moreover, we propose a new metric called influence relay to analyze the flow of influence between networks. Through comprehensive experiments, we discover crucial properties of the multiplex networks in diffusing the information. Our main contributions are summarized as follows:

- Proposing a model representation via various coupling schemes to reduce the problem in multiplex networks to an equivalent problem on a single network. Coupling schemes can be applied for most popular diffusion models including: Linear Threshold model, Stochastic Threshold model, and Independent Cascading model.
- Propose a new metric called *influence relay* to analyze the influence diffusion process both a single network and multiplex networks.
- Explore and verify through extensive experiments that multiplex networks significantly support each other to propagate the influence.

The rest of the paper is organized as follows. In Section II, we present the influence propagation model in multiplex networks and define the problem. Our coupling schemes for Linear Threshold model are proposed in Section III and Section IV. We then define the *influence relay* metric in Section V. Section VI shows the experimental results on the performance of different algorithms and coupling schemes. Finally, Section VII concludes the paper.

II. Model and Problem Definition

A. Graph notations

We consider k networks $G^1, G^2, \ldots, G^k$, each of which is modeled as a weighted directed graph $G^i = (V^i, E^i, \theta^i, W^i)$. The vertex set $V^i = \{u\text{'s}\}$ represents the participation of $n^i = |V^i|$ users in the network G^i, and the edge set $E^i = \{(u,v)\text{'s}\}$ contains $m^i = |E^i|$ oriented connections (e.g., friendships or relationships) among network users. $W^i = \{w^i(u,v)\text{'s}\}$ is the (normalized) weight function associated to all edges in the i^{th} network. In our model, weight $w^i(u,v)$ can also interpreted as the strength of influence (or the strength of the relationship) a user u has on another user v in the i^{th} network. The sets of incoming and outgoing neighbors of vertex u in network G^i are denoted by N_u^{i-} and N_u^{i+}, respectively. In addition, each user u is associated with a threshold $\theta^i(u)$ indicating the persistence of his opinions. The higher $\theta^i(u)$ is, the more unlikely that u will be influenced by the opinions of his friends. Furthermore, the users that actively participate in multiple networks are referred to as *overlapping users* and can be identified using methods in [11], [5] (Note that identifying overlapping users is not the focus of this paper). Those users are considered as bridge users for information propagation across networks. Finally, we denote by $G^{1\ldots k}$ the system consisting of k networks, and by U the exhaustive set of all users $U = \cup_{i=1}^{k} V^i$.

B. Influence Propagation Model

We first describe the *linear threshold model* (LT-model) [20], [8], a popular model for information and influence diffusion in a single network, and then discuss how this LT model can be extended to cope with multiplex networks. In the original LT model, each network user u is either in an *active* or *inactive* state: u is in an *active* state if he originally adopts the information, or the total influence from his direct neighbors

exceeds his threshold $\theta(u)$, i.e, $\sum_{v \in N(u)} w(v,u) \geq \theta(u)$. Otherwise, u is in an *inactive* state.

In a big picture, given a system of k networks, the information is propagated separately in each network and can only be transferred from one network to another via the overlapping users of these networks. The information starts to spread out from a set of seed users S i.e. all users in S have active state and the remaining users are inactive. At time t, a user u becomes active if the total influence from its active neighbors surpasses its threshold in some network i.e. there exists i such that:

$$\sum_{v \in N_u^{i-}, v \in A} w^i(v,u) \geq \theta^i(u)$$

where A is the set of active users after time $(t-1)$.

After each time step, new inactive users are activated and they continue to activate other users. The process will continue until no more inactive users can be activated. If we limit the propagation time to d, then the process will stop after $t = d$ time steps. The set of active users caused by the seed set S after time d is denoted as $A^d(G^{1\cdots k}, S)$. Note that d is also the number of hops in networks up to which the influence can be propagated from the seed set, so d is called the number of propagation hops.

C. Problem definition

In this paper, we address the fundamental problem of viral marketing in multiplex networks: the **Least Cost Influence** problem. The problem asks to find a seed set of minimum cardinality which influences a large fraction of users.

Definition 1. *(Least Cost Influence (LCI) Problem) Given a system of k networks $G^{1\cdots k}$ with the set of users U, a positive integer d, and $0 < \beta \leq 1$, the LCI problem asks to find a seed set $S \subset U$ of minimum cardinality such that the number of active users after d hops according to LT model is at least β fraction of users i.e. $|A^d(G^{1\cdots k}, S)| \geq \beta |U|$.*

When $k = 1$, we have the variant of the problem on a single network which NP-hard to solve [6] but it is easier to design heuristic algorithm in the single network. In next sections, we present different coupling strategies to reduce the problem in multiplex networks to one in a single network in order to utilize the algorithm design.

III. Lossless coupling schemes

In this section, we present the lossless scheme to couple multiple networks into a new single network with respect to the influence diffusion process on each participant network. A notable advantage of this newly coupled graph is that we can use any existing algorithm on a single network to produce the solution in multiplex networks with the same quality. Unfortunately, we encounter a series of the challenging issues in designing such coupling schemes:

(1) The heterogeneity of user participation. A user may join in one, two, or more networks. How can we recognize and differentiate these users in the coupled network? How to capture the roles of each user in the multiplex networks?
(2) The process of information and influence propagation among networks. In multiplex networks, when a user is influenced he tends to immediately propagate the information on all networks that he is a part of. How can we describe this immediate transmission of the information between networks in just a single network?
(3) Preserving properties of individual networks. The coupled network should be a good representative of all the individual networks. It should preserve the diffusion properties of all the networks. That enables us to establish the relationship between solution of the problem on the coupled network and the original networks. How can we design a coupling scheme that addresses this issue?

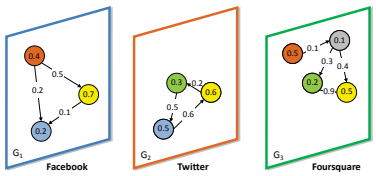
A. A coupling scheme for LT-model

In LT-model, the first issue is solved by introducing dummy nodes for each user u in networks that it does not belong to. These dummy nodes are isolated. Now the vertex set V^i of i^{th} network can be represented by $V^i = \{u_1^i, u_2^i, \ldots, u_n^i\}$ where $U = \{u_1, u_2, \ldots, u_n\}$ is the set of all users. u_p^i is called the *representative vertex* of u_p in network G^i. In the new representation, there is an edge from u_p^i to u_q^i if u_p and u_q are connected in G^i. Now we can union all k networks to form a new network G. The approach to overcome the second challenge is to allow nodes $u^1, u^2, \ldots, u^k$ of a user u to influence each other e.g. adding edge (u^i, u^j) with weight $\theta(u^j)$. When u^i is influenced, u^j is also influenced in the next time step as they are actually a single overlapping user u, thus the information is transferred from network G^i to G^j. But an emerged problem is that the information is delayed when it is transferred between two networks. Right after being activated, u^i will influence its neighbors while u^j needs one more time step before it starts to influence its neighbors. It would be better if both u^i and u^j start to influence their neighbors in the same time. For this reason, new *gateway vertex* u^0 is added to G such that both u^i and u^j can only influence other vertices through u^0. In particular, all edges (u^i, v^i) $((u^j, z^j))$ will be replaced by edges (u^0, v^i) $((u^0, z^j))$. In addition, more edges are added between u^0, u^i, and u^j to let them influence each other. After forming the topology of the coupled network, we assign edge weights and vertex thresholds as following:

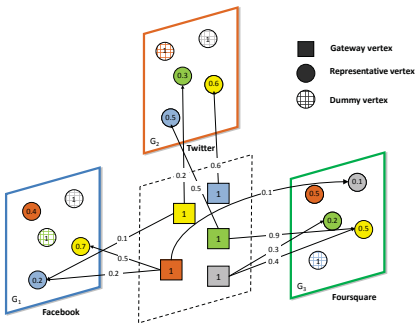
Vertex thresholds. All dummy vertices and gateway vertices have the threshold of 1. Any remaining representative vertex u_p^i has the same threshold as u_p in G^i, i.e., $\theta(u_p^i) = \theta^i(u_p)$.

Edge weights. If there is an edge between user u and v in G^i, then the edge (u^0, v^i) has weight $w(u^0, v^i) = w^i(u, v)$. The edges between gateway and representative vertices of the same user u are assigned as $w(u^i, u^j) = \theta(u^j), \forall 0 \leq i, j \leq k, i \neq j$ to synchronize their state together. A simple example of the scheme is illustrated in Fig. 3.

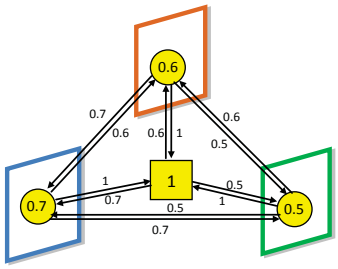
Next we will show that the propagation process in the original multiplex networks and the coupled network is actually the same. Influence is alternatively propagated between gateway and representative vertices, so the problem with d hops in the multiplex networks is equivalent to the problem with $2d$ hops in the coupled network.



(a) An instance of multiplex networks with 4 users. Each user is represented by vertices of the same color with different thresholds in different networks e.g. green user has thresholds of 0.3 and 0.2 in G^2 and G^3.



(b) The influence of gateway vertices on representative vertices represent the influence between users in multiplex networks.



(c) The connection between gateway and representative vertices of the same user.

Fig. 3. An example of the *lossless coupling scheme*

Lemma 1. *Suppose that the propagation process in the coupled network G starts from the seed set which contains only gateway vertices $S = \{s_1^0, \ldots, s_p^0\}$, then representativeices are activated only at even propagation hops.*

Proof: Suppose that a gateway vertex u^0 is the first gateway vertex that is activated at the odd hops $2d+1$. u^0 must be activated by some vertex u^i and u^i is the first activated vertex among vertices $u^1, u^2, \ldots, u^k$. It means that u^i is activated in hop $2d$. Since all incoming neighbors of u^i are gateway vertices, some gateway vertex becomes active in hop $2d-1$ (contradiction). ■

Lemma 2. *Suppose that the propagation process on $G^{1\ldots k}$ and G starts from the same seed set S, then following conditions are equivalent:*

(1) *User u is active after d propagation hops in $G^{1\ldots k}$.*
(2) *There exists i such that u^i is active after $2d-1$ propagation hops in G.*
(3) *Vertex u^0 is active after $2d$ propagation hops in G.*

Proof: We will prove this lemma by induction. Suppose it is correct for any $1 \le d \le t$, we need to prove it is correct for $d = t+1$. Denote $A^{1\ldots k}(t)$ and $A(t)$ as the set of active users and active vertices after t propagation hops in $G^{1\ldots k}$ and G, respectively.

$(1) \Rightarrow (2)$: If user u is active at time $t+1$ in $G^{1\ldots k}$, it must be activated in some network G^j. We have:

$$\sum_{v \in N_u^{j-} \cap A^{1\ldots k}(t)} w^j(v,u) \ge \theta^j(u)$$

Due to the induction assumption, for each $v \in A^{1\ldots k}(t)$, we also have $v^0 \in A(2t)$ in G. Thus:

$$\sum_{v^0 \in N_{u^j}^- \cap A(2t)} w(v^0, u^j) = \sum_{v \in N_u^{j-} \cap A^{1\ldots k}(t)} w^j(v,u) \ge \theta^j(u) = \theta(u^j)$$

It means that u^j is active after $(2(t+1)-1)$ propagation hops.

$(2) \Rightarrow (3)$: If there exists i such that u^i is active after $2(t+1)-1$ propagation hops on G, then u^i will activate u^0 in hop $2(t+1)$

$(3) \Rightarrow (1)$: Suppose that $u^0 \notin S$ is active after $2(t+1)$ propagation hops in G, then there exists u^j which activates u^0 before. This is equivalent to:

$$\sum_{v \in N_{u^j}^-, v \in A(2t)} w(v, u^j) \ge \theta(u^j)$$

For each $v \in A(2t)$, we also have $v \in A^{1\ldots k}(t)$. Replace this into the above inequality we have:

$$\sum_{v \in N_u^{j-} \cap A^{1\ldots k}(t)} w^j(v,u) = \sum_{v^0 \in N_{u^j}^- \cap A(2t)} w(v^0, u^j) \ge \theta(u^j) = \theta^j(u)$$

Thus, u is active in network G^j after $t+1$ hops. ■

Next, we will show that the number of influenced vertices in the coupled network is $(k+1)$ times the number of influenced users in multiplex networks as stated in Theorem 1.

Theorem 1. *Given a system of k networks $G^{1\ldots k}$ with the user set U, the coupled network G produced by the lossless coupling scheme, and a seed set $S = \{s_1, s_2, \ldots, s_p\}$, if $A^d(G^{1\ldots k}, S) = \{a_1, a_2, \ldots, a_q\}$ is the set of active users caused by S after d propagation hops in multiplex networks, then $A^{2d}(G,S) = \{a_1^0, a_1^1, \ldots, a_1^k, \ldots, a_q^0, a_q^1, \ldots, a_q^k\}$ is the set of active vertices caused by S after $2d$ propagation hops in the coupled network.*

Proof: For each user $a_i \in A^d(G^{1\ldots k}, S)$ i.e. a_i is active after d hops in $G^{1\ldots k}$, then there exists a_i^j which is active after $2d-1$ hops in G according to the Lemma 2. As a result, all $a_i^0, a_i^1, \ldots, a_i^k$ are active after $2d$ hops. So $B = \{a_1^0, a_1^1, \ldots, a_1^k, \ldots, a_q^0, a_q^1, \ldots, a_q^k\} \subseteq A^{2d}(G,S)$.

Let consider a vertex of $A^{2d}(G,S)$ which is:

Case 1. A gateway vertex u^0 which is active after $2d$ hops in G, so vertex u must be active after d hops in $G^{1\ldots k}$. This implies $u \in A^d(G^{1\ldots k}, S)$, thus $u^0 \in B$.

Case 2. An representative vertex u^i. If u^i is active after $2d-1$ hops, then u must be active after d hops due to Lemma

2, thus $u \in A^d(G^{1\cdots k}, S)$. Otherwise, u^i is activated at hop $2d$, it must be activated by some vertex u^j, $j > 0$ since all gateway vertices only change their state at even hops. Again, $u \in A^d(G^{1\cdots k}, S)$. This results in $u^i \in B$.

From two above cases, we also have $A^{2d}(G, S) \subseteq B$. So that $A^{2d}(G, S) = B$, the proof is completed. ■

Theorem 1 provides the basis to derive the solution for LCI in multiplex networks from the solution on a single network. It implies an important algorithmic property of the *lossless coupling scheme* regarding the relationship between the solutions of LCI in $G^{1\cdots k}$ and G. The equivalence of two solutions is stated below:

Theorem 2. *When the lossless scheme is used, the set $S = \{s_1, s_2, \ldots, s_p\}$ influences β fraction of users in $G^{1\cdots k}$ after d propagation hops if and only if $S' = \{s_1^0, s_2^0, \ldots, s_p^0\}$ influences β fraction of vertices in coupled network G after $2d$ propagation hops.*

Size of the coupled network. Each user u has $k + 1$ corresponding vertices $u^0, u^1, \ldots, u^k$ in the coupled network, thus the number of vertices is $|V| = (k+1)|U| = (k+1)n$. The number of edges equals the total number of edges from all input networks plus the number of new edges for synchronizing. Thus the total number of edges is $|E| = \sum_{i=1}^{k} |E^i| + nk(k+1)$.

B. Extensions to other diffusion models

In this section, we show that we can design lossless coupling schemes for some other well-known diffusion models in each component network. As a result, top influential users can be identified under these diffusion models. In particular, we investigate two most popular stochastic diffusion models which are Stochastic Threshold and Independent Cascading models [12].

- *Stochastic Threshold model.* This model is similar to the Linear Threshold model but the threshold $\theta^i(u^i)$ of each node u^i of G^i is a random value in the range $[0, \Theta^i(u^i)]$. Node u^i will be influenced when $\sum_{v^i \in N^-_{u^i}, v \in A} w^i(v^i, u^i) \geq \theta^i(u^i)$
- *Independent Cascading model.* In this model, there are only edge weights representing the influence between users. Once node u^i of G^i is influenced, it has a single chance to influence its neighbor $v^i \in N^+(u^i)$ with probability $w^i(u^i, v^i)$.

For both models, we use the same approach of using gateway vertices, representative vertices and the synchronization edges between gateway vertices and their representative vertices. The weight of edge (u^i, u^j), $0 \leq i \neq j \leq k$ will be $\Theta(u^j)$ for Stochastic Threshold model and 1 for Independent Cascading model. Once u^i is influenced, u^j will be influenced with probability 1 in the next time step. The proof for the equivalence of the coupling scheme is similar to ones for LT-model.

IV. Lossy coupling schemes

In the preceding coupling scheme for LT-model, a complicated coupled network is produced with large numbers of auxiliary vertices and edges. It is ideal to have a coupled network which only contain users as vertices. This network provides a compact view of the relationship between users crossing the whole system of networks. The loss of the information is unavoidable when we try to represent the information of multiplex networks in a compact single network. The goal is to design a scheme that minimizes the loss as much as possible i.e. the solution for the problem in the coupled network is very close to one in the original system. Next, we present these schemes based on the following key observations.

Observation 1. User u will be activated if there exists i such that: $\sum_{v \in N_u^{i-} \cap A} w^i(v, u) \geq \theta^i(u)$ where A is the set of active users. We can relax the condition to activate u with positive parameters $\alpha^1(u)$, $\alpha^2(u)$, …, $\alpha^k(u)$ as follows:

$$\sum_{i=1}^{k} \left(\alpha^i(u) \sum_{v \in N_u^{i-} \cap A} w^i(v, u) \right) \geq \sum_{i=1}^{k} \alpha^i(u) \theta^i(u) \quad (1)$$

Proposition 1. *Given a system of networks $G^{1\cdots k}$, if the condition (1) is satisfied, then user u is activated.*

Proof: When the condition is satisfied, there must exist i such that $\alpha^i(u) \sum_{v \in N_u^{i-} \cap A} w^i(v, u) \geq \alpha^i(u)\theta^i(u)$. As a result, the condition to activate u is satisfied since $\alpha^i(u) > 0$ ■

Note that sometimes the condition to activate u is met, but the condition (1) still needs more influence from u's friends to satisfy. The more this need for extra influence is, the looser condition (1) is. We can reduce this redundancy by increasing the value of $\alpha^i(u)$ proportional to the value of $\sum_{v \in N_u^{i-} \cap A} w^i(v, u) - \theta^i(u)$. In the special case, if $\sum_{v \in N_u^{i-} \cap A} w^i(v, u) > \theta^i(u)$ and we choose $\alpha^i(u) \gg \alpha^j(u)$, $\forall j \neq i$, then there is no redundancy. Unfortunately, we do not know before hand in which network user u will be activated, so we can only choose parameters heuristically.

Observation 2. When user u participates in multiple networks, it is easier to influence u in some network than the others. The following simple case illustrate such situation. Suppose that we have two networks. In network 1, $\theta^1(u) = 0.1$ and u has 8 in-neighbors, each neighbor v influences u with $w^1(v, u) = 0.1$. In network 2, $\theta^2(u) = 0.7$ and u has 8 in-neighbors, each neighbor v influences u with $w^2(v, u) = 0.1$. The number of active neighbors to activate u is 1 and 7 in network 1 and 2, respectively.

Easiness. Intuitively, we can say that u is easier to be influenced in the first network. We quantify the *easiness* $\epsilon^i(u)$ that u is influenced in network i as the ratio between the total influence from friends and the threshold to be influenced: $\epsilon^i(u) = \frac{\sum_{v \in N_u^{i-}} w^i(v,u)}{\theta^i(u)}$. We can use the easiness of a user in networks as the parameters of the condition 1.

Based on above observations, we couple multiplex networks into one using parameters $\{\alpha^i(u)\}$. The vertex set is the set of users $V = \{u_1, u_2, \ldots, u_n\}$. The threshold of vertex u is set to $\theta(u) = \sum_{i=1}^{k} \alpha^i(u)\theta^i(u)$

The weight of the edge (v, u) is: $w(v, u) = \sum_{i=1}^{k} \alpha^i(u) w^i(v, u)$ where $w^i(v, u) = 0$ if there is no edge from v to u in i^{th} network.

Then the set of edges is $E = \{(v, u) | w(v, u) > 0\}$. Fig. 4 illustrates the loosy coupled network of networks in Fig. 3.

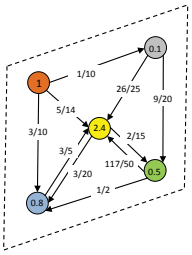


Fig. 4. Lossy coupled network using easiness parameters.

Besides easiness, other metrics can be used for the same purpose. We enumerate here some other metrics.

Involvement. If a user is surrounded by a group of friends who have high influence on each other, he tends to be influenced. When a few of his friends are influenced, the whole group involving him is likely to be influenced. We estimate *involvement* of a node v in a network G^i by measuring how strongly the 1-hop neighborhood v is connected and to what extent influence can propagate from one node to another in the 1-hop neighborhood. Formally we can define *involvement* of a node v in network G^i as: $\sigma_v^i = \sum_{x,y \in N_v^i \cup \{v\}} \frac{w^i(x,y)}{\theta_y^i}$ where $N_v^i = N_v^{i+} \cup N_v^{i-}$ is the set of all neighbors of v (both in-coming and out-going).

Average. All parameters have the same value $\alpha^i(u) = 1$

Next we show the relationship between the solution for LCI in the lossy coupled network and the original system of networks. As discussed in the above observations, if the propagation process starts from the same set of users in $G^{1\dots k}$ and the coupled network G, then the active state of a user in G implies its active state in $G^{1\dots k}$. It means that if the set of users S activates β fraction of users in G, it also activates at least β fraction of users in $G^{1\dots k}$. It implies that if a seed set is a feasible solution in G, it is also a feasible solution in $G^{1\dots k}$. Thus we have the following result.

Theorem 3. *When the lossy coupling scheme is used, if the set of users S activates β fraction of users in G, then it activates at least β fraction of users in $G^{1\dots k}$.*

V. Influence Relay

We propose the *influence relay* metric to quantify the role of users in propagating information. When the information is diffused in multiplex networks, it may flow within a single network or travel through multiplex networks. This raises several questions including: What is the contribution of each component network in the influence process? How much information flows inside a network, or between the networks? Quantifying these values is the key to understand the insights of the diffusion process in multiplex networks.

Influence relay. The influence relay of vertices is recursively defined based on the order of activation. Given a seed set S in the coupled network G, assume that the activation process stops after d hops of propagation. All inactive vertices in $V \setminus A^d(G,S)$ have influence relay 0. For each activated vertex $u \in A^d(G,S)$, the influence relay of u, denoted by $IR(u)$, is a linear combination of the influence relay of its outgoing neighbors that are activated after u. Formally, the influence relay of $u \in A^d(G,S)$ is defined as

$$IR(u) = 1 + \sum_{v \in N_u^+ \wedge h(v) > h(u)} \frac{w(u,v) IR(v)}{\sum_{z \in N_v^- \wedge h(z) < h(v)} w(z,v)} \quad (2)$$

where $h(u)$ is the hop at which u is activated.

The influence relay captures the amount of influence a vertex "'relays" to other vertices after adopting the information. The influence relay of a vertex u depends largely on the influence relay of vertices that u helps to activate and the weights of edges between u and them. Note that each outgoing neighbor v of u might also be under the influence of other vertices. Among them, u contributes the impact of $w(u,v)$ to influence v, hence u is responsible for only $\frac{w(u,v)}{\sum_{z \in N_v^- \wedge h(z) < h(v)} w(z,v)}$ of v's influence relay. Moreover, we add 1 to the influence relay of u since u also contributes itself to the set of activated vertices.

Computing influence relay. We compute the influence relay of nodes in the reverse order of the diffusion process. First we construct the influence graph $IG_S = (V_S, E_S)$ from the seed set S to represent the diffusion process and compute the influence relay of all nodes in V_S. The vertex set V_S of n_S nodes is $A^d(G,S)$. There is an edge from u to v in E_S with the same weight $w(u,v)$ in G if u has passed the information to v, i.e., $u, v \in A^d(G,S)$ and $h(v) > h(u)$. The graph IG_S is a directed acyclic graph, thus we can compute in linear time the influence relay of all vertices in the reverse topological ordering of IG_S as described in *CIR* algorithm (Algorithm 1 in the appendix).

One of the important properties of the influence relay is that it preserves the number of activated vertices as stated in the following theorem.

Theorem 4. *The total influence relay of seeding vertices is equal to the total number of activated vertices.*

$$\sum_{u \in S} IR(u) = |A^d(G,S)|$$

Proof: See appendix for the proof. ■

Influence contribution. To obtain the contribution of a network to the diffusion process, we sum up the influence relay of all seed vertices of that network.

Internal and external influence. We propose the concepts of internal and external influence to quantify the amount of information flowing within and between networks. When the information is propagated from a seed set in a component network called the target network to its vertices, there are two kinds of influence paths: *internal paths* include only edges in the target network and *external paths* include some edges of other networks. An external path is formed when some of its vertices are activated outside the target network. The total influence which passes through such paths can be considered as the support of other networks to propagate information in the target network. We adapt the relay influence to measure the internal influence (passes through internal paths) and external influence (passes through external paths) of the seed set in the target network as follows. Each vertex u has internal influence

Networks	#Nodes	#Edges	Avg. Degree
Twitter	48277	16304712	289.7
FSQ	44992	1664402	35.99
CM	40420	175692	8.69
Het	8360	15751	1.88
NetS	1588	2742	1.73

TABLE I
DATA-SETS

$IR^{in}(u)$ and external influence $IR^{ex}(u)$. Both values are computed backward from activated vertices under u's influence similarly to Eq 2. Only activated vertex u in the target network receives 1 more influence unit to $IR^{in}(u)$ since we only consider the influence propagation in the target network. Moreover, if a vertex is activated outside the target network, all internal influence is converted to external influence. All the influence relayed through this vertex can be credited for the external support.

VI. EXPERIMENTS

In this section, we show the experimental results to compare the proposed coupling schemes and utilize these coupling schemes to analyze the influence diffusion in multiplex networks. First, we compare lossless and lossy coupling schemes to measure the trade-off between the running time and the quality of solutions. After the networks are coupled, we run the greedy algorithm, which select vertices iteratively based on the marginal gains, to identify the seed set. Despite the simplicity, the greedy algorithm provides consistently the highest quality in all previous researches [12], [13], [7]. Second, we investigate the relationship between networks in the information diffusion to address the following questions: (1) What is the role of overlapping users in diffusing the information? (2) What do we miss when considering each network separately? (3) How and to what extent does the diffusion on one network provide a burst of information in other networks?

A. Datasets

Real networks. We perform experiments on two datasets:

- *Foursquare* (*FSQ*) and *Twitter* networks [18]
- Co-author networks in the area of Condensed Matter(CM) [15], High-Energy Theory(Het) [15], and Network Science (NetS) [16].

The statistics of those networks are described in Table I. The number of overlapping users in the first dataset FSQ-Twitter is 4100 [18]. For the second dataset, we match overlapping users based on authors' names. The numbers of overlapping users of the network pairs CM-Het, CM-NetS, and Het-NetS are 2860, 517, and 90, respectively. While the edge weights are provided for co-author networks, only the topology is available for Twitter and Foursquare networks. Thus we assign the weight of each edge randomly from 0 to 1 for Twitter and Foursqure networks. We then normalize the edge weights so that the total weight of in-coming edges is 1 for each node. This is suitable since the influence of user u on user v tends to be small if v is under the influence of many friends. Finally, the threshold of each node is a random value from 0 to 1.

Synthesized networks. We also use synthesized networks generated by Erdos-Renyi random network model [9] to test on networks with controlled parameters. There are two networks with 10000 nodes which are formed by randomly connecting each pair of nodes with probability $p_1 = 0.0008$ and $p_2 = 0.006$. The average degrees, 8 and 60, reflect the diversity of network densities in reality. Then, we select randomly f fraction of nodes in two networks as overlapping nodes. We shall refer to $0 \leq f \leq 1$ as the *overlapping fraction*. The edge weights and node thresholds are assigned as Twitter.

Setup. We ran all our experiments on a desktop with an Intel(R) Xeon(R) W350 CPU and 12 GB RAM. The number of hops is $d = 4$ and the *influenced fraction* $\beta = 0.8$, unless otherwise mentioned.

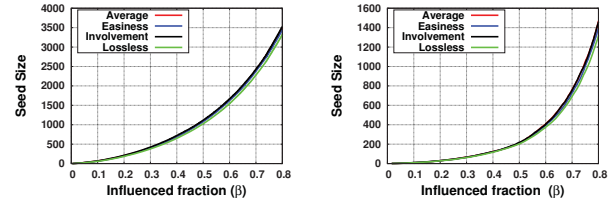
B. Comparison of coupling schemes

We evaluate the impact of the coupling schemes on the running time and the solution quality of the greedy algorithm to solve the LCI problem.

Solution quality. As shown in Figs. 5(a) and 5(b), the greedy algorithm provides larger seed sets but runs faster in lossy coupled networks than lossless coupled networks. In both Twitter-FSQ and the co-author networks, the seed size is smallest when the lossless coupling scheme is used. It is as expected since the lossless coupling scheme reserves all the influence information which is exploited later to solve LCI. However, the seed sizes are only a bit larger using the lossy coupling schemes. In the lossy coupling schemes, the information is only lost at overlapping users which occupy a small fraction the total number of users (roughly 5% in FSQ-Twitter and 7% in co-author networks). Thus, the impact of the lossy coupling schemes on the solution quality is small especially when seed sets are big to influence a large fraction of users.

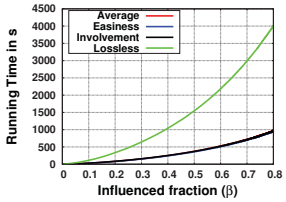
A closer examination reveals the relative effectiveness of the coupling methods on the seed size. That is when the seed size is significantly small, the lossess coupling outperforms all the lossy methods and, in turn, the lossy Easiness method beats the other two lossy methods: Involment and Average. The small seed size is obtained through two different means: 1) increasing the fraction of overlapping users (as shown in Fig. 6(a)) and 2) increasing the number of propagation hops (as shown in Fig. 7). The relative order is significant and consistent in both cases. For example, when the overlapping fraction $f = 0.8$, the solution using the lossless coupling is roughly 55% of that in the solution using the (lossy) Easiness, and the solution using Easiness is about 15% smaller than the other two lossy methods (Fig. 6(a)). Similarly, when the number of propagation hops $d = 5$, the best lossy scheme (Easiness) produces 50% and 26% larger solutions in the co-author networks and FSQ-Twitter (Fig. 7) in comparison to the lossless scheme.

Running time. The greedy algorithm runs much faster in the lossy coupled networks than in the lossless coupled networks, in general. As shown in Figs. 5(c) and 5(d), using the lossy coupled networks reduces the running times by a factor of 2 in FSQ-Twitter and a factor 4 in the co-author networks in comparison to using the lossless coupled networks. The major

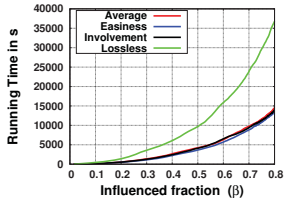


(a) Seed size - co-author networks

(b) Seed size - FSQ-Twitter

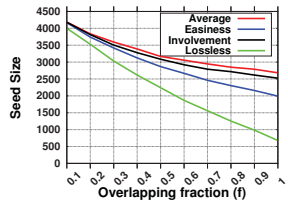


(c) Running time - co-author networks

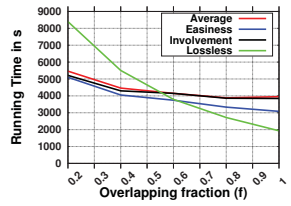


(d) Running time - FSQ-Twitter

Fig. 5. Impact of coupling schemes on finding the minimum seed set



(a) Seed size



(b) Running time

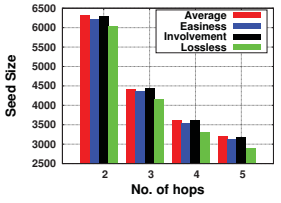
Fig. 6. Comparing coupling schemes in the synthesized networks

disadvantages of the lossless coupling scheme are the doubled number of hops and the number of extra nodes and edges. In the co-author dataset, the number of extra edges is relatively high comparing to the total number of edges in all networks, so the speeding up factor is higher in the co-author networks. We therefore can infer that the lossy coupling schemes work well on real datasets in which networks are sparse and the number of overlapping users is small.

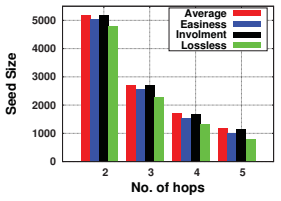
However, it is interesting that the lossless coupling scheme is, in some cases, more efficient in terms of the running time. The running time in the synthesized networks with different overlapping fraction f are shown in Fig. 6(b). The running time in the lossless coupled networks is initially higher than that in the lossy coupled networks but it gradually catches up and overtakes the later networks at $f = 0.4$. The key point is the size of the seed set. The larger f is, the larger the ratio between the seed size in the lossless and lossy coupled networks is. As the running time depends on the seed size, the running time in the lossless coupled network reduces faster.

Overall, the lossless coupling scheme is the method of choice as it leads to higher quality solutions, especially when the seed set is small. However, if the running time and the memory are of priority, the lossy Easiness coupling scheme offers an attractive alternative.

In the next parts, we utilize the lossless coupling method, which leads to the highest quality solutions, to analyze the influence propagation in multiplex networks. First, we show the benefits of using coupled networks and then analyze in details the composition of efficient seed sets to influence users in multiplex networks.

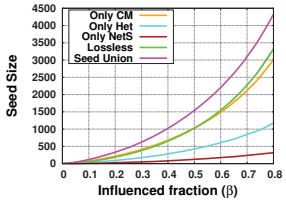


(a) Co-author networks

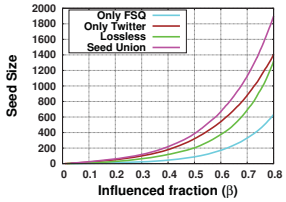


(b) FSQ and Twitter

Fig. 7. Seed size with different number of propagation hops d



(a) Co-author networks



(b) FSQ and Twitter

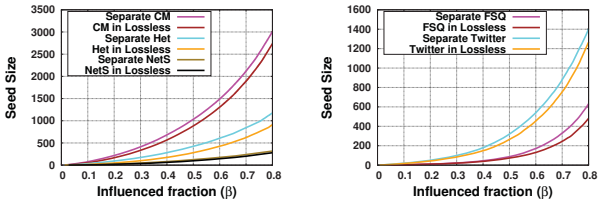
Fig. 8. The quality of seed sets with and without using the coupled network

C. Advantages of using coupled networks.

The coupling schemes provide efficient tools to study the influence propagation in multiplex networks. To understand their benefits, we compare the results obtained using our coupling schemes with the results obtained when each network is examined separately. We consider two different scenarios: 1) when we want to influence a fraction β of the nodes in *all networks* (i.e. the LCI problem); and 2) when we want to influence a fraction β of the nodes in *a particular network*.

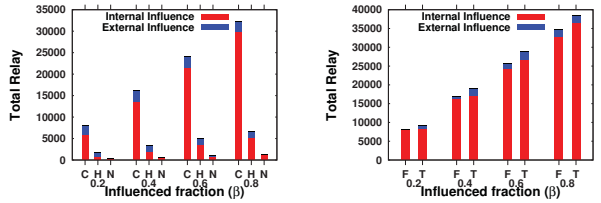
The results for the first scenario are shown in Fig. 8. The results using our lossless coupling method outperform the results when we run the greedy algorithm on each network separately and take the union of the produced seed sets (shown in Fig. 8 as Union). In the co-author networks, the size of the union set is approximately 30% larger than the size of the seed sets found using our coupling method. It is 47% in FSQ-Twitter. The reason is that the coupled network can capture the collaboration of networks to propagate the information and exploit it to reduce the seed size. When we find the seed set in each network separately, we ignore this property. As a result, we endure a penalty on the size of the union set which is high if networks can propagate the information well.

In the second scenario, we modify the greedy algorithm to find the smallest seed set to influence β fraction of a particular network using the lossless coupling method (labeled as 'CM in Lossess', 'Het in Lossless', and so on in Fig. 9) and compare it to the seed set found when the network is considered as a standalone network (labeled as Only CM, Only Het, and so on in Fig. 9). The seed size decreases up to 9%, 25%, 17%, and 26% in CM, Het, FSQ, and Twitter, respectively, when we consider these networks in the connection with other networks. The improvement in NetS is small due to the small number of overlapping users with other networks. When the network sizes are unbalanced, Het – the network with the smaller number of users seems to get better improvement ratio than the bigger network CM. The back and forth propagation of the information between networks forms the base for the outside support of the target network.



(a) Co-author networks (b) FSQ and Twitter

Fig. 9. The quality of seed sets with and without using the coupled network



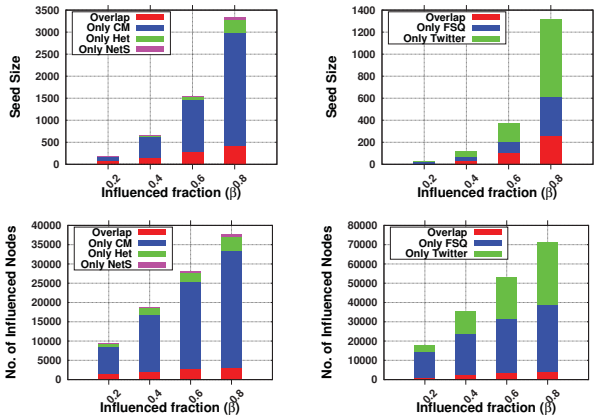
(a) Co-author networks (b) FSQ-Twitter

Fig. 10. The internal/external influence. C, H, N stand for CM, Het, NetS; F and T stand for FSQ and Twitter.

Fig. 10 shows the amount of internal and external influence (presented in Section V) when $d = 4$. The external influence is substantial and accounts for large portions in many cases. For instance, when the influenced fraction $\beta = 0.2$, the external influence accounts for 27.3%, 52.7%, and 30.0% the total influence in CM, Het, and NetS, respectively. *Therefore, it is not sufficient to study each network separately, ignoring the interdependency among the networks.*

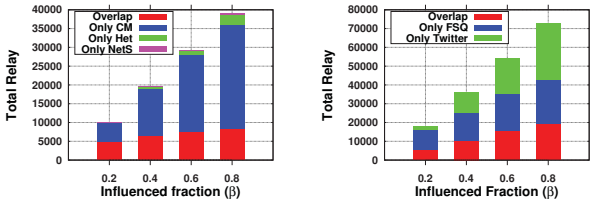
D. Analysis of seed sets

We analyze seed sets with different influenced fraction β to find out: the composition of the seed set and the influenced set; and the influence contribution of each network. As illustrated in Fig. 11, a significant fraction of the seed set is overlapping nodes although only 5% (7%) users of FSQ-Twitter (the co-author networks) are overlapping users. With $\beta = 0.4$, the fraction of overlapping seed vertices is around 24.9% and 25% in the co-author and FSQ-Twitter networks, respectively. As overlapping users can influence friends in different networks, they are more likely to be selected in the seed set than ones



(a) Co-author networks (b) FSQ-Twitter

Fig. 11. The bias in selecting seed nodes



(a) Co-author networks (b) FSQ and Twitter

Fig. 12. The influence contribution of seed vertices from component networks

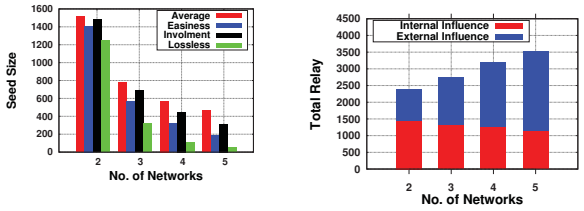
participating in only one network. Fig. 12 demonstrates the high influence contribution of the overlapping users, especially when β is small (contribute more than 50% of the total influence when $\beta = 0.2$). However, when β is large, good overlapping users are already selected, so overlapping users are not favored any more.

Additionally, there is an imbalance between the number of selected vertices and influenced vertices in each networks. In the co-author dataset, CM contributes a large number of seed vertices and influenced vertices since the size of CM is significantly larger than other networks. When $\beta = 0.8$, 76.7% of seed vertices and 80.5% of influenced vertices are from CM. In contrast, the number of seed vertices from FSQ is small but the number of influenced vertices in FSQ is much higher than Twitter. With $\beta = 0.4$, 27% (without overlapping vertices) of seed vertices belong to FSQ while 70% of influenced vertices are in FSQ. After the major of vertices in FSQ are influenced, the algorithm starts to select more vertices in Twitter to increase the influence fraction. This implies that it is easier for the information to propagate in one network than the other, even when we consider the overlapping between them. Moreover, we can target the overlapping users in one network (e.g. Twitter) to influence users in another network (e.g. FSQ).

E. Mutual impact of networks

We evaluate the mutual impact between networks when the number of network k increases. We use a user base of 10000 users to synthesize networks for the experiment. For each network, we randomly select 4000 users from the user base and connect each pair of selected users randomly with probability 0.0025. Thus all networks have the same size and the expected average outgoing (incoming) degree of 10. The expected overlapping fraction of any network pair is 16%. We measure the seed size to influence 60% of users (6000 users) with the different number of networks (Fig. 13(a)). When k increases from 2 to 5, the seed size decreases several times. It implies that the introduction of a new OSN increases the diffusion of information significantly.

We also compute how much new networks help the existing one to propagate the information. Using the same seed set found by the greedy algorithm to influence 60% (2400 users) of the target network (the first created network), we compute the total number of influenced vertices in that network as well as the external influence. Fig. 13(b) shows that the number of influenced vertices is raised 46% with the support of 3 new networks when k is changed from 2 to 5. In addition, the fraction of external influence is also increased dramatically



(a) The impact of additional networks on the seed size

(b) The impact of additional networks on the influence diffusion in a network

Fig. 13. The impact of additional networks

from 39% when $k = 2$ to 67% when $k = 5$. It means that the majority of influence can be obtained via the support of other networks. On the hand, these results suggest that the existing networks may benefit from the newly introduced competitor.

VII. Conclusions

In this paper, we study the least cost influence problem in multiplex networks. To tackle the problem, we introduced novel coupling schemes to reduce the problem to a version on a single network. Then we design a new metric to quantify the flow of influence inside and between networks based on the coupled network. Exhaustive experiments provide new insights to the information diffusion in multiplex networks.

In the future, we plan to investigate the problem in multiplex networks with heterogeneous diffusion models in which each network may have its own diffusion model. Can we still represent them efficiently? Does there exist a method to couple them into one network?

Acknowledgment

This work is supported in part by NSF CAREER 0953284 and DTRA HDTRA-1-10-1-0050.

References

[1] 216 social media and internet statistics. http://thesocialskinny.com/216-social-media-and-internet-statistics-september-2012/.

[2] 99 new social media stats for 2012. http://thesocialskinny.com/99-new-social-media-stats-for-2012/.

[3] Overlap among major social network services. http://www.tomhcanderson.com/2009/07/09/overlap-among-major-social-network-services/.

[4] J. J. Brown and P. H. Reingen. Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, 14(3):pp. 350–362, 1987.

[5] F. Buccafurri, G. Lax, A. Nocera, and D. Ursino. Discovering links among social networks. In *ECML PKDD*. 2012.

[6] N. Chen. On the approximability of influence in social networks. In *SODA*, 2008.

[7] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD*, 2010.

[8] T. N. Dinh, D. T. Nguyen, and M. T. Thai. Cheap, easy, and massively effective viral marketing in social networks: truth or fiction? In *Proceedings of the 23rd ACM conference on Hypertext and social media*, HT '12, 2012.

[9] P. Erdos and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

[10] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[11] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff. Identifying users across social tagging systems. In *ICWSM*, 2011.

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

[13] D. Kempe, J. Kleinberg, and E. Tardos. Influential nodes in a diffusion model for social networks. In *ICALP*, 2005.

[14] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han. Event-based social networks: linking the online and offline social worlds. In *KDD*, 2012.

[15] M. E. Newman. The structure of scientific collaboration networks. *PNAS*, 98(2):404–409, 2001.

[16] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, 2006.

[17] D. T. Nguyen, S. Das, and M. T. Thai. Inuence maximization in multiple online social networks. In *GLOBECOM*, 2013.

[18] Y. Shen, T. N. Dinh, H. Zhang, and M. T. Thai. Interest-matching information propagation in multiple online social networks. In *CIKM*, 2012.

[19] O. Yagan, D. Qian, J. Zhang, and D. Cochran. Information diffusion in overlaying social-physical networks. In *CISS*, 2012.

[20] F. Zou, Z. Zhang, and W. Wu. Latency-bounded minimum influential node selection in social networks. In *WASA*, 2009.

Appendix

Algorithm 1 Computing Influence Relay (CIR)

Input: A network G, a seed set S and the number of hops d.
Output: The influence relay IR of all vertices.

```
IG_S ← The influence graph caused by S on G
for each u ∈ V_S do
    IR(u) ← 0
end for
Compute the topological ordering u_1, u_2, ..., u_{n_S} of vertices in V_S
for i = n_S down to 1 do
    IR(u_i) ← IR(u_i) + 1
    total ← 0
    for each v ∈ N^-(u_i) do
        total ← total + w(v, u_i)
    end for
    for each v ∈ N^-(u_i) do
        IR(v) ← IR(v) + w(v,u_i) IR(u_i) / total
    end for
end for
Return IR
```

Time complexity. The topological ordering of a directed acyclic graph can be computed in linear time and the number of updates in the main loop equals to the number of edges of IG_S, so the *CIR* algorithm runs in linear time.

Proof of Theorem 4. The proof is based on an invariant of variables $IR(u_1), \ldots, IR(u_n)$ in *CIR* algorithm. The information is propagated from the seed set, thus all seed vertices do not have incoming neighbors in IG_S and occupy smallest indices in the topological ordering. Let u_p be the highest index seed vertex. We will prove that after the loop $i = k + 1$ we have:

$$\sum_{j=1}^{k} IR(u_j) = n_S - k, \forall p \leq k \leq n_S$$

Before the loop $i = n$, it is obviously true. After the loop $i = k$, the value of variable $IR(u_{k+1})$ is increased by 1 and redistributed to its incoming neighbors, thus $\sum_{j=1}^{k-1} IR(u_j)$ equals $\sum_{j=1}^{k} IR(u_j)$ plus 1. It implies that $\sum_{j=1}^{k-1} IR(u_j) = n_S - (k-1)$ after the loop $i = k$.

After the loop $i = p + 1$, we have $\sum_{i=1}^{p} IR(u_i) = n_S - p$. At each loop $i = p$ down to $i = 1$, the value of $IR(u_i)$ is increased by 1. Thus, when the algorithm stops we have:

$$\sum_{u \in S} IR(u) = \sum_{i=1}^{p} IR(u_i) = n_S - p + p = |A^d(G, S)| \quad \square$$