Discover Tipping Users For Cross Network Influencing

Qianyi Zhan*, Jiawei Zhang†, Philip S. Yu†, Sherry Emery† and Junyuan Xie*
* National Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
† University of Illinois at Chicago, Chicago, IL, USA
Email: zhanqianyi@gmail.com, jzhan9@uic.edu,
psyu@uic.edu, slemery@uic.edu, jyxie@nju.edu.cn

***Abstract*—Traditional viral marketing problem aims at selecting a set of influential seed users to maximize the awareness of products and ideas in one single social network. However, in real scenarios, users' profiles in the target social network (e.g., Facebook) are usually confidential to the public, which block the conventional viral marketing strategies reaching the target consumers effectively. Instead, since users nowadays are usually involved in multiple social networks simultaneously, the viral marketing can actually be performed in other public networks. These networks with public profile information are referred as the source networks, from which information can diffuse to and activate users in the target network indirectly. Thus in the cross-network information diffusion, besides the influential seed users, those who act as bridges propagating information between networks actually play a more important role and some can trigger the *tipping point* in the target network, who are named as the *tipping users* formally.**

Motivated by this, in this paper, we studied the "Discovering Tipping Users for Cross Network Influencing" (TURN) problem across multiple aligned heterogeneous social networks. To depict the information diffusion process across aligned heterogeneous social networks, we propose a novel network information diffusion model, "Cross Network Information Diffusion" (CONFORM). In CONFORM, various diffusion links in the heterogeneous networks are extracted and fused by weight to calculate the users' activation probabilities. To address the TURN problem, a new method called "Tipping Users Discovery Algorithm (TUDOR)" is proposed to identify the tipping users who bring about the largest influence gain, which is a new concept first introduced in this paper. Extensive experiments are conducted on real-world social network datasets, which demonstrate the effectiveness and efficiency of TUDOR.

I. INTRODUCTION

One of the great advantages about the online social networks is that when an idea takes off, it can propel a brand to seemingly instant fame and fortune with quite a low cost [17]. This is called *viral marketing* based on the *word of mouth communication* among users in the networks [8]. Traditional viral marketing problems and methods mainly focus on selecting the *seed users* for the products and ideas to be promoted within *one single network* only [9].

However in the real scenarios, it is difficult to apply viral marketing on some online social networks which create intimate and private communicating circles within the users' choice of close friends. For example, since 2012, Centers for Disease Control and Prevention (CDC) has launched the national tobacco education campaign "Tips From Former Smokers"[1] to encourage smokers to quit. One of the aims of the campaign is to enlarge their exposures and influence in online social networks such as Facebook[2] and Twitter[3]. The campaign has achieved great success, however, the advertising effect in Facebook is not as good as other networks. HMC[4] from UIC, which is supported to evaluate the advertising effects, explains the reason that Facebook enables users to choose their own privacy settings and choose who can see specific parts of their profile. Therefore due to the privacy and security policies of the network, ad companies do not have access to users' profiles and cannot easily get word out to the audience whom they most wish to reach. Hence traditional single-network viral marketing strategies which influence target users directly can no longer perform well.

Meanwhile, users nowadays are usually involved in multiple online social networks simultaneously, and those joining in Facebook are also using other networks, such as Twitter, Foursquare and Instagram, at the same time. These social networks, whose information is more public and users are much easier to reach, provide good channels for ad companies to communicate with audience. Moreover, most of them have provided the services that users can share the posts, images and videos to their Facebook homepage effortlessly. In such a roundabout way, information can be propagated to the Facebook network that we target on from these public social networks. In other words, viral marketing can actually be performed in these public social networks instead, i.e., information from which can diffuse to and activate users in the target network indirectly. To differentiate them from the target social network (e.g., Facebook), these social networks with easy access are named as the *source networks* in this paper.

However, in such a cross-network viral marketing setting, besides the influence sources (i.e., seed users) to be selected in traditional viral marketing problems, the anchor users who acts as the bridges to propagate information between the source networks and the target network play a more significant role.

[1] http://www.cdc.gov/tobacco/campaign/tips/index.html
[2] https://www.facebook.com/cdctobaccofree/
[3] https://twitter.com/CDCTobaccoFree
[4] https://www.healthmediacollaboratory.org/

In traditional sociology studies, the concept *tipping point* denotes a time point when a large number of group members rapidly and dramatically change their behavior by widely adopting a previously rare practice [7]. Actually, such mysterious behavior changes of individuals are happening in online social networks everyday, e.g., the sudden emergence of fashion trends [23], the quick swing of public opinion [14], as well as the transformation of an unknown person to be the center of attention [3]. If we let the rare social action be the individuals' purchase of the products in the target network, the *tipping point* leading to the massive adoption of the target products will be of great interests to the companies carrying out the promotion activities. In addition, when triggering the *tipping point* in the target network, those critical anchor users (in the source networks) who occupy the crucial positions in diffusing information across the networks are formally named as the *tipping users*.

In this paper, we aim at identifying the *tipping users* from the source networks, and the problem is formally referred as "discovering **T**ipping **U**sers for c**R**oss **N**etwork influencing" (TURN) problem.

The TURN problem studied in this paper is a novel problem, and it is totally different from the conventional information diffusion ([1], [6], [13]) and viral marketing problems ([2], [4], [24]). Instead of finding the influence sources (i.e., the seed users), the TURN problem aims at identifying the tipping users from the source network. Therefore, the models proposed for traditional viral marketing problems cannot be applied to address the TURN problem.

Despite its importance and novelty, the TURN problem studied in this paper is very challenging to solve due to the following reasons:

- Problem Definition: To the best of our knowledge, we are the first to propose the TURN problem. A clear definition of the *tipping user* concept as well as the TURN problem is needed, which is still an open problem to this context so far.
- Information Diffusion Model across Heterogeneous Networks: The networks studied in this paper are all heterogeneous information networks, involving various types of nodes and complex intra-network connections. Moreover, due to the shared common users, these networks are partially aligned via the anchor links [29] as well. How to effectively model the information diffusion process across the aligned heterogeneous networks is a great challenge.
- NP-hard: The TURN problem based on our cross-network diffusion model to be introduced in Section 3 is proved to be NP-hard, thus there exist no polynomial-time algorithms for TURN if P $\neq$ NP.

To address these challenges, we propose a novel method "**T**ipping **U**sers **D**iscovery alg**OR**ithm" (TUDOR) in the paper. This method is based on a novel information diffusion model, named "**C**r**O**ss **N**etwork in**FORM**ation diffusion" (CONFORM), proposed in this paper. CONFORM effectively aggregates both intra-network and inter-network diffusion links , which are weighted differently according to their importance in the aggregation. Based on CONFORM, the TURN problem is proved to be NP-hard. However TUDOR applies a novel step-wise greedy algorithm to identify the tipping users, which can resolve the TURN problem efficiently and is proven to achieve a $(1-1/e)$-approximation of the optimal result.

The rest of this paper is organized as follows. In Section 2, we give the concept definitions and problem formulation. In Section 3 and 4, the CONFORM model is introduced in detail. The TUDOR method is proposed in Section 5. We evaluate the performance of TUDOR with extensive experiments in Section 6. Finally, we give the related works in Section 7 and conclude the paper in Section 8.

II. PROBLEM FORMULATION

A. Basic Terms

In this paper, we will follow the definitions of concepts *"heterogeneous networks"*, *"anchor user" "anchor link"*, etc. proposed in [29].

DEFINITION 1. ***Partially aligned heterogeneous networks:*** *A pair of partially aligned heterogeneous networks is* $\mathcal{H} = (G^{(s)}, G^{(t)}, \mathcal{L})$, *where* $G^{(s)} = (\mathcal{V}^{(s)}, \mathcal{E}^{(s)})$ *is the source network and* $G^{(t)} = (\mathcal{V}^{(t)}, \mathcal{E}^{(t)})$ *is the target network, and their user sets are represented as* $\mathcal{U}^{(s)} \subset \mathcal{V}^{(s)}$ *and* $\mathcal{U}^{(t)} \subset \mathcal{V}^{(t)}$ *respectively.* $\mathcal{L}$ *denotes the set of anchor links which connect anchor user sets* $\mathcal{A}^{(s)} \subseteq \mathcal{U}^{(s)}$ *of* $G^{(s)}$ *and* $\mathcal{A}^{(t)} \subseteq \mathcal{U}^{(t)}$ *of* $G^{(t)}$.

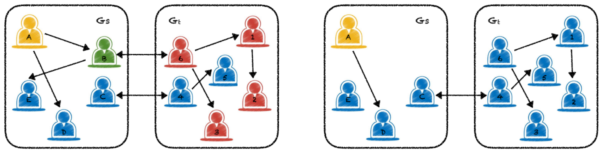
In the cross-network information diffusion process, messages propagate in discrete steps beginning with a group of seed users $\mathcal{S} \subseteq \mathcal{U}^{(s)}$ in the source network, to the target network $G^{(t)}$ via both the *heterogeneous intra and inter network links*. Users in networks have two statuses: *active* and *inactive*, where *active* status represents the user has adopted the certain product. The notations used in this paper are listed in Table I.

Table I: Notation

Notation	Definition
$G^{(s)}$, $G^{(t)}$	the heterogeneous source or target network
$\mathcal{A}^{(s)}$, $\mathcal{A}^{(t)}$	set of anchor users in the source or target network
$\mathcal{L}$	set of anchor links connected the source and the target network
$\mathcal{H}$	partially aligned heterogeneous networks
$\mathcal{S}$	set of seed users which all belong to the source network, i.e., $\mathcal{S} \in \mathcal{U}^{(s)}$
$\mathcal{Z}$	set of tipping users which are all anchor users in the source network, i.e., $\mathcal{Z} \in \mathcal{A}^{(s)}$
$\delta(\mathcal{S}\|\mathcal{H})$	influence function which maps the seed user set $\mathcal{S}$ to the number of activated users in the target network
$G^{(s)}_{\mathcal{Z}}$	the reduced source network
$\mathcal{H}_{\mathcal{Z}}$	the reduced partially aligned heterogeneous networks

B. Tipping Users and Influence Gain

As introduced in the previous section, in the cross-network information diffusion process, the existence of certain users occupying the crucial positions in the source network can trigger the *tipping point* [7] of the massive adoption of the product in the target network $G^{(t)}$. These



(a) Original Network Structure (b) Reduced Network Structure

Figure 1: Reduced network structure and cross network influence gain

users who are a group of anchor users and cause a substantial increase of the number of activated users in the target network $G^{(t)}$ are called the *tipping users* in this paper.

To identify *tipping users*, a new metric, called *cross network influence gain*, is proposed to evaluate the contribution of $\mathcal{Z}$ and it is based on the *reduced network* and the *influence function*.

Definition 2. ***Reduced Network:*** *Given a heterogeneous network $G = (\mathcal{V}, \mathcal{E})$ and an group of anchor users $\mathcal{Z} \subseteq \mathcal{A} \setminus \mathcal{S}$, the reduced network is $G_{\mathcal{Z}} = (\mathcal{V} - \mathcal{Z}, \mathcal{E} - \mathcal{E}_{\mathcal{Z}})$, where $\mathcal{E}_{\mathcal{Z}} = \{(v, u) | v \in \mathcal{Z} \vee u \in \mathcal{Z}\}$ is the set of edges connecting with users in $\mathcal{Z}$.*

Definition 3. ***Reduced Partially Aligned Networks:*** *Given a pair of partially aligned heterogeneous networks $\mathcal{H} = (G^{(s)}, G^{(t)}, \mathcal{L})$ and a user set $\mathcal{Z} \subseteq \mathcal{U}^{(s)} \setminus \mathcal{S}$, the reduced partially aligned network is $\mathcal{H}_{\mathcal{Z}} = (G^{(s)}_{\mathcal{Z}}, G^{(t)}, \mathcal{L} - \mathcal{L}_{\mathcal{Z}})$, where $G^{(s)}_{\mathcal{Z}}$ is the reduced network, and $\mathcal{L}_{\mathcal{Z}} = \{(v, u) | v \in \mathcal{Z}\}$ is the set of anchor links connecting with users in $\mathcal{Z}$.*

The *reduced networks* $G_{\mathcal{Z}}$ and $\mathcal{H}_{\mathcal{Z}}$ actually denote the residual network structures after removing users in $\mathcal{Z}$ and all the attached edges from networks G and H respectively.

Definition 4. ***Influence Function:*** *Given a pair of partially aligned heterogeneous social networks $\mathcal{H} = (G^{(s)}, G^{(t)}, \mathcal{L})$ and a seed user set $\mathcal{S} \subseteq \mathcal{U}^{(s)}$, let $\delta(\mathcal{S}|\mathcal{H})$ be the influence function which maps the seed set $\mathcal{S}$ in $G^{(s)}$ to the number of activated users in the target network $G^{(t)}$ by the same seed user set $\mathcal{S}$ based on $\mathcal{H}$.*

Therefore the *cross network influence gain* is defined as:

Definition 5. ***Cross Network Influence Gain*** *Given a pair of partially aligned heterogeneous social networks $\mathcal{H} = (G^{(s)}, G^{(t)}, \mathcal{L})$, a pre-determined seed user set $\mathcal{S} \subseteq \mathcal{U}^{(s)}$, and a group of anchor user $\mathcal{Z} \in \mathcal{A}^{(s)} \setminus \mathcal{S}$, $\delta(\mathcal{S}|\mathcal{H}_{\mathcal{Z}})$ is the number of activated users in $G^{(t)}$ based on the reduced network $\mathcal{H}_{\mathcal{Z}}$. The cross network influence gain of user set $\mathcal{Z}$ is defined as the difference between $\delta(\mathcal{S}|\mathcal{H})$ and $\delta(\mathcal{S}|\mathcal{H}_{\mathcal{Z}})$, denoted as $I(\mathcal{Z}) = \delta(\mathcal{S}|\mathcal{H}) - \delta(\mathcal{S}|\mathcal{H}_{\mathcal{Z}})$.*

Figure 1 explains the concepts of *reduced network structure* and *cross network influence gain*. In both Fig. 1(a) and Fig. 1(b), the left rectangle represents the source network $G^{(s)}$ and the right one denotes the target network $G^{(t)}$. Each not yet activated user is represented by a *blue* icon, and the icon becomes *red* when the user is activated. Let *yellow* icons be seed users and *green* icons denote the candidates of tipping users. Thus, in our example, based on seed set $\mathcal{S} = \{A\}$, we aim to discover one tipping users ($k = 1$) from two anchor users of $G^{(s)}$ and user B is current candidate, i.e., $\mathcal{Z} = \{B\}$. The Fig. 1(a) shows the original network structure $\mathcal{H} = (G^{(s)}, G^{(t)}, \mathcal{L})$, and under this circumstance, four users $\{1, 2, 3, 6\}$ in $G^{(t)}$ can be activated, i.e., $\delta(\mathcal{S}|\mathcal{H}) = 4$. Consider, for instance, if we remove user B and all edges connecting it (including the anchor link), the obtained *reduced network structure* can be represented as $\mathcal{H}_{\mathcal{Z}} = (G^{(s)}_{\mathcal{Z}}, G^{(t)}, \mathcal{L} - \mathcal{L}_{\mathcal{Z}})$, as shown in Fig. 1(b). Due to the removal of B, information can no longer be propagated from the seed user A to users in $G^{(t)}$. In other words, no user in $G^{(t)}$ can be activated and $\delta(\mathcal{S}|\mathcal{H}_{\mathcal{Z}}) = 0$. Therefore, the cross network influence gain introduced by $\mathcal{Z}$ is $I(\mathcal{Z}) = I(\{B\}) = \delta(\mathcal{S}|\mathcal{H}) - \delta(\mathcal{S}|\mathcal{H}_{\mathcal{Z}}) = 4$.

C. Problem Formulation

Based on the above concepts, in this paper, tipping users are formally defined as a group of anchor users who has the largest cross network influence gain, which means without tipping users, the number of activated users in the target network will decrease significantly.

Definition 6. ***Tipping Users:*** *Given a pair of partially aligned heterogeneous networks $\mathcal{H} = (G^{(s)}, G^{(t)}, \mathcal{L})$ and the seed user set $\mathcal{S} \subseteq \mathcal{U}^{(s)}$, tipping users are a group of anchor users who can maximize the cross-network influence gain in the source network, i.e., $\mathcal{Z} = argmax_{\mathcal{Z}} I(\mathcal{Z})$, where $\mathcal{Z} \subseteq \mathcal{A}^{(s)} \setminus \mathcal{S}$.*

Therefore based on above definitions, we formulate the problem of "discovering **T**ipping **U**sers for c**R**oss **N**etwork influencing" (TURN) as follows:

Definition 7. ***The* TURN *Problem:*** *Given the partially aligned heterogeneous networks $\mathcal{H} = (G^{(s)}, G^{(t)}, \mathcal{L})$, the seed user set $\mathcal{S} \subseteq \mathcal{U}^{(s)}$ and the per-specified number of tipping users k, the* TURN *problem aims at discovering k tipping users in the source network, $\mathcal{Z}^* \subseteq \mathcal{A}^{(s)} \setminus \mathcal{S}$, who can maximize the cross-network influence gain $I(Z)$, i.e.,.*

$$\mathcal{Z}^* = argmax_{\mathcal{Z}} I(\mathcal{Z}) = argmax_{\mathcal{Z}} \delta(\mathcal{S}|\mathcal{H}) - \delta(\mathcal{S}|\mathcal{H}_{\mathcal{Z}}) \quad (1)$$

III. Information Diffusion Model

In the CONFORM model, information propagation process is traced by *meta paths*, which are based on the *network schema*.

Definition 8. ***Network Schema:*** *Given a heterogeneous network $G = (\mathcal{V}, \mathcal{E})$, its network schema is defined as $S = (\mathcal{O}, \mathcal{R})$, where $\mathcal{O}$ and $\mathcal{R}$ denote the type sets of entities and links in G respectively.*

Definition 9. ***Meta Path:*** *A meta path Ω, based on the given network schema $S = (\mathcal{O}, \mathcal{R})$, is denoted in the form of $\Omega = o_1 \xrightarrow{r_1} o_2 \xrightarrow{r_2} \cdots \xrightarrow{r_{k-1}} o_k$ where entity $o_1, o_k = User \in \mathcal{O}$, $o_i \in \mathcal{O} - \{User\}, i = 2, \cdots, k-1$ and $r_i \in \mathcal{R}, i \in \{1, 2, \cdots, k-1\}$.*

Table II: Atomic Meta Paths of Foursquare and Twitter

Link Type	Network	Physical Meaning	Meta Path
Intra-network	Foursquare	follow	User $\xrightarrow{follow^{-1}}$ User
		co-location check-ins	User $\xrightarrow{checkin}$ Location $\xrightarrow{checkin^{-1}}$ User
		co-location via shared lists	User $\xrightarrow{create/like}$ List $\xrightarrow{contain}$ Location $\xrightarrow{contain^{-1}}$ List $\xrightarrow{create/like^{-1}}$ User
	Twitter	follow	User $\xrightarrow{follow^{-1}}$ User
		co-location check-ins	User $\xrightarrow{checkin}$ Location $\xrightarrow{checkin^{-1}}$ User
		contact via tweet	User $\xrightarrow{write}$ Tweet $\xrightarrow{retweet}$ Tweet $\xrightarrow{write^{-1}}$ User
Inter-network		anchor	User of Twitter $\xrightarrow{Anchor}$ User of Foursquare
		anchor	User of Foursquare $\xrightarrow{Anchor}$ User of Twitter

In a meta path, the types of start and end entity are both *User* and types of other entities are not *User*. The instances of meta paths are called *diffusion links*, which start and end with two specific users. It is obvious that network schemes, meta paths and diffusion links of diverse networks are quite different. Therefore information propagation inside each network and across two networks should be considered respectively, thus we classify meta paths into two categories: intra-network meta path and inter-network meta path.

DEFINITION 10. ***Intra-network and Inter-network Meta Path:*** *Given partially aligned heterogeneous networks $\mathcal{H} = (G^{(s)}, G^{(t)}, \mathcal{L})$ with their network schemes $S^{(s)} = (\mathcal{O}^{(s)}, \mathcal{R}^{(s)})$ and $S^{(t)} = (\mathcal{O}^{(t)}, \mathcal{R}^{(t)})$. To a meta path $\Omega^{(i)} = o_1 \xrightarrow{r_1} o_2 \xrightarrow{r_2} \cdots \xrightarrow{r_{k-1}} o_k$, if the start and end users belong to the same network, i.e., $o_1, o_k = User \in \mathcal{O}^{(i)}$, the path is called an intra-network meta path. While when the start and end users of a meta path belong to different networks, i.e., $o_1 = User \in \mathcal{O}^{(i)}$, $o_k = User \in \mathcal{O}^{(j)}$, the path is called an inter-network meta path.*

Besides following links, a heterogeneous online social network usually has various kinds of intra-network meta paths. Let $m^{(i)}$ be the number of intra-network meta paths, and $\Omega_k^{(i)}$ denotes the kth path of the network $G^{(i)}$. The set of links connecting two users u and v in the same network $G^{(i)}$ by a specific meta path $\Omega_k^{(i)}$ is represented as $\mathcal{Q}_k^{(i)}(u \rightsquigarrow v)$. We define $\omega_k^{(i)}(u,v)$ as the amount of information propagating from u to v through meta path $\Omega_k^{(i)}$, which is calculated as:

$$\omega_k^{(i)}(u,v) = \frac{2|\mathcal{Q}_k^{(i)}(u \rightsquigarrow v)|}{|\mathcal{Q}_k^{(i)}(u \rightsquigarrow \cdot)| + |\mathcal{Q}_k^{(i)}(\cdot \rightsquigarrow v)|},$$

where $\mathcal{Q}_k^{(i)}(u \rightsquigarrow \cdot)$, $\mathcal{Q}_k^{(i)}(\cdot \rightsquigarrow v)$ are the sets of diffusion links with $\Omega_k^{(i)}$ which start from u and end at v respectively.

We now aggregate information received from the above channels. For a specific user $v \in \mathcal{U}^{(i)}$, the sum of information from all intra-network diffusion links denoted as $W_v^{(i),(intra)}$, is defined as follows.

$$W_v^{(i),(intra)} = \sum_{k=1}^{m^{(i)}} \alpha_k^{(i)} \times \sum_{u \in \Gamma_k^{(i)}(v)} \omega_k^{(i)}(u,v) \tag{2}$$

where $\Gamma_k^{(i)}(v)$ is the set of neighbors connected to v with intra-network links of meta path type $\Omega_k^{(i)}$, and $\alpha_k^{(i)}$ is the weight of $\Omega_k^{(i)}$ in the linear aggregation.

As to inter-network meta paths, similar to intra-network ones, let $\Omega_k^{(j,i)}$ be the kth inter-network meta path from the network $G^{(j)}$ to $G^{(i)}$, and the number of inter-network meta paths is $m^{(j,i)}$. Therefore the amount of information received by $v \in \mathcal{U}^{(i)}$ from $u \in \mathcal{U}^{(j)}$ through inter-network diffusion links with meta path type $\Omega_k^{(j,i)}$ is

$$\omega_k^{(j,i)}(u,v) = \frac{2|\mathcal{Q}_k^{(j,i)}(u \rightsquigarrow v)|}{|\mathcal{Q}_k^{(j,i)}(u \rightsquigarrow \cdot)| + |\mathcal{Q}_k^{(j,i)}(\cdot \rightsquigarrow v)|}$$

Therefore, the total amount of information propagated by v via inter-network diffusion links from the other networks is:

$$W_v^{(i),(inter)} = \sum_{k=1}^{m^{(j,i)}} \beta_k^{(j,i)} \times \sum_{u \in \Gamma_k^{(j,i)}(v)} \omega_k^{(j,i)}(u,v) \tag{3}$$

where $\Gamma_k^{(j,i)}(v)$ is the set of neighbors in $G^{(j)}$ connected to $v \in \mathcal{U}^{(i)}$ with inter-network diffusion links of meta path type $\Omega_k^{(j,i)}$, and $\beta_k^{(j,i)}$ is the weight of $\Omega_k^{(j,i)}$.

We take Foursquare and Twitter as examples of partially aligned heterogeneous networks. Inside both networks, users can follow others and check-in at locations. Meanwhile, (1) in Foursquare, users can create/like lists containing a set of locations; (2) while in Twitter, users can repost other users' tweets. The intra-network meta paths considered in this paper as well as their physical meanings are listed in Table (II). These two networks are connected by anchor links. The intra-network and inter-network meta paths considered in this paper as well as their physical meanings are listed in Table (II).

Finally we can define the activation function $p_v = f(\cdot) : \mathbb{R} \to \{0,1\}$ and logistic function $f(x) = \frac{e^x}{1+e^x}$ is used in this paper. This function maps the received information amount to activation probability of user v of $G^{(i)}$, and is calculated as:

$$p_v^{(i)} = \frac{e^{(W_v^{(i),(intra)} + W_v^{(i),(inter)})}}{1 + e^{(W_v^{(i),(intra)} + W_v^{(i),(inter)})}} \tag{4}$$

IV. Diffusion Links Weighting

In the last section, we extract different diffusion links to describe the information propagation process in partially aligned heterogeneous networks and we set weights to these diffusion links when aggregate information. In this section, the optimal weight values are learned from the user activation log data.

In the CONFORM model, each weight measures the significance of correspond link in the diffusion process, thus different links will be ranked according to their importance, and top links will be selected to increase individuals' activation probabilities.

For the network $G^{(t)}$, we construct a column vector $\mathbf{P}^{(t)} \in \mathbb{R}^{|\mathcal{U}^{(t)}| \times 1}$, where $\mathbf{P}^{(t)}[v]$ records user v's activation probability, calculated by (4). At the same time ground-truth extracted from the log data is represented by a binary vector $\mathbf{T}^{(t)} \in \mathbb{R}^{|\mathcal{U}^{(t)}| \times 1}$, which $\mathbf{T}^{(t)}[v] = 1$ means user v is activated finally, otherwise $\mathbf{T}^{(t)}[v] = 0$. The learned weights aim to narrow the gap between the prediction and the ground-truth, i.e., $\|\mathbf{P}^{(t)} - \mathbf{T}^{(t)}\|_F^2$, where $\| \cdot \|_F$ is the *Frobenius norm* of the matrix.

The other concern is that the behavior of an anchor user should be consistent. Though we treat an anchor user as two independent users in respective networks, it is intuitive that one person shows the consistent interest to the same topic in real life. Therefore if anchor user $v^{(s)}$ in $G^{(s)}$ is active, there is a high probability for v in other networks, such as $G^{(t)}$. Thus the values of anchor user $\mathbf{P}^{(s)}[v]$ and $\mathbf{P}^{(t)}[v]$ should be close. The restriction is the sum of all weights of diffusion links which connect to a specific network $G^{(t)}$ or $G^{(s)}$ should equals 1 and all weights should be no-negative.

Therefore, the optimal weights of information delivered in different diffusion links can be obtained by solving the following objective function:

$$\begin{aligned} \min\ & \|\mathbf{P}^{(s)} - \mathbf{T}^{(s)}\|_F^2 + \|\mathbf{P}^{(t)} - \mathbf{T}^{(t)}\|_F^2 \\ & + \|(\mathbf{A}^{(t,s)})^T \mathbf{P}^{(t)} - (\mathbf{A}^{(t,s)})^T \mathbf{A}^{(t,s)} \mathbf{P}^{(s)}\|_F^2 \\ \text{s.t.}\ & \sum_{k=1}^{m^{(s)}} \alpha_k^{(s)} + \sum_{k=1}^{m^{(t,s)}} \beta_k^{(t,s)} = 1, \\ & \sum_{k=1}^{m^{(t)}} \alpha_k^{(t)} + \sum_{k=1}^{m^{(s,t)}} \beta_k^{(s,t)} = 1, \\ & \alpha_1^{(s)}, \cdots, \alpha_{m^{(s)}}^{(s)}, \alpha_1^{(t)}, \cdots, \alpha_{m^{(t)}}^{(t)} \geq 0, \\ & \beta_1^{(t,s)}, \cdots, \beta_{m^{(t,s)}}^{(t,s)}, \beta_1^{(s,t)}, \cdots, \beta_{m^{(s,t)}}^{(s,t)} \geq 0. \end{aligned} \tag{5}$$

where anchor matrix $\mathbf{A}^{(t,s)} \in \mathbb{R}^{|\mathcal{U}^{(t)}| \times |\mathcal{U}^{(s)}|}$ represents the anchor users of two networks, which can be constructed based on the anchor link set $\mathcal{L}$ and $\sum_i \mathbf{A}^{(t,s)}[v][i] = 1$ denotes v is an anchor user. The matrix $(\mathbf{A}^{(t,s)})^T \mathbf{P}^{(t)} - (\mathbf{A}^{(t,s)})^T \mathbf{A}^{(t,s)} \mathbf{P}^{(s)}$ extracts the difference of activation probability of anchor users in $\mathbf{P}^{(s)}$ and $\mathbf{P}^{(t)}$. The constraints make sure users' activation probabilities in each network are all between the range $[0, 1]$ and all weights are no-negative.

By the triangle inequality and positive homogeneity, every norm is a convex function. Since constraints have inequalities, we need to extend the *Lagrange Multiplier Method* to the *Karush-Kuhn-Tucker (KKT)* conditions to find local and also globe minim of the objective function:

$$\begin{aligned} \mathcal{L}(\alpha, \beta, \lambda) = & \|\mathbf{P}^{(s)} - \mathbf{T}^{(s)}\|_F^2 + \|\mathbf{P}^{(t)} - \mathbf{T}^{(t)}\|_F^2 \\ & + \|(\mathbf{A}^{(t,s)})^T \mathbf{P}^{(2)} - (\mathbf{A}^{(t,s)})^T \mathbf{A}^{(t,s)} \mathbf{P}^{(s)}\|_F^2 \\ & + \lambda_1 (\sum_{k=1}^{m^{(s)}} \alpha_k^{(s)} + \sum_{k=1}^{m^{(t,s)}} \beta_k^{(t,s)} - 1) \\ & + \lambda_2 (\sum_{k=1}^{m^{(t)}} \alpha_k^{(t)} + \sum_{k=1}^{m^{(s,t)}} \beta_k^{(s,t)} - 1) \\ & + \sum \mu_i \alpha_i + \sum \nu_i \beta_i \end{aligned} \tag{6}$$

Setting the gradient $\nabla_{\alpha,\beta,\lambda,\mu,\nu} \mathcal{L}(\alpha, \beta, \lambda, \mu, \nu) = 0$ can yield a group of equations about variables $\alpha, \beta, \lambda, \mu$ and ν. They are implicit functions and can be solved with the source toolkit, e.g., *Scipy nonlinear solver*. Given the parameters with initial values, multiple solutions can be obtained by resolving the objective function.

V. Proposed Algorithm

In this section, the TUDOR method is proposed to address the TURN problem. Before introducing TUDOR, we first analyze the TURN problem.

A. Problem Analysis

THEOREM 1. *The* TURN *Problem based on the* CONFORM *model is NP-hard.*

Proof. In the TURN problem, since when the seed user set $\mathcal{S}$ is pre-determined, its final activate user set, denoted as $\mathcal{D}$ is fixed. Consider an instance of the NP-hard *Vertex Cover* problem defined by a graph $G = (\mathcal{D}, E)$ and integer k. It aims to identify a set of k vertices $\mathcal{C}$ such that each edge of the graph is incident to at least one vertex of the $\mathcal{D}$. If there is a vertex cover $\mathcal{C}$ of size k in $G^{(s)}$, we can block all activated users, i.e.$\mathcal{D}$, by setting the tipping user set $\mathcal{Z} = \mathcal{C}$. This shows the *Vertex Cover* problem can be viewed as a special case of the TURN problem, therefore, the TURN problem is NP-hard. □

Therefore there is no known polynomial algorithm which can give the optimal solution to the TURN problem and the brute force method tends to grow very quickly as the size of the network increases. However the influence gain function is proofed to be monotone and submodular motivated by [9].

THEOREM 2. *The influence gain function $I(\mathcal{Z})$ is monotone.*

Proof: To a specific u, the amount of information got from any intra-network diffusion link $\Omega_k^{(i)} \geq 0$ and the its corresponding weight $\alpha_k^{(i)} \geq 0$. Therefore the sum of information received from all intra-network diffusion links $W_v^{(i),(intra)} \geq 0$ and the function (2) is monotone.

Similarly, $W_v^{(j,i),(inter)} \geq 0$ and the function (3) is also monotone. Since the logistic function is monotone, the activation probability function (4) is monotone too, which deducting more links will descend the value of $p_v^{(i)}$. So if we block more users, the amount of activated users will not increase, in other words, when $\mathcal{Z}^1 \subset \mathcal{Z}^2$, their influence gain $I(\mathcal{Z}^1) \leq I(\mathcal{Z}^2)$. Therefore the influence gain function is monotone.

Corollary 1. *The influence gain function $I(\mathcal{Z})$ is non-negative.*

Proof: If $\mathcal{Z} = \emptyset$, $I(\mathcal{Z}) = 0$, which means no active user will be blocked when there is no tipping users. As we proved above, the influence gain function $I(\mathcal{Z})$ is monotone, therefore given any tipping user set $\mathcal{Z}$, $I(\mathcal{Z}) \geq 0$.

Theorem 3. *The influence gain function $I(\mathcal{Z})$ is submodular.*

Proof: Given a pre-determined seed user set $\mathcal{S}$ in $G^{(s)}$, its final active user set $\mathcal{D}$ are fixed. Our aim is to discover a subset of anchor users who can block most of active users in the target network. According to the Claim 2.6 in [9], the block set obtained by running the CONFORM model, which is similar to the LT model, is equivalent to reachability via live-edge paths defined in [9]. Moreover, an active user u in $G^{(t)}$ can be blocked if and only if there is a live-edge path from some node in $\mathcal{Z}$ to u based on Claim 2.3 in [9].

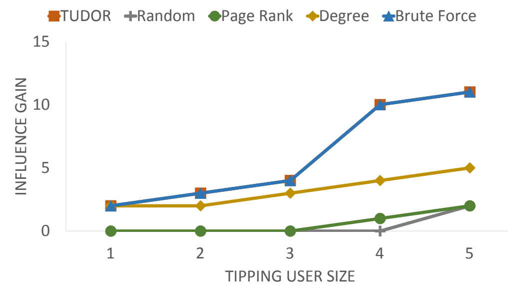
Let $R(u)$ denote the set of all nodes that can be blocked from u via the live-edge path. $\mathcal{Z}^1, \mathcal{Z}^2$ are two candidate sets of tipping user set and $\mathcal{Z}^1 \subseteq \mathcal{Z}^2$. $I(\mathcal{Z}^1 + \{u\}) - I(\mathcal{Z}^1)$ denotes the number of nodes in $R(u)$ which are not in the union set $\cup_{v \in \mathcal{Z}^1} R(v)$. Since $\mathcal{Z}^1 \subseteq \mathcal{Z}^2$, $\cup_{v \in \mathcal{Z}^2} R(v)$ contains more or equal nodes than $\cup_{v \in \mathcal{Z}^1} R(v)$, thus $I(\mathcal{Z}^1 + \{u\}) - I(\mathcal{Z}^1) \geq I(\mathcal{Z}^2 + \{u\}) - I(\mathcal{Z}^2)$, where $I(\mathcal{Z}^2 + \{u\}) - I(\mathcal{Z}^2)$ denotes the number of nodes in $R(u)$ which are not in the union set $\cup_{v \in \mathcal{Z}^2} R(v)$. Therefore the influence gain function is submodular.

B. Proposed Algorithm

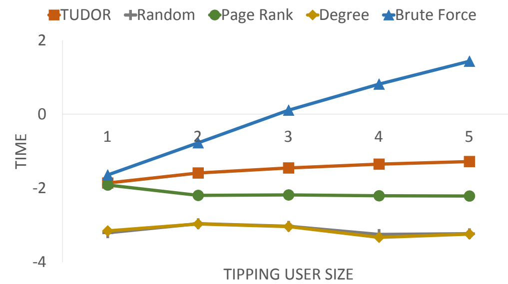
Algorithm 1 Tipping Users Discovery Algorithm

Input: $\mathcal{H} = (G^{(s)}, G^{(t)}, \mathcal{L})$, seed set $\mathcal{S}$, size of tipping users k
Output: tipping user set $\mathcal{Z}$
initialize $\mathcal{Z} = \emptyset$, TIPPING POINTS index $i = 0$;
get the network schema $S^{(s)}$, $S^{(t)}$ and user set $\mathcal{U}^{(s)}, \mathcal{U}^{(t)}$;
extract both intra-network and inter-network meta paths of $G^{(s)}, G^{(t)}$
calculate $\omega_k^{(i)}(u, v)$ and $\omega_k^{(j,i)}(u, v)$ for each user v
learn the weights of meta paths α and β
while $i < k$ **do**
  for $v \in \mathcal{A}^{(s)}$ **do**
    add v into $\mathcal{Z}$
    estimate $\mathcal{Z}$'s inter-network influence gain $I(\mathcal{Z})$
    remove v from $\mathcal{Z}$
  end for
  select $z = \arg\max_v I(\mathcal{Z})$
  $Z = Z \cup \{z\}$ and $i = i + 1$.
end while

For a non-negative, monotone, submodular function $I(\mathcal{Z})$, let $\mathcal{Z}$ be a node set of size k obtained by choosing one node each round, which maximizes the marginal increase in the function value. Let $\mathcal{Z}^*$ be the set with largest value of all k-node sets. Thus $\mathcal{Z}$ provides a $(1 - 1/e)$-approximation, i.e., $I(\mathcal{Z}) \geq (1 - 1/e) \cdot I(\mathcal{Z}^*)$ [16]. Alg. 1 shows the TUDOR algorithm's pseudo-code.



(a) Influence Gain



(b) Running Time

Figure 2: Performance on small dataset when the target network is Foursquare.

VI. Experiment

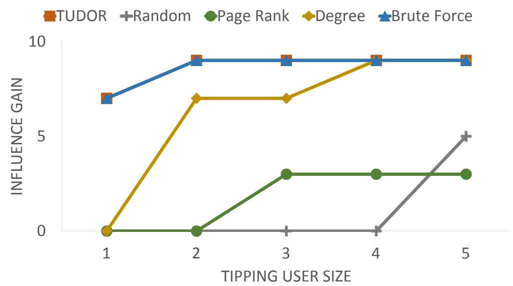
In this section, we present empirical validation of our methods and results from experiments on real network data sets. We first sample a small dataset to compare the performance of TUDOR with the optimal solution. Then the experiments are conducted on original real-world aligned social networks.

Table III: Dataset Description

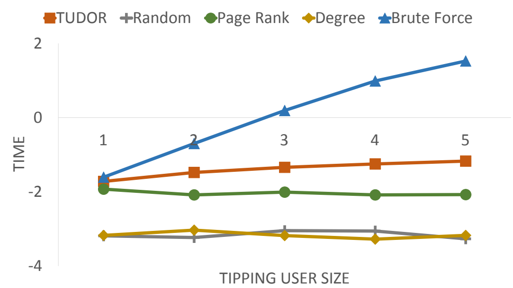
		# dataset	# small dataset
	anchor link	698	22
Twitter	user	1000	30
	follow link	14138	83
	co-location link	13678	48
	contact link	3513	27
Foursquare	user	1000	30
	follow link	13255	73
	co-list link	2806	6
	co-tip link	1024	2

A. Small Dataset Experiment

Dataset Description: We sample a small dataset from Foursquare and Twitter, both of which are famous heterogeneous online social networks. Users of Foursquare can link accounts with their Twitter accounts, which are shown in their profile. Based on these anchor links, we crawled information of users and links in two networks. In the small dataset, we sample total 60 users, 30 of each network, and among them 22 are anchor users. In Twitter, there are 83



(a) Influence Gain



(b) Seed User Analysis

Figure 3: Performance on small dataset when the target network is Twitter.

social links, 48 co-location links and 27 contact links. Users from Foursquare generate 73 social links, 6 co-list links and 2 co-tips links. The statistical information also listed in small dataset column in Table III, and the physical meaning of each link is listed in Table II.

Comparison Methods: We implemented our proposed method and compared with other baselines. Since we are the first to address the TURN problem, there are barely other algorithms can be directly used. Therefore the baselines in this paper are classical algorithms including:

- TUDOR: TUDOR is the method proposed in this paper, which can discover tipping users based on the given seed users.
- Random: This method selects k anchor users randomly.
- Page Rank: Anchor users in $G^{(t)}$ are sorted according to their page rank scores and tipping users are the top k among them.
- Out-Degree (Degree): Rank anchor users in $G^{(t)}$ based on their out-degrees and choose top k as tipping users.
- Brute Force: To achieve the global optimal result, we list all possible combinations of k anchor users and select one which has the largest influence gain.

Setups: The weights of diffusion links are first learned based on Section 3.2 and they are used to calculate user activation probability by logistic function. Since all variables of (4) are in $[0, 1]$ due to the normalization, its result, the activation probability, is in $[0.5, 0.75]$. Since each user has a threshold towards the message in the CONFORM model, value of user's threshold samples from a uniform distribution with range $[0.5, 0.75]$.

There are Foursquare and Twitter two social networks and each time we regard one as the target network and the other as the source network. The size of seed user set is 5 and the number of tipping users changes in range of $\{1, 2, 3, 4, 5\}$. To evaluate the performance of all these methods, we will calculate the influence gain of selected tipping users and we also record the running time of algorithms to compare their efficiency.

Results: When the target network is Foursquare, the experiment results of different comparison methods are given in Fig. 2. Fig. 2(a) shows the influence gain with changing tipping user sizes and Fig. 2(b) shows running time of all algorithms. Due to the huge difference between Brute Force and other methods, the horizontal axis of 2(b) is scaled logarithmically.

Based on the results shown in Fig. 2(a), the influence gains increase as more tipping users are selected for all methods. The performance of TUDOR is exactly the same with Brute Force, which achieves the most influence gain. When the number of tipping users is 5, their influence gains are both 11, which is more than 2 times the result of Degree. Numbers of Page rank and Random algorithm are only 1.

The running time is shown in Fig. 2(b). Random and Degree choose tipping users in a heuristic way and consume the least time, while Page rank needs update users' scores iteratively, so it costs a little more time than the former two. Same as expected, the time cost by Brute Force grows exponentially, when tipping user size is 5, it cost 27.08 seconds. Actually, when the network size extends to 50 users, it is difficult for Brute Force to find out 5 tipping users. TUDOR requires simulating the diffusion processes, which is time-consuming, but TUDOR just use 0.05 second to find the same optimal solution as Brute Force, which demonstrates both the efficiency and effectiveness of TUDOR.

When Twitter as the target network, the outcome of influence gain is available in Fig. 3(a) and Fig. 3(b) shows the time cost, which also uses the logarithmic y-axis.

The results shown in Fig. 3(a) illustrate the influence gains rise with the increase of tipping user size, which is in agreement with the monotone property of influence gain function. Unlike the big difference in Foursquare, all methods except Random achieve the optimal result, because Twitter network is denser than Foursquare, and choosing more users has high probability to get the same influence gain. But among them, only TUDOR can find the optimal solution even when the tipping user size is only 2.

The running time comparison is similar to that of Foursquare. Page rank selects tipping users with simple metric and cost more than Random and Degree methods. The time of Brute Force rises dramatically and reach 33.57 seconds to find 5 tipping users. TUDOR still balance the time and performance well, which cost 0.067 second to discover the optimal 5 tipping users.

The seed users selected by different methods are quite distinct from each other. In Table IV, we show the intersections of tipping user set selected by different methods in two networks, where the tipping user size is 5. We can observe that the tipping user set selected by TUDOR always

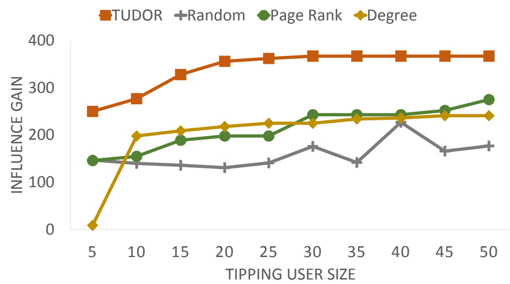
Table IV: Intersection of seed users selected by different comparison methods in the viral marketing problem.

Foursquare

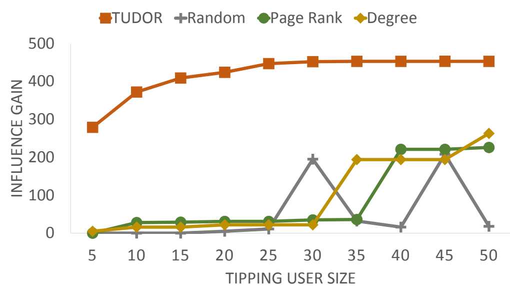
	TUDOR	Random	Page Rank	Degree	Brute Force
TUDOR	5	1	2	4	5
Random		5	0	1	1
Page Rank			5	3	2
Degree				5	4
Brute Force					5

Twitter

	Brute Force	Degree	Page Rank	Random	TUDOR
TUDOR	5	1	0	1	5
Random	0	4	2	5	
Page Rank	1	3	5		
Degree	0	5			
Brute Force	5				



(a) Foursquare



(b) Twitter

Figure 4: Performance of Influence Gain on two target networks respectively.

coincides with the optimal set generated by Brute Force. The interesting thing is that TUDOR shares most tipping users with Degree in Foursquare, but their influence gains differ greatly. While tipping user set of TUDOR is distinct from those selected by Degree in Twitter, however they achieve the same influence gain. This is because the small network size, seed user size and tipping users size magnify the effect of each seed user and tipping use, and make the experiment result sensitive.

B. Original Dataset Experiment

Dataset Description: This dataset is also sampled from Foursquare and Twitter, which the numbers of users in Foursquare and Twitter are both 1000, among which 689 users have accounts in both networks. There exist 4138 social links, 13678 co-location links and 3513 contact links in Twitter network and we crawled 13255 social links, 2806 co-list links and 1024 co-tip links which connect users from Foursquare. Column of large dataset in Table (III) shows the detail statistics information.

Comparison Methods: Obviously, Brute Force method cannot be applied in this dataset due to the large user size. Therefore we compare TUDOR with other baselines of the small dataset besides Brute Force methods.

Setups: Most of experiment setting are the same with that of the small dataset. For example users' threshold values are also set within $[0.5, 0.75]$ and we also treat Foursquare and Twitter as target network respectively. But the seed set contains 50 users from source network and the amount of tipping users changes in range of $[5, 50]$ with step 5. With the large dataset, we just compare their influence gain generated by selected tipping users.

Results: The results of influence gain in Foursquare are shown in Fig. 4(a). Except Random, influence gains increase when more tipping users are selected. However for TUDOR, after a size point, 20 in Fig. 4(a), the growing speed of influence gain slows down, which is because the saturation of tipping users. Comparing with other methods, the result of TUDOR has obvious advantages on others. When the tipping user size is 5, influence gain of TUDOR is 250, however Degree's is only 9. While when the size is enlarged to 50, the result of TUDOR is 367, which is still 53% more than 240 of Degree.

When the experiment conducted on Twitter, outcomes are presented in Fig. 4(b), which shows the advantage of TUDOR clearly. Like in Foursquare, with more tipping users, the influence gain of TUDOR rises, but after the size extending to 20, the growth rate becomes smaller, which is also caused by saturation of tipping users. When comparing with other methods, TUDOR enjoys a big lead with all tipping user size. When the tipping user size is 50, the influence gain of TUDOR is almost 2 times of others' and the superiority is much more obvious with other sizes.

The overlap of tipping users selected by different methods is given in Table (V). From the results, we observe that the tipping users selected by TUDOR is quite distinct from others in both networks. For example, in Foursquare dataset, the intersection between tipping user sets achieved by TUDOR and Page rank is 6 among 50 and so is the case in Twitter.

C. Parameter Analysis

To analyze how parameter the seed user size affects the performance of all methods, we fix the tipping user size and conduct experiment on two datasets. In the small dataset, the tipping user size is 5, and size of seed users changes in the range of $\{1, 2, 3, 4, 5\}$. The results are shown in Fig. 5(a) and Fig. 5(b). In the other dataset, the tipping user size is fixed at 50, and the seed users size is in the range of $[5, 50]$ with step 5. Fig. 6(a) and Fig. 6(b) present the outcomes of influence gain. Here we mainly discuss about the large dataset.

Based on the result shown in Fig. 6(a) and Fig. 6(b), we observe that with more seed users, influence gain shows a uptrend at first, but the value drops afterwards. In Twitter (Fig. 6(b)), for example, the influence gain of TUDOR rises

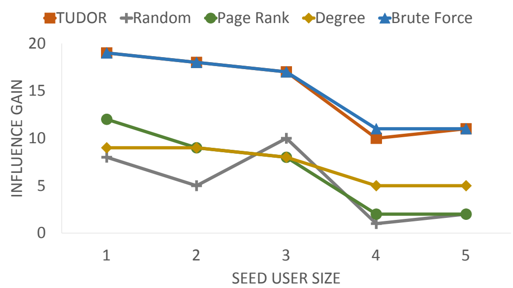
Table V: Intersection of seed users selected by different comparison methods in the viral marketing problem.

Foursquare

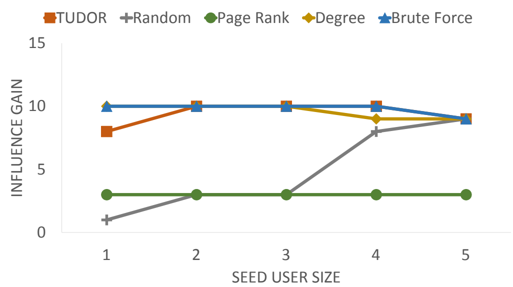
	TUDOR	Random	Page Rank	Degree
TUDOR	50	2	6	3
Random		50	2	2
Page Rank			50	12
Degree				50

Twitter

	Degree	Page Rank	Random	TUDOR
TUDOR	4	5	3	50
Random	6	2	50	
Page Rank	11	50		
Degree	50			

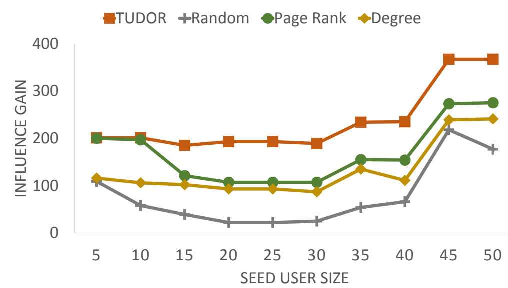


(a) Foursquare

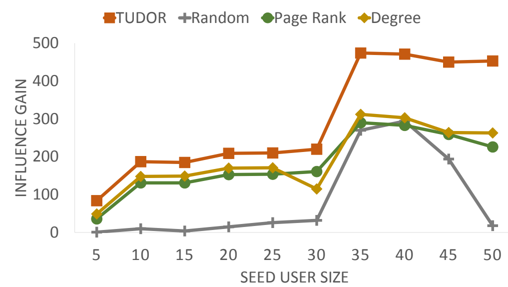


(b) Twitter

Figure 5: Parameter Analysis on the small dataset of two target networks respectively.



(a) Foursquare



(b) Twitter

Figure 6: Parameter Analysis on two target networks respectively.

from 84 to 474 when there are 35 seed users, and decrease to 453 at last. The reason is explained as following: influence gain is the number of activated users who are blocked in the target network when tipping users are removed from the source network. When the size of seed users in the source network increases, there are more original activated users in the target network. Removing a large number of tipping users declines the activated users substantially, and the influence gain rises fast. But when seed user size is large enough, the saturation of seed user causes the number of original activated users keeping the same. However removing tipping users breaks the saturation and let seed users who used to be redundant active others again. Therefore the number of blocked activated users decreases and influence gain drops.

We analyze the results of experiment on small dataset briefly, which are presented in Fig. 5(a) and Fig. 5(b). As we explained before, the number of network size, seed user size and tipping user size are small, which make the result is very sensitive. Therefore the result is different from that of large dataset but the drop of influence gain still can be explained by the same reason.

D. Discussion

According to the result of extensive experiment above, we list some advises of using TUDOR for viral marketing:

1) TUDOR works better when the target network is more intense. The performance in Twitter is better than Foursquare demonstrates this.
2) "More is better" does not apply to tipping user size. The influence gain grows slowly when having enough tipping users.
3) Seed users have great effect. If seed users can active many users, small group of tipping users still can help diffusion, but when seed users cannot do good job, selecting more tipping users is a better choice.

VII. Related Work

Generally, our work is related to the following topics.

Heterogeneous Network Analysis: A heterogeneous information network is composed of multiple types of objects. [13], [18], [25] investigated the principles and methodologies of mining heterogeneous information networks in recent years. Sun et al. studied similarity search that is defined among the same type of objects in heterogeneous networks in [19]. Liu et al. [12] propose a generative graphical model which utilizes the heterogeneous link information to mine topic-level direct influence.

Multiple Social Networks Alignment: Many works study data mining with fusing and utilizing multiple networks [24], [26], [28]. Zhang et al identify connections between the shared users' accounts in multiple social networks in [28] and predict the formation of social links among users in the target network with other external social networks in [27]. [24] studies the influence maximization problem in multiple partially aligned heterogeneous OSNs. However our paper is the first to discover the tipping users in cross network information propagation process based on aligned networks.

Influence Maximization Problem: As an application, influence maximization problem attracts many researchers' interests. Since the seminal paper [9], massive algorithms aim to propose more effective ways to find the seed users for viral marketing ([2], [4]). A huge number of papers extend the original setting to more complicated environment, such as in networks with friendship and foe relationship [11], in continuous time diffusion networks [20] and location-aware networks [10]. However the goal of influence maximization problem is selecting a group of initial users to spread the information, which is totally different from the TURN problem.

Tipping Point: The concept *tipping point* denotes a point in time when a large number of group members rapidly and dramatically change their behavior by widely adopting a previously rare practice in sociology [7]. The phrase has extended beyond its original meaning and been applied to any process in which, beyond a certain point, the rate of the process increases dramatically [22]. It has been applied in many fields, such as economics [15], human ecology [21], and epidemiology [5]. But we are the first to extend *tipping point* into the data mining area and define the new concept *tipping users* based on it.

VIII. CONCLUSION

We study the TURN problem in this paper, which provide a novel way to conduct viral marketing. TURN aims at finding tipping users in the source network to influence users in the target network based on the given seed users. We design the CONFORM model to describe the cross network information diffusion in a pair of heterogeneous networks. The TUDOR method is proposed to address the TURN problem, which can achieve a $(1-1/e)$-approximation of the optimal results. Extensive experiments on real-world social network datasets demonstrate the superior performance of TUDOR in addressing the TURN problem.

ACKNOWLEDGEMENT

This work is supported in part by NSF through grants III-1526499. It is also supported by grants from the Centers for Disease Control and Prevention (CDC), the National Cancer Institute (NCI), and the FDA Center for Tobacco Products (CTP) under award numbers U01CA154254-05S1, U01CA154254, and P50CA179546. The content is solely the responsibility of the authors and does not necessarily represent the official views of CDC, NCI or FDA.

REFERENCES

[1] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *WWW*, 2012.

[2] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier. Maximizing social influence in nearly optimal time. In *SODA*, 2014.

[3] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10(10-17):30, 2010.

[4] W. Chen, Y. Yuan, and L. Zhang. Scalable influence maximization in social networks under the linear threshold model. In *ICDM*, 2010.

[5] K. S. Coyne, C. C. Sexton, C. L. Thompson, I. Milsom, D. Irwin, Z. S. Kopp, C. R. Chapple, S. Kaplan, A. Tubaro, L. P. Aiyer, et al. The prevalence of lower urinary tract symptoms (luts) in the usa, the uk and sweden: results from the epidemiology of luts (epiluts) study. *BJU international*, 104(3):352–360, 2009.

[6] N. Du, L. Song, H. Woo, and H. Zha. Uncover topic-sensitive information diffusion networks. In *AISTATS*, 2013.

[7] M. Gladwell. *The tipping point: How little things can make a big difference*. Little, Brown, 2006.

[8] 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.

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

[10] G. Li, S. Chen, J. Feng, K.-l. Tan, and W.-s. Li. Efficient location-aware influence maximization. In *SIGMOD*, 2014.

[11] Y. Li, W. Chen, Y. Wang, and Z.-L. Zhang. Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In *WSDM*, 2013.

[12] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks. In *CIKM*, 2010.

[13] T. Lou and J. Tang. Mining structural hole spanners through information diffusion in social networks. In *WWW*, 2013.

[14] M. McCombs. *Setting the agenda: The mass media and public opinion*. John Wiley & Sons, 2013.

[15] J. Murray and D. King. Climate policy: Oil's tipping point has passed. *Nature*, 481(7382):433–435, 2012.

[16] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978.

[17] D. M. Scott. *The new rules of marketing and PR: how to use social media, blogs, news releases, online video, and viral marketing to reach buyers directly*. John Wiley & Sons, 2009.

[18] Y. Sun and J. Han. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2):1–159, 2012.

[19] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *VLDB*, 2011.

[20] Y. Tang, Y. Shi, and X. Xiao. Influence maximization in near-linear time: A martingale approach. In *SIGMOD*, 2015.

[21] A. J. Veraart, E. J. Faassen, V. Dakos, E. H. van Nes, M. Lürling, and M. Scheffer. Recovery rates reflect distance to a tipping point in a living system. *Nature*, 481(7381):357–359, 2012.

[22] Wikipedia. Tipping point (sociology), 2004.

[23] J. Wolny and C. Mueller. Analysis of fashion consumers' motives to engage in electronic word-of-mouth communication through social media platforms. *Journal of Marketing Management*, 29(5-6):562–583, 2013.

[24] Q. Zhan, J. Zhang, S. Wang, S. Y. Philip, and J. Xie. Influence maximization across partially aligned heterogenous social networks. In *Advances in Knowledge Discovery and Data Mining*, pages 58–69. Springer, 2015.

[25] J. Zhang and P. Y. Mcd. Mutual clustering across multiple heterogeneous networks. In *IEEE BigData Congress*, 2015.

[26] J. Zhang, W. Shao, S. Wang, X. Kong, and S. Y. Philip. Pna: Partial network alignment with generic stable matching. In *IRI*, 2015.

[27] J. Zhang and P. S. Yu. Integrated anchor and social link predictions across partially aligned social networks. In *IJCAI*, 2015.

[28] J. Zhang and P. S. Yu. Multiple anonymized social networks alignment. In *ICDM*, 2015.

[29] J. Zhang, P. S. Yu, and Z.-H. Zhou. Meta-path based multi-network collective link prediction. In *KDD*, 2014.