Cross-Platform Social Network Analysis

Jiawei Zhang, Philip S. Yu

1 Synonyms

Multiple Aligned Social Network Analysis
Heterogeneous Information Networks
Meta Path based Heterogeneous Social Network Analysis

2 Glossary

SN: Social Network
HIN: Heterogeneous Information Network
MP: Meta Path
INMP: Inter-Network Meta Path

3 Definition

As shown in Figure 1(a), online social networks usually contain heterogeneous information involving different types of nodes, e.g., users, posts, words, timestamps and location checkins, as well as complex links among the nodes, e.g., friendship links among users, write links between users and posts, and the contain/attach links

Jiawei Zhang
Department of Computer Science, University of Illinois at Chicago, IL, USA. e-mail: jzhan9@uic.edu

Philip S. Yu
Department of Computer Science, University of Illinois at Chicago, IL, USA. e-mail: psyu@cs.uic.edu

between posts and words, timestamps and checkins. Formally, such a kind of online social network can be represented as the *heterogeneous information networks*.

Definition 1. (Heterogeneous Information Networks): A *heterogeneous information network* can be represented as $G = (\mathcal{V}, \mathcal{E})$, where the nodes in set $\mathcal{V} = \bigcup_i \mathcal{V}_i$ and the links in set $\mathcal{E} = \bigcup_i \mathcal{E}_i$ are of different categories respectively.

Users nowadays are usually involved in multiple online social networks simultaneously to enjoy more social network services. Formally, the online social networks sharing common users can be defined as the *multiple aligned social networks* [16], which are connected by the *anchor links* [42] between the accounts of shared users, i.e., the *anchor users* [50].

Definition 2. (Multiple Aligned Social Networks): The *multiple aligned social networks* can be represented as $\mathcal{G} = (\{G^i\}_i, \{\mathcal{A}^{(i,j)}\}_{i,j})$, where $G^i = (\mathcal{V}^i, \mathcal{E}^i)$ denotes the i_{th} *heterogeneous information network* and $\mathcal{A}^{(i,j)}$ represents the set of undirected anchor links between networks G^i and G^j.

Definition 3. (Anchor Link): Between networks G^i and G^j, the set of undirected anchor links $\mathcal{A}^{(i,j)}$ can be represented as $\mathcal{A}^{(i,j)} = \{(u_m^i, v_n^j) | u_m^i \in \mathcal{U}^i, v_n^j \in \mathcal{U}^i, u_m^i$ and v_n^j are the accounts of the same user$\}$, where $\mathcal{U}^i \subset \mathcal{V}^i$ and $\mathcal{U}^j \subset \mathcal{V}^j$ are the user node sets in networks G^i and G^j respectively.

One way to model the heterogeneous information available across the *multiple aligned social networks* is *meta path* [34, 50, 47], which abstracts the connections among the different categories of nodes as sequences of *link types* connected by the *node types*. For instance, given the social network with its schema shown in Figure 1, a summary of the intra-network social meta paths extracted from the network is provided in Table 1.

Definition 4. (Intra-Network Meta Path): Given a *heterogeneous information network* $G^i = (\mathcal{V}^i, \mathcal{E}^i)$, we can represents its *networks schema* as $S(G^i) = (\mathcal{T}^i, \mathcal{R}^i)$, where $\mathcal{T}^i$ denotes the types of nodes in $\mathcal{V}^i$ and $\mathcal{R}^i$ denotes the types of links in $\mathcal{E}^i$. Formally, based on the *network schema*, we can define the *meta path* as a sequence $P : T_1^i \xrightarrow{R_1^i} T_2^i \xrightarrow{R_2^i} \cdots \xrightarrow{R_m^i} T_{m+1}^i$, where $T_m^i \in \mathcal{T}^i$ and $R_n^i \in \mathcal{R}^i$ are the node and link types available in network G^i respectively.

Besides the *intra-network meta paths*, via the anchor links and other shared information entities, nodes across different networks can also get connected by the *inter-network meta paths*.

Definition 5. (Inter-Network Meta Path): Given a meta path P consisting of sequences of link types, P is an *inter-network meta path* between networks G^i and G^j iff P involves the node types and link types from the schema of both network G^i and network G^j.

The simplest *inter-network meta path* between networks G^i and G^j will be the *anchor meta path* [44, 50] involving the user node types from G^i and G^j and the anchor link type between G^i and G^j. Some *inter-network meta path* examples are summarized in Table 2.

4 Introduction

Looking from a global perspective, the landscape of online social networks is highly fragmented. A large number of online social networks have appeared and achieved prosperous developments in recent years. Meanwhile, in such an age of online social media, users usually participate in multiple online social networks simultaneously to enjoy more social networks services, who can act as bridges connecting different networks together. Formally, the online social networks sharing common users are named as the *aligned social networks* [16], and these shared users who act like anchors aligning the networks together are called the *anchor users* in existing works [50].

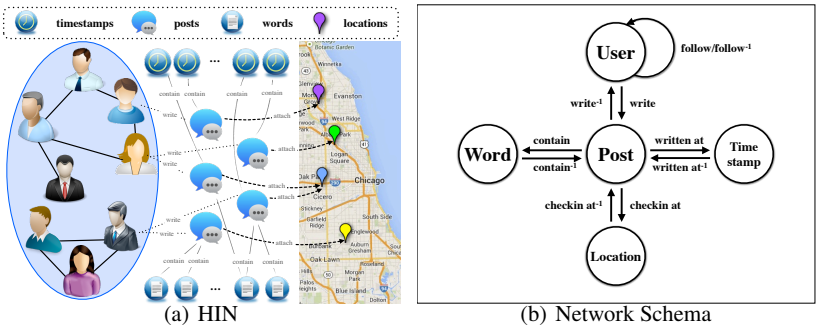
The modeling of *multiple aligned social networks* provides social network practitioners and researchers with the opportunities to study both individual user's social behaviors across multiple social platforms and the propagation of information across multiple social sites. Generally, with the social information from different social sites, we can gain a more comprehensive knowledge about individual's social behavior patterns, which will be helpful for the networks to provide personalized social network services for them. What's more, the social information generated either by the users themselves or from the external offline social events will be able to propagate not only within one single social network, but also across the different social platforms at the same time. By studying the multiple aligned networks simultaneously, we can actually model the information diffusion process much better, which will benefit many social information propagation based applications and services.

However, in the real world, the accounts of individuals in different social sites are mostly isolated without any known correspondence relationships between them. Discovering the correspondence relationships between accounts of the same user can be a crucial step for effective cross-platform social network services and applications, including *friend recommendation*, *social community detection*, *information diffusion and propagation*.

5 Key Points

In this article, we will focus on the cross-platform social network analysis problems, whose prerequisite step is to align the different networks together, i.e., the network alignment step. Meanwhile, to investigate users' social activities and the propagation of information across different social platforms, several application problems will also be introduce in this article after aligning the networks, which include *link prediction*, *community detection*, and *viral marketing*. The formulation of these problems are provided as follows:

- *network alignment*: In the *network alignment* problem, we aim at identifying the common users' accounts (i.e., the anchor links) across different social platforms.



(a) HIN (b) Network Schema

Fig. 1 An example of HIN and the corresponding network schema.

Formally, given networks $G^1, G^2, \cdots, G^n$ together with information available in them, the *network alignment* problem aims at identifying the anchor link sets $\mathscr{A}^{(1,2)}, \mathscr{A}^{(1,3)}, \cdots, \mathscr{A}^{(n-1,n)}$ between pairwise networks.

- *link prediction*: Given multiple aligned networks $\mathscr{G} = (\{G^1, G^2, \cdots, G^n\}, \{\mathscr{A}^{(1,2)}, \mathscr{A}^{(1,3)}, \cdots, \mathscr{A}^{(n-1,n)}\})$, the objective of the *cross-network link prediction* problem is to infer the potential social connections which will be formed in the near future in networks $G^1, G^2, \cdots, G^n$ respectively.
- *community detection*: Given multiple aligned networks $\mathscr{G} = (\{G^1, G^2, \cdots, G^n\}, \{\mathscr{A}^{(1,2)}, \mathscr{A}^{(1,3)}, \cdots, \mathscr{A}^{(n-1,n)}\})$, the *cross-network community detection* problem aims at detecting the community structures of networks $G^1, G^2, \cdots, G^n$ respectively.
- *viral marketing*: Across the multiple aligned networks $\mathscr{G} = (\{G^1, G^2, \cdots, G^n\}, \{\mathscr{A}^{(1,2)}, \mathscr{A}^{(1,3)}, \cdots, \mathscr{A}^{(n-1,n)}\})$, the *cross-network viral marketing* problem aims at modeling the information propagation process across the aligned networks and selecting the optimal seed users who will introduce the maximum influence.

6 Historical Background

Social Network Analysis Cross Aligned Network. Social activity analysis across *aligned social networks* has become a hot research topic in recent years and many pioneer works have been done on this topic. Zhang et al. propose to study the network alignment problem between pairwise fully aligned networks [16], pairwise partially aligned networks [44, 46, 49] and multiple partially aligned networks [48]. Based on the aligned networks, various kinds of application problems have been studied across multiple social platforms, including friend recommendation and social link prediction for new users[42] and emerging networks [43, 50, 46], location recommendation [43], community detection for emerging networks [45] and synergistic clustering across networks [11, 47, 30], information diffusion [40, 41], viral marketing [40], and tipping user identification [41].

Meta Path Applications. Meta path first proposed by Sun et al. for heterogeneous information networks (HIN) in [37] is a powerful tool, which can be applied in link prediction problems [35, 36], clustering problems [37, 34], searching and ranking problems [39, 21] as well as collective classification problem [15] in HIN. However, most of these applications are within one single network only, meta path extracted from which are called the intra-network meta path. In our works, we are the first to extend the meta path concept to inter-network scenario [50, 44] and apply them to address various synergistic knowledge discovery problems across partially aligned heterogeneous social networks, which include network alignment [44], link recommendation [50], community detection [47] and information diffusion [40, 41].

Network Alignment and Stable Matching. Network alignment problem has been well studied in bioinformatics, e.g., protein-protein interaction (PPI) network alignment [13, 32, 33, 18, 14, 22]. Most network alignment approaches focus on finding approximate isomorphism between two graphs [33, 18, 14]. Because of the intractability of the problem, existing methods usually rely on practical heuristics to solve the problem [14, 22]. Meanwhile, in recent years, some works have been done on aligning social networks [16, 17, 26]. Various network alignment models have been proposed to address the problem, which include the supervised classification based network alignment methods [16, 44], PU (positive and unlabeled) classification based method [46], and unsupervised matrix estimation based methods [48, 49].

Link Prediction and Recommendation: Link prediction in social networks first proposed by Liben-Nowell [23] has been a hot research topic and many different methods have been proposed. Liben-Nowell [23] proposes many unsupervised link predicators to predict the social connections among users. Later, Hasan [9] proposes to predict links by using supervised learning methods. An extensive survey of link prediction works is available in [10, 8]. Most existing link prediction works are based on one single network but many researchers start to shift their attention to multiple networks. Dong et al. [6] propose to do link prediction with multiple information sources. Zhang et al. introduce the link prediction problem across aligned networks for new users [42] and emerging networks [43, 46] based on supervised classification models [42] and PU classification models [43, 46] respectively.

Clustering and Community Detection. Clustering is a very broad research area, which includes various types of clustering problems, e.g., consensus clustering [25, 24], multi-view clustering [1, 2], multi-relational clustering [38], co-training based clustering [19], at the same time. Clustering based community detection in online social networks is a hot research topic and many different models have already been proposed to optimizing certain evaluation metrics, e.g., modularity function [29], and normalized cut [31]. A detailed survey about existing community detection works is available in [28, 27]. Meanwhile, based on the information available in multiple aligned networks, Jin [11], Zhang et al. [47] and Shao et al. [30] propose to do synergistic community detection across multiple aligned social networks. Via the anchor links, Zhang et al. also propose to transfer information from developed networks to detect social community structures in emerging networks in [45].

Influence Maximization and Information Diffusion. Influence maximization problem is first proposed by Domingos et al. [5]. It is first formulated as an optimization

Table 1 Summary of Intra-Network Social Meta Paths.

ID	Notation	Intra-Network Social Meta Path	Semantics
1	U $\rightarrow$ U	User $\xrightarrow{follow}$ User	Follow
2	U $\rightarrow$ U $\rightarrow$ U	User $\xrightarrow{follow}$ User $\xrightarrow{follow}$ User	Follower of Follower
3	U $\rightarrow$ U $\leftarrow$ U	User $\xrightarrow{follow}$ User $\xleftarrow{follow}$ User	Common Out Neighbor
4	U $\leftarrow$ U $\rightarrow$ U	User $\xleftarrow{follow}$ User $\xrightarrow{follow}$ User	Common In Neighbor
5	U $\rightarrow$ P $\rightarrow$ W $\leftarrow$ P $\leftarrow$ U	User $\xrightarrow{write}$ Post $\xrightarrow{contain}$ Word $\xleftarrow{contain}$ Post $\xleftarrow{write}$ User	Posts Containing Common Words
6	U $\rightarrow$ P $\rightarrow$ T $\leftarrow$ P $\leftarrow$ U	User $\xrightarrow{write}$ Post $\xrightarrow{contain}$ Time $\xleftarrow{contain}$ Post $\xleftarrow{write}$ User	Posts Containing Common Timestamps
7	U $\rightarrow$ P $\rightarrow$ L $\leftarrow$ P $\leftarrow$ U	User $\xrightarrow{write}$ Post $\xrightarrow{attach}$ Location $\xleftarrow{attach}$ Post $\xleftarrow{write}$ User	Posts Attaching Common Location Check-ins

problem in [12], where Kempe et al. propose two stochastic influence diffusion models, the *independent cascade (IC) model* and *linear threshold (LT) model*, to depict the information propagation process. Viral marketing algorithms are usually of very high time complexiety, and a considerable number of works focusing on speeding up the seed selection have been introduced already, which include the CELF model [20] and the heuristic algorithms for both IC model [4] and LT model [3]. However, most of the existing works mainly focus on information diffusion within one single network but fail to consider the propagation of information across different social platforms. Zhan et al. [40, 41] propose to study the cross-network information diffusion problems to identify both the optimal seed users [40] and tipping users [41] from online social networks respectively.

7 Cross-Network Information Fusion and Mining

In this section, we will briefly introduce several different information fusion problems across multiple social sites. The problem studied in this section include (1) *network alignment*, (2) *social link prediction*, (3) *social community detection*, and (4) *information diffusion and viral marketing*. Before diving into the details about the problems and methods, we will first introduce the meta paths extracted from the aligned heterogeneous social networks at the beginning.

7.1 Social Meta Path Description

Meta paths can actually connect various categories of node types from the network, and those starting and ending with user node types are formally named as the *social meta paths* [47] specifically. In this article, we will use the Foursquare and Twitter networks as the example of *multiple aligned social networks*, which actually share a large amount of common users. As shown in Figure 1(a), both the Foursquare and Twitter networks can be represented as a heterogeneous information network $G = (\mathscr{V}, \mathscr{E})$, where the node set $\mathscr{V} = \mathscr{U} \cup \mathscr{P} \cup \mathscr{L} \cup \mathscr{T} \cup \mathscr{W}$

Table 2 Summary of Inter-Network Social Meta Paths.

ID	Notation	Intra-Network Social Meta Path	Semantics
1	$U^i \to U^i \leftrightarrow U^j \leftarrow U^j$	$\text{User}^i \xrightarrow{follow} \text{User}^i \xleftrightarrow{Anchor} \text{User}^j \xleftarrow{follow} \text{User}^j$	Inter-Network Common Out Neighbor
2	$U^i \leftarrow U^i \leftrightarrow U^j \to U^j$	$\text{User}^i \xleftarrow{follow} \text{User}^i \xleftrightarrow{Anchor} \text{User}^j \xrightarrow{follow} \text{User}^j$	Inter-Network Common In Neighbor
3	$U^i \to U^i \leftrightarrow U^j \to U^j$	$\text{User}^i \xrightarrow{follow} \text{User}^i \xleftrightarrow{Anchor} \text{User}^j \xrightarrow{follow} \text{User}^j$	Inter-Network Common Out In Neighbor
4	$U^i \leftarrow U^i \leftrightarrow U^j \leftarrow U^j$	$\text{User}^i \xleftarrow{follow} \text{User}^i \xleftrightarrow{Anchor} \text{User}^j \xleftarrow{follow} \text{User}^j$	Inter-Network Common In Out Neighbor
5	$U^i \to P^i \to L \leftarrow P^j \leftarrow U^j$	$\text{User}^i \xrightarrow{write} \text{Post}^i \xrightarrow{checkin\ at} \text{Location} \xleftarrow{checkin\ at} \text{Post}^j \xleftarrow{write} \text{User}^j$	Inter-Network Common Location Checkins
7	$U^i \to P^i \to T \leftarrow P^j \leftarrow U^j$	$\text{User}^i \xrightarrow{write} \text{Post}^i \xrightarrow{at} \text{Time} \xleftarrow{at} \text{Post}^j \xleftarrow{write} \text{User}^j$	Inter-Network Common Timestamps
8	$U^i \to P^i \to W \leftarrow P^j \leftarrow U^j$	$\text{User}^i \xrightarrow{write} \text{Post}^i \xrightarrow{contain} \text{Word} \xleftarrow{contain} \text{Post}^j \xleftarrow{write} \text{User}^j$	Inter-Network Common Words

involves the nodes of users, posts, locations, timestamps and words, while the link set $\mathscr{E} = \mathscr{E}_{u,u} \cup \mathscr{E}_{u,p} \cup \mathscr{E}_{p,l} \cup \mathscr{E}_{p,t} \cup \mathscr{E}_{p,w}$ contains the links among users, between users and posts, and those between posts and locations, timestamps, words respectively. The corresponding network schema of the HIN is shown in Figure 1(b). Based on the network schema, a set of *intra-network social meta paths* can be extracted and defined from the network, which are shown in Table 1.

Besides the *intra-network social meta paths*, in Table 2, we also show a list of *inter-network social meta paths* connecting user node types in networks G^i and G^j respectively. These *inter-network social meta paths* connect user nodes across networks via either the *anchor links* or other common information entities, e.g., location checkins, words and timestamps.

7.2 Cross-Network Network Alignment

As introduced in Section 5, let $\mathscr{A}^{(i,j)}$ be the set of anchor links to be inferred between networks G^i and G^j, which maps users between networks G^i and G^j. Considering that users in different social networks are associated with both links and attribute information, the quality of the inferred anchor links $\mathscr{A}^{(i,j)}$ can be measured by the costs introduced by such mappings calculated with users' link and attribute information, i.e.,

$$cost(\mathscr{A}^{(i,j)}) = \text{ cost in links } (\mathscr{A}^{(i,j)}) + \alpha \cdot \text{ cost in attributes}(\mathscr{A}^{(i,j)}),$$

where α denotes the weight of the cost obtained from the attribute information.

7.2.1 Social Structure Information based Network Alignment

Based on the social links among users in both G^i and G^j (i.e., $\mathscr{E}^i_{u,u}$ and $\mathscr{E}^j_{u,u}$ respectively), we can construct the binary *social adjacency matrices* $\mathbf{S}^i \in \mathbb{R}^{|\mathscr{U}^i| \times |\mathscr{U}^i|}$ and $\mathbf{S}^j \in \mathbb{R}^{|\mathscr{U}^j| \times |\mathscr{U}^j|}$ for networks G^i and G^j respectively. Entries in $\mathbf{S}^i$ and $\mathbf{S}^j$ (e.g., $\mathbf{S}^i(p,q)$ and $\mathbf{S}^j(l,m)$) will be assigned with value 1 iff the corresponding social links

(u_p^i, u_q^i) and (u_l^j, u_m^j) exist in G^i and G^j, where $u_p^i, u_q^i \in \mathscr{U}^i$ and $u_l^j, v_m^j \in \mathscr{U}^j$ are users in networks G^i and G^j.

Via the inferred user anchor links $\mathscr{A}^{(i,j)}$, users as well as their social connections can be mapped between networks G^i and G^j. We can represent the inferred user anchor links $\mathscr{A}^{(i,j)}$ with binary *user transitional matrix* $\mathbf{P} \in \mathbb{R}^{|\mathscr{U}^i| \times |\mathscr{U}^j|}$, where the (i_{th}, j_{th}) entry $\mathbf{P}(p,q) = 1$ iff link $(u_p^i, u_q^j) \in \mathscr{A}^{(i,j)}$. Considering that the constraint on user anchor links is *one-to-one*, each column and each row of $\mathbf{P}$ can contain at most one entry being assigned with value 1, i.e.,

$$\mathbf{P}\mathbf{1}^{|\mathscr{U}^j|\times 1} \leq \mathbf{1}^{|\mathscr{U}^i|\times 1},\ \mathbf{P}^\top \mathbf{1}^{|\mathscr{U}^i|\times 1} \leq \mathbf{1}^{|\mathscr{U}^j|\times 1},$$

where $\mathbf{P}\mathbf{1}^{|\mathscr{U}^j|\times 1}$ and $\mathbf{P}^\top\mathbf{1}^{|\mathscr{U}^i|\times 1}$ can get the sum of rows and columns of matrix $\mathbf{P}$ respectively. Equation $\mathbf{P}\mathbf{1}^{|\mathscr{U}^j|\times 1} \leq \mathbf{1}^{|\mathscr{U}^i|\times 1}$ denotes that every entry of the left vector is no greater than the corresponding entry in the right vector.

Matrix $\mathbf{P}$ is an equivalent representation of user anchor link set $\mathscr{A}^{(i,j)}$. Next, we will infer the optimal *user transitional matrix* $\mathbf{P}$, from which we can obtain the optimal anchor link set $\mathscr{A}^{(i,j)}$.

The optimal user anchor links are those which can minimize the inconsistency of mapped social links across networks and the cost introduced by the inferred user anchor link set $\mathscr{A}^{(i,j)}$ with the link information can be represented as

$$\text{cost in link}(\mathscr{A}^{(i,j)}) = \text{cost in link}(\mathbf{P}) = \left\| \mathbf{P}^\top \mathbf{S}^i \mathbf{P} - \mathbf{S}^j \right\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm of the corresponding matrix and $\mathbf{P}^\top$ is the transpose of matrix $\mathbf{P}$.

7.2.2 Social Attribute Information based Network Alignment

With these different attribute information (i.e., username, temporal activity and text content), we can calculate the similarities between users across networks G^i and G^j based on the inter-network social meta paths. To measure the social closeness among users across directed heterogeneous information networks, we propose a new closeness measure named *INMP-Sim* (Inter-Network Meta Path based Similarity) as follows.

Definition 6. (INMP-Sim): Let $\mathscr{P}_i(x \rightsquigarrow y)$ and $\mathscr{P}_i(x \rightsquigarrow \cdot)$ be the sets of path instances of *inter-network meta paths # i* going from x to y and those going from x to other nodes in the network. The INMP-Sim of node pair (x,y) is defined as

$$\text{INMP-Sim}(x,y) = \sum_i \omega_i \left(\frac{|\mathscr{P}_i(x \rightsquigarrow y)| + |\mathscr{P}_i(y \rightsquigarrow x)|}{|\mathscr{P}_i(x \rightsquigarrow \cdot)| + |\mathscr{P}_i(y \rightsquigarrow \cdot)|} \right),$$

where ω_i is the weight of *inter-network meta paths # i* and $\sum_i \omega_i = 1$.

Formally, we represent such similarity matrix as $\Lambda \in \mathbb{R}^{|\mathscr{U}^i|\times|\mathscr{U}^j|}$, where entry $\Lambda(p,q)$ is the similarity between u_p^i and u_q^j. Similar users across social networks are more likely to be the same user and user anchor links $\mathscr{A}_u^{(i,j)}$ that align similar users together should lead to lower cost. In this paper, the cost function introduced by the inferred user anchor links $\mathscr{A}_u^{(i,j)}$ in attribute information is represented as

$$\text{cost in attribute}(\mathscr{A}_u^{(i,j)}) = \text{cost in attribute}(\mathbf{P}) = -\left\|\mathbf{P}\circ\Lambda\right\|_1,$$

where $\|\cdot\|_1$ is the L_1 norm of the corresponding matrix, entry $(\mathbf{P}\circ\Lambda)(i,l)$ can be represented as $P(i,l)\cdot\Lambda(i,l)$ and $\mathbf{P}\circ\Lambda$ denotes the Hadamard product of matrices $\mathbf{P}$ and Λ.

7.2.3 Joint Objective Function for Network Alignment

Both link and attribute information is important for user anchor link inference. By taking these two categories of information into consideration simultaneously, we can represent the optimal *user transitional matrix* $\mathbf{P}^*$ which can lead to the minimum cost as follows:

$$\begin{aligned}
\mathbf{P}^* &= \arg\min_{\mathbf{P}} cost(\mathscr{A}_u^{(i,j)}) \\
&= \arg\min_{\mathbf{P}} \left\|\mathbf{P}^\top\mathbf{S}^i\mathbf{P} - \mathbf{S}^j\right\|_F^2 - \alpha\cdot\left\|\mathbf{P}\circ\Lambda\right\|_1 \\
s.t.\quad & \mathbf{P}\in\{0,1\}^{|\mathscr{U}^i|\times|\mathscr{U}^j|}, \\
& \mathbf{P}\mathbf{1}^{|\mathscr{U}^j|\times 1} \le \mathbf{1}^{|\mathscr{U}^i|\times 1}, \mathbf{P}^\top\mathbf{1}^{|\mathscr{U}^i|\times 1} \le \mathbf{1}^{|\mathscr{U}^j|\times 1}.
\end{aligned}$$

The objective function is an constrained $0-1$ integer programming problem, which is hard to address mathematically. Many relaxation algorithms have been proposed so far. For more information about how to resolve the objective function as well as its effectiveness evaluation on real-world datasets, please refer to [49].

7.3 Cross-Network PU Link Prediction

Given a network screenshot, we propose to label the existing and non-existing social links among users as positive and unlabeled instances respectively, where the unlabeled links involve both positive and negative links at the same time. In this section, we will introduce the PU link prediction framework for multiple aligned networks proposed in [50].

7.3.1 PU Link Prediction Feature Extraction

Meta paths introduced in the previous sections can actually cover a large number of path instances connecting users across the network. Formally, we denote that node n (or link l) is an instance of node type T (or link type R) in the network as $n \in T$ (or $l \in R$). Identity function $I(a,A) = \begin{cases} 1, & \text{if } a \in A \\ 0, & otherwise, \end{cases}$ can check whether node/link a is an instance of node/link type A in the network. To consider the effect of the unconnected links when extracting features for social links in the network, we formally define the *Social Meta Path based Features* to be:

Definition 7. (Social Meta Path based Features): For a given link (u,v), the feature extracted for it based on meta path $P = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$ from the networks is defined to be the expected number of formed path instances between u and v across the networks:

$$x(u,v) = I(u,T_1)I(v,T_k) \sum_{n_1 \in \{u\}, n_2 \in T_2, \cdots, n_k \in \{v\}} \prod_{i=1}^{k-1} p(n_i,n_{i+1}) I((n_i,n_{i+1}),R_i),$$

where $p(n_i,n_{i+1}) = 1.0$ if $(n_i,n_{i+1}) \in E_{u,u}$ and otherwise, $p(n_i,n_{i+1})$ denotes the *formation probability* of link (n_i,n_{i+1}) to be introduced in Subsection 7.3.3.

Based on the above *social meta path based feature* definition and the extracted *intra-network* and *inter-network* meta paths, a set of features can be extracted for user pairs with the information across the aligned networks.

7.3.2 Meta Path based Feature Selection

Meanwhile, information transferred from aligned networks via the features extracted based on the *inter-network social meta path* can be helpful for improving link prediction performance in a given network but can be misleading as well, which is called the *network difference problem*. To solve the *network difference problem*, we propose to rank and select top K features from the feature vector extracted based on the *intra-network* and *inter-network social meta paths*, $\mathbf{x}$, from the multiple *partially aligned heterogeneous networks*.

Let variable $X_i \in \mathbf{x}$ be a feature extracted based on meta paths #i and variable Y be the *label*. $P(Y = y)$ denotes the *prior probability* that links in the training set having label y and $P(X_i = x)$ represents the *frequency* that feature X_i has value x. Information theory related measure *mutual information* (mi) is used as the ranking criteria:

$$mi(X_i) = \sum_x \sum_y P(X_i = x, Y = y) \log \frac{P(X_i = x, Y = y)}{P(X_i = x)P(Y = y)}$$

Let $\bar{\mathbf{x}}$ be the features of the top K *mi* score selected from $\mathbf{x}$. In the next subsection, we will use the selected feature vector $\bar{\mathbf{x}}$ to build a novel PU link prediction model.

7.3.3 PU Link Prediction Method

As introduced at the beginning of this section, from a given network, e.g., G, we can get two disjoint sets of links: connected (i.e., formed) links $\mathscr{P}$ and unconnected links $\mathscr{U}$. To differentiate these links, we define a new concept "*connection state*", z, in this paper to show whether a link is connected (i.e., formed) or unconnected in network G. For a given link l, if l is connected in the network, then $z(l) = +1$; otherwise, $z(l) = -1$. As a result, we can have the "*connection states*" of links in $\mathscr{P}$ and $\mathscr{U}$ to be: $z(\mathscr{P}) = +\mathbf{1}$ and $z(\mathscr{U}) = -\mathbf{1}$.

Besides the "*connection state*", links in the network can also have their own "*labels*", y, which can represent whether a link is to be formed or will never be formed in the network. For a given link l, if l has been formed or to be formed, then $y(l) = +1$; otherwise, $y(l) = -1$. Similarly, we can have the "*labels*" of links in $\mathscr{P}$ and $\mathscr{U}$ to be: $y(\mathscr{P}) = +\mathbf{1}$ but $y(\mathscr{U})$ can be either $+1$ or -1, as $\mathscr{U}$ can contain both links to be formed and links that will never be formed.

By using $\mathscr{P}$ and $\mathscr{U}$ as the positive and negative training sets, we can build a *link connection prediction model* $\mathscr{M}_c$, which can be applied to predict whether a link exists in the original network, i.e., the *connection state* of a link. Let l be a link to be predicted, by applying $\mathscr{M}_c$ to classify l, we can get the *connection probability* of l to be:

Definition 8. (Connection Probability): The probability that link l's *connection states* is predicted to be *connected* (i.e., $z(l) = +1$) is formally defined as the *connection probability* of link l: $p(z(l) = +1|\bar{\mathbf{x}}(l))$.

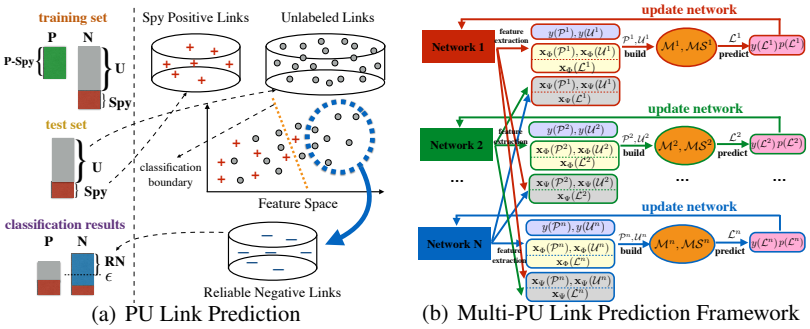
Meanwhile, if we can obtain a set of links that "will never be formed", i.e., "-1" links, from the network, which together with $\mathscr{P}$ ("+1" links) can be used to build a *link formation prediction model*, $\mathscr{M}_f$, which can be used to get the *formation probability* of l to be:

Definition 9. (Formation Probability): The probability that link l's *label* is predicted to be *formed or will be formed* (i.e., $y(l) = +1$) is formally defined as the *formation probability* of link l: $p(y(l) = +1|\bar{\mathbf{x}}(l))$.

However, from the network, we have no information about "links that will never be formed" (i.e., "-1" links). As a result, the *formation probabilities* of potential links that we aim to obtain can be very challenging to calculate. Meanwhile, the correlation between link l's *connection probability* and *formation probability* has been proved in existing works [7] to be:

$$p(y(l) = +1|\bar{\mathbf{x}}(l)) \propto p(z(l) = +1|\bar{\mathbf{x}}(l)).$$

In other words, for links whose *connection probabilities* are low, their *formation probabilities* will be relatively low as well. This rule can be utilized to extract links which can be more likely to be the reliable "-1" links from the network. We propose to apply the the *link connection prediction model* $\mathscr{M}_c$ built with $\mathscr{P}$ and $\mathscr{U}$ to classify links in $\mathscr{U}$ to extract the *reliable negative link set*. Formally, such a kind of



(a) PU Link Prediction (b) Multi-PU Link Prediction Framework

Fig. 2 PU Link Prediction Framework across Multiple Aligned Networks.

negative link extraction method is called the *spy technique based reliable negative link extraction*. For more detailed information about method, please refer to [50].

With the extracted *reliable negative link set* $\mathscr{RN}$, we can solve the *PU link prediction* problem with *classification based link prediction methods*, where $\mathscr{P}$ and $\mathscr{RN}$ are used as the positive and negative training sets respectively. Meanwhile, when applying the built model to predict links in $\mathscr{L}^i$, the optimal labels, $\hat{\mathscr{Y}}^i$, of $\mathscr{L}^i$, should be those which can maximize the following *formation probabilities*:

$$\begin{aligned}\hat{\mathscr{Y}}^i &= \arg\max_{\mathscr{Y}^i} p(y(\mathscr{L}^i) = \mathscr{Y}^i | G^1, G^2, \cdots, G^k) \\ &= \arg\max_{\mathscr{Y}^i} p(y(\mathscr{L}^i) = \mathscr{Y}^i | \bar{\mathbf{x}}(\mathscr{L}^i))\end{aligned}$$

where $y(\mathscr{L}^i) = \mathscr{Y}^i$ represents that links in $\mathscr{L}^i$ have labels $\mathscr{Y}^i$.

7.3.4 Multi-Network Link Prediction Framework

Method proposed in [50] is a general link prediction framework and can be applied to predict social links in *n partially aligned networks* simultaneously. When it comes to n partially aligned network, the optimal labels of potential links $\{\mathscr{L}^1, \mathscr{L}^2, \cdots, \mathscr{L}^n\}$ of networks $G^1, G^2, \cdots, G^n$ will be:

$$\begin{aligned}&\hat{\mathscr{Y}}^1, \hat{\mathscr{Y}}^2, \ldots, \hat{\mathscr{Y}}^n \\ &= \arg \max_{\mathscr{Y}^1, \mathscr{Y}^2, \cdots, \mathscr{Y}^n} p(y(\mathscr{L}^1) = \mathscr{Y}^1, y(\mathscr{L}^2) = \mathscr{Y}^2, \cdots, y(\mathscr{L}^n) = \mathscr{Y}^n | G^1, G^2, \cdots, G^n)\end{aligned}$$

The above target function is very complex to solve and, in this paper, we propose to obtain the solution by updating one variable, e.g., $\mathscr{Y}^1$, and fix other variables, e.g., $\mathscr{Y}^2, \cdots, \mathscr{Y}^n$, alternatively with the following equation [43]:

$$\begin{cases} (\hat{\mathscr{Y}}^1)^{(\tau)} & = \arg\max_{\mathscr{Y}^1} p(y(\mathscr{L}^1) = \mathscr{Y}^1 | G^1, \cdots, G^n, (\hat{\mathscr{Y}}^2)^{(\tau-1)}, (\hat{\mathscr{Y}}^3)^{(\tau-1)}, \cdots, (\hat{\mathscr{Y}}^n)^{(\tau-1)}) \\ (\hat{\mathscr{Y}}^2)^{(\tau)} & = \arg\max_{\mathscr{Y}^2} p(y(\mathscr{L}^2) = \mathscr{Y}^2 | G^1, \cdots, G^n, (\hat{\mathscr{Y}}^1)^{(\tau)}, (\hat{\mathscr{Y}}^3)^{(\tau-1)}, \cdots, (\hat{\mathscr{Y}}^n)^{(\tau-1)}) \\ & \cdots\cdots \\ (\hat{\mathscr{Y}}^n)^{(\tau)} & = \arg\max_{\mathscr{Y}^n} p(y(\mathscr{L}^n) = \mathscr{Y}^n | G^1, \cdots, G^n, (\hat{\mathscr{Y}}^1)^{(\tau)}, (\hat{\mathscr{Y}}^2)^{(\tau)}, \cdots, (\hat{\mathscr{Y}}^{(n-1)})^{(\tau)}) \end{cases}$$

The structure of the link prediction framework is shown in Figure 2(b). When predicting social links in network G^i, we can extract features based on the *intra-network social meta path* extracted from G^i and those extracted based on the *inter-network social meta path* across G^1, G^2, $\cdots$, G^{i-1}, G^{i+1}, $\cdots$, G^n for links in $\mathscr{P}^i$, $\mathscr{U}^i$ and $\mathscr{L}^i$. Feature vectors $\mathbf{x}(\mathscr{P})$ and $\mathbf{x}(\mathscr{P})$ as well as the labels, $y(\mathscr{P})$, $y(\mathscr{U})$, of links in $\mathscr{P}$ and $\mathscr{U}$ are passed to the PU link prediction model $\mathscr{M}^i$ and the meta path selection model $\mathscr{MS}^i$. The formation probabilities of links in $\mathscr{L}^i$ predicted by model $\mathscr{M}^i$ will be used to update the network by replace the weights of $\mathscr{L}^i$ with the newly predicted formation probabilities. The initial weights of these potential links in $\mathscr{L}^i$ are set as 0 (i.e., the *formation probability* of links mentioned in Definition 11). After finishing these steps on G^i, we will move to conduct similar operations on G^{i+1}. We iteratively predict links in G^1 to G^n alternatively in a sequence until the results in all of these networks converge.

7.4 Cross-Network Community Detection

The goal of *cross-network community detection* is to distill relevant information from another social network to compliment knowledge directly derivable from each network to improve the clustering or community detection, while preserving the distinct characteristics of each individual network. To solve the Mutual Clustering problem, a novel community detection method, MCD, is proposed in [47]. By mapping the social network relations into a heterogeneous information, the proposed method in [47] uses the concept of social meta path to define closeness measure among users. Based on this similarity measure, the proposed method [47] can preserve the network characteristics and utilize the information in other networks to refine community structures mutually at the same time. In this section, we will introduce the mutual community detection framework proposed in [47] briefly.

7.4.1 Network Characteristic Preservation Clustering

Clustering each network independently can preserve each networks characteristics effectively as no information from external networks will interfere with the clustering results. Partitioning users of a certain network into several clusters will cut connections in the network and lead to some costs inevitably. Optimal clustering results can be achieved by minimizing the clustering costs.

Let $\mathbf{A}_i$ be the *adjacency matrix* corresponding to the *intra-network meta path* # i among users in the network and $\mathbf{A}_i(m,n) = k$ iff there exist k different path instances

of *intra-network meta path # i* from user m to n in the network. Furthermore, the similarity score matrix among users of meta path # i can be represented as $\mathbf{S}_i = \left(\mathbf{D}_i + \bar{\mathbf{D}}_i\right)^{-1}\left(\mathbf{A}_i + \mathbf{A}_i^T\right)$, where $\mathbf{A}_i^T$ denotes the transpose of $\mathbf{A}_i$, diagonal matrices $\mathbf{D}_i$ and $\bar{\mathbf{D}}_i$ have values $\mathbf{D}_i(l,l) = \sum_m \mathbf{A}_i(l,m)$ and $\bar{\mathbf{D}}_i(l,l) = \sum_m (\mathbf{A}_i^T)(l,m)$ on their diagonals respectively. The meta path based similarity matrix of the network which can capture all possible connections among users is represented as follows:

$$\mathbf{S} = \sum_i \omega_i \mathbf{S}_i = \sum_i \omega_i \left(\left(\mathbf{D}_i + \bar{\mathbf{D}}_i\right)^{-1}\left(\mathbf{A}_i + \mathbf{A}_i^T\right)\right).$$

For a given network G, let $\mathscr{C} = \{U_1, U_2, \ldots, U_k\}$ be the community structures detected from G. Term $\overline{U_i} = \mathscr{U} - U_i$ is defined to be the complement of set U_i in G. Various cost measure of partition $\mathscr{C}$ can be used, e.g., *cut* and *normalized cut*:

$$cut(\mathscr{C}) = \frac{1}{2}\sum_{i=1}^{k} S(U_i, \overline{U_i}) = \frac{1}{2}\sum_{i=1}^{k} \sum_{u \in U_i, v \in \overline{U_i}} S(u,v),$$

$$Ncut(\mathscr{C}) = \frac{1}{2}\sum_{i=1}^{k} \frac{S(U_i, \overline{U_i})}{S(U_i, \cdot)} = \sum_{i=1}^{k} \frac{cut(U_i, \overline{U}_i)}{S(U_i, \cdot)},$$

where $S(u,v)$ denotes the similarity between u, v and $S(U_i, \cdot) = S(U_i, \mathscr{U}) = S(U_i, U_i) + S(U_i, \overline{U}_i)$.

For all users in $\mathscr{U}$, their clustering result can be represented in the *result confidence matrix* $\mathbf{H}$, where $\mathbf{H} = [\mathbf{h_1}, \mathbf{h_2}, \ldots, \mathbf{h_n}]^{\mathbf{T}}$, $n = |\mathscr{U}|$, $\mathbf{h}_i = (h_{i,1}, h_{i,2}, \ldots, h_{i,k})$ and $h_{i,j}$ denotes the confidence that $u_i \in \mathscr{U}$ is in cluster $U_j \in \mathscr{C}$. The optimal $\mathbf{H}$ that can minimize the normalized-cut cost can be obtained by solving the following objective function:

$$\min_{\mathbf{H}} \ \mathrm{Tr}(\mathbf{H}^T \mathbf{L} \mathbf{H}),$$
$$s.t. \ \ \mathbf{H}^T \mathbf{D} \mathbf{H} = \mathbf{I}.$$

where $\mathbf{L} = \mathbf{D} - \mathbf{S}$, diagonal matrix $\mathbf{D}$ has $\mathbf{D}(i,i) = \sum_j \mathbf{S}(i,j)$ on its diagonal, and $\mathbf{I}$ is an identity matrix.

7.4.2 Discrepancy based Clustering of Multiple Aligned Networks

Besides the shared information due to common network construction purposes and similar network features [45], anchor users can also have unique information (e.g., social structures) across aligned networks, which can provide us with a more comprehensive knowledge about the community structures formed by these users. Meanwhile, by maximizing the consensus (i.e., minimizing the "*discrepancy*") of the clustering results about the anchor users in multiple partially aligned networks, we refine the clustering results of the anchor users with information in other aligned

networks mutually. We can represent the clustering results achieved in G^i and G^j as $\mathscr{C}^i = \{U_1^i, U_2^i, \cdots, U_{k^i}^i\}$ and $\mathscr{C}^j = \{U_1^j, U_2^j, \cdots, U_{k^j}^j\}$ respectively.

Let u_p and u_q be two anchor users in the network, whose accounts in G^i and G^j are u_p^i, u_p^j, u_q^i and u_q^j respectively. If users u_p^i and u_q^i are partitioned into the same cluster in G^i but their corresponding accounts u_p^j and u_q^j are partitioned into different clusters in G^j, then it will lead to a *discrepancy* between the clustering results of u_p^i, u_p^j, u_q^i and u_q^j in aligned networks G^i and G^j.

Definition 10. (Discrepancy): The discrepancy between the clustering results of u_p and u_q across aligned networks G^i and G^j is defined as the difference of confidence scores of u_p and u_q being partitioned in the same cluster across aligned networks. Considering that in the clustering results, the confidence scores of u_p^i and u_q^i (u_p^j and u_q^j) being partitioned into k^i (k^j) clusters can be represented as vectors $\mathbf{h}_p^i$ and $\mathbf{h}_q^i$ ($\mathbf{h}_p^j$ and $\mathbf{h}_q^j$) respectively, while the confidences that u_p and u_q are in the same cluster in G^i and G^j can be denoted as $\mathbf{h}_p^i(\mathbf{h}_q^i)^T$ and $\mathbf{h}_p^j(\mathbf{h}_q^j)^T$. Formally, the discrepancy of the clustering results about u_p and u_q is defined to be $d_{p,q}(\mathscr{C}^i, \mathscr{C}^j) = \left(\mathbf{h}_p^i(\mathbf{h}_q^i)^T - \mathbf{h}_p^j(\mathbf{h}_q^j)^T\right)^2$ if u_p, u_q are both anchor users; and $d_{p,q}(\mathscr{C}^i, \mathscr{C}^j) = 0$ otherwise. Furthermore, the discrepancy of $\mathscr{C}^i$ and $\mathscr{C}^j$ will be:

$$d(\mathscr{C}^i, \mathscr{C}^j) = \sum_p^{n^i} \sum_q^{n^j} d_{p,q}(\mathscr{C}^i, \mathscr{C}^j),$$

where $n^i = |\mathscr{U}^i|$ and $n^j = |\mathscr{U}^j|$.

However, considering that $d(\mathscr{C}^i, \mathscr{C}^j)$ is highly dependent on the number of anchor users and anchor links between G^i and G^j, minimizing $d(\mathscr{C}^i, \mathscr{C}^j)$ can favor highly consented clustering results when the anchor users are abundant but have no significant effects when the anchor users are very rare. To solve this problem, we propose to minimize the *normalized discrepancy* instead.

Definition 11. (Normalized Discrepancy) The normalized discrepancy measure computes the differences of clustering results in two aligned networks as a fraction of the discrepancy with regard to the number of anchor users across partially aligned networks:

$$Nd(\mathscr{C}^i, \mathscr{C}^j) = \frac{d(\mathscr{C}^i, \mathscr{C}^j)}{\left(\left|A^{(i,j)}\right|\right)\left(\left|A^{(i,j)}\right| - 1\right)}.$$

Optimal consensus clustering results of G^i and G^j will be $\hat{\mathscr{C}}^i, \hat{\mathscr{C}}^j$:

$$\hat{\mathscr{C}}^i, \hat{\mathscr{C}}^j = \arg\min_{\mathscr{C}^i, \mathscr{C}^j} Nd(\mathscr{C}^i, \mathscr{C}^j).$$

Similarly, the normalized-discrepancy objective function can also be represented with the *clustering results confidence matrices* $\mathbf{H}^i$ and $\mathbf{H}^j$ as well. Meanwhile, considering that the networks studied in this paper are partially aligned, matrices $\mathbf{H}^i$

and $\mathbf{H}^j$ contain the results of both anchor users and non-anchor users, while non-anchor users should not be involved in the discrepancy calculation according to the definition of discrepancy. After pruning the non-anchor users from the confidence matrices, we can represent the pruned confidence matrices as $\bar{\mathbf{H}}^i$ and $\bar{\mathbf{H}}^j$.

Furthermore, the objective function of inferring clustering confidence matrices, which can minimize the normalized discrepancy can be represented as follows

$$\min_{\mathbf{H}^i,\mathbf{H}^j} \frac{\left\|\bar{\mathbf{H}}^i\left(\bar{\mathbf{H}}^i\right)^T-\bar{\mathbf{H}}^j\left(\bar{\mathbf{H}}^j\right)^T\right\|_F^2}{\left\|\mathbf{T}^{(i,j)}\right\|_F^2\left(\left\|\mathbf{T}^{(i,j)}\right\|_F^2-1\right)},$$
$$s.t.\ \ (\mathbf{H}^i)^T\mathbf{D}^i\mathbf{H}^i=\mathbf{I},(\mathbf{H}^j)^T\mathbf{D}^j\mathbf{H}^j=\mathbf{I}.$$

where $\mathbf{D}^i$, $\mathbf{D}^j$ are the corresponding diagonal matrices of similarity matrices of networks G^i and G^j respectively.

7.4.3 Joint Optimization Objective Function

Taking both of these two issues into considerations, the optimal mutual clustering results $\hat{\mathscr{C}}^i$ and $\hat{\mathscr{C}}^j$ of aligned networks G^i and G^j can be achieved as follows:

$$\arg\min_{\mathscr{C}^i,\mathscr{C}^j} \alpha\cdot Ncut(\mathscr{C}^i)+\beta\cdot Ncut(\mathscr{C}^j)+\theta\cdot Nd(\mathscr{C}^i,\mathscr{C}^j)$$

where α, β and θ represents the weights of these terms and, for simplicity, α, β are both set as 1 in this paper.

By replacing $Ncut(\mathscr{C}^i)$, $Ncut(\mathscr{C}^j)$, $Nd(\mathscr{C}^i,\mathscr{C}^j)$ with the objective equations derived above, we can rewrite the joint objective function as follows:

$$\min_{\mathbf{H}^i,\mathbf{H}^j}\ \alpha\cdot\mathrm{Tr}((\mathbf{H}^i)^T\mathbf{L}^i\mathbf{H}^i)+\beta\cdot\mathrm{Tr}((\mathbf{H}^j)^T\mathbf{L}^j\mathbf{H}^j)+\theta\cdot\frac{\left\|\bar{\mathbf{H}}^i\left(\bar{\mathbf{H}}^i\right)^T-\bar{\mathbf{H}}^j\left(\bar{\mathbf{H}}^j\right)^T\right\|_F^2}{\left\|\mathbf{T}^{(i,j)}\right\|_F^2\left(\left\|\mathbf{T}^{(i,j)}\right\|_F^2-1\right)},$$
$$s.t.\ \ (\mathbf{H}^i)^T\mathbf{D}^i\mathbf{H}^i=\mathbf{I},(\mathbf{H}^j)^T\mathbf{D}^j\mathbf{H}^j=\mathbf{I},$$

where $\mathbf{L}^i=\mathbf{D}^i-\mathbf{S}^i$, $\mathbf{L}^j=\mathbf{D}^j-\mathbf{S}^j$ and matrices $\mathbf{S}^i$, $\mathbf{S}^j$ and $\mathbf{D}^i$, $\mathbf{D}^j$ are the similarity matrices and their corresponding diagonal matrices defined before.

The objective function is a complex optimization problem with orthogonality constraints, which can be very difficult to solve because the constraints are not only non-convex but also numerically expensive to preserve during iterations. Please refer to [47] for more information about the solution to the objective function.

7.5 Cross-Network Influence Maximization

Via anchor users, information can propagate not only within but also across social networks. The anchor users' social influence have been seriously underestimated in traditional single-network setting. By identifying seeds that have cross-network impacts, we reduce the number of seeds to affect the same number of people. Alternatively, we can also use an easily accessible network such as Twitter to impact other networks such as Foursquare or Facebook. In this section, we will introduce the *cross-network influence maximization* problem studied in [40], and its objective is to identify the optimal seed users who will introduce the maximum influence across aligned networks.

7.5.1 Information Propagation Model across Aligned Heterogeneous Social Networks

Meanwhile, in heterogeneous social networks, each meta path defines an influence propagation channel among users, based on which, we can construct multi-aligned multi-path networks for the aligned heterogeneous networks. The formal definition of multi-aligned multi-path networks is given as follows:

Definition 12. (Multi-Aligned Multi-Relational Networks (MMNs)) For two given heterogenous networks G^i and G^j, we can define the multi-aligned multi-relational network constructed based on the above intra and inter network social meta paths as $G = (\mathscr{U}, \mathscr{E}, \mathscr{R})$, where $\mathscr{U} = \mathscr{U}^i \cup \mathscr{U}^j$ denote the user nodes in the MMNs G. Set $\mathscr{E}$ is the set of links among nodes in $\mathscr{U}$ and element $e \in \mathscr{E}$ can be represented as $e = (u, v, r)$ denoting that there exists at least one link (u, v) of link type $r \in \mathscr{R} = \mathscr{R}^i \cup \mathscr{R}^j \cup \{Anchor\}$, where $\mathscr{R}^i$, $\mathscr{R}^j$ are the intra-network link types of networks G^i, G^j and the inter-network *Anchor* link between G^i and G^j respectively.

The authors of [40] propose to extend the LT model into the MMNs case and propose a new information diffusion model, MMLT (MMNs based LT model). In particular, under MMNs, they generalize the definition of neighbor to be anyone that can be connected through a given set of meta paths, e.g., anyone in the same network sharing the same posting words under the *intra-network common word meta path*, or across networks under the *inter-network common word meta path*. To simplify the presentation, they assume that the threshold of every object follows a uniform distribution in $[0, 1]$, such that the weighted percentage of the activated neighbors determines the object activation probability, where the weight is determined by the weight of the link. Next, they focus on calculating the object activation probability of all users in the network with the influence propagated based on the MMLT model in multiple meta paths across networks. If the individual's activation probability can exceed his threshold, he will be activated in the MMLT model.

Meanwhile, based on the MMNs $M = (U, E, R)$, the amount of influence propagated between pairs of users in different meta paths in/across the network can be quantified by Pathsim [37]. Formally, the amount of intra-network (inter-network)

influence propagated between user u and v in network G^i with *intra-network meta path* # l and *inter-network meta path* # m can be represented as:

$$\phi^{i,l}_{(u,v)} = \frac{2|\mathscr{P}^{i,l}_{(u,v)}|}{|\mathscr{P}^{i,l}_{(u,\cdot)}| + |\mathscr{P}^{i,l}_{(\cdot,v)}|},\ \psi^{i,m}_{(u,v)} = \frac{2|\mathscr{Q}^{i,m}_{(u,v)}|}{|\mathscr{Q}^{i,m}_{(u,\cdot)}| + |\mathscr{Q}^{i,m}_{(\cdot,v)}|},$$

where $\mathscr{P}^{i,l}_{(u,v)}$ ($\mathscr{Q}^{i,m}_{(u,v)}$) denotes the set of intra-network (inter-network) diffusion channels in meta path # l (and # m) starting from u and ending at v respectively.

Furthermore, in the MMLT model, information diffuses in discrete step and the activation probability of individuals in network G^i at step $t+1$ based on the influence in intra-network (and inter-network) meta path # l (and # m) can be denoted as:

$$g^{i,l}_v(t+1) = \frac{\sum_{u\in\Gamma^{i,l}_{in}(v)} \phi^{i,l}_{(u,v)} I(u,t)}{\sum_{u\in\Gamma^{i,l}_{in}(v)} \phi^{i,l}_{(u,v)}},\ h^{i,m}_{v,j}(t+1) = \frac{\sum_{u\in\Gamma^{i,m}_{in}(v)} \psi^{i,m}_{(u,v)} I(u,t)}{\sum_{u\in\Gamma^{i,m}_{in}(v)} \psi^{i,m}_{(u,v)}},$$

where $\Gamma^{i,l}_{in}(v)$ (and $\Gamma^{i,m}_{in}(v)$) are the neighbor sets of user v in intra-network meta path # l (and inter-network meta path # m) and function $I(u,t) = 1$ if user u is activated at step t, and 0 otherwise.

By aggregating all kinds of intra-network and inter-network relations, they can obtain the integrated activation probability of v^i, where the logistic function is used as the aggregation function.

$$p^i_v(t+1) = \frac{e^{\sum_l \rho^{i,l} g^{i,l}_v(t+1) + \sum_m \omega^{i,m} h^{i,m}_v(t+1)}}{1 + e^{\sum_l \rho^{i,l} g^{i,l}_v(t+1) + \sum_m \omega^{i,m} h^{i,m}_v(t+1)}},$$

where $\rho^{i,l}$ and $\omega^{i,m}$ denote the weights of *intra-network* and *inter-network* relationships in diffusion process, whose value satisfy $\sum_l \rho^{i,l} + \sum_m \omega^{i,m} = 1, \rho^{i,l} \geq 0, \omega^{i,m} \geq 0$. Similarly, we can get activation probability of a user $v^{(j)}$ in $G^{(j)}$.

7.5.2 Seed User Selection

Formally, let mapping $\sigma : \mathscr{Z} \to \mathbb{R}$ denote the influence function which projects the seed user set to the number of users who can get activated by $\mathscr{Z}$. As proposed in [40], based on the cross-network information propagation model introduced in the previous subsection, the identification of the optimal seed user set of certain size who can introduce the maximum influence is NP-hard. Meanwhile, they also show that based on the information diffusion model, the influence function is both *monotone* and *submodular*. In such a case, the conventional stepwise greedy seed user selection method which select the users who can lead to the maximum increase of influence can achieve a $1 - \frac{1}{e}$-approximation of the optimal solution. The pseudo-code of the algorithm is available in Algorithm 1.

Algorithm 1 M&M Greedy Algorithm for AHI problem

Input: $G^{(1)}$, $G^{(2)}$, anchor user matrix $A_{n_{(1)} \times n_{(2)}}$, d
Output: seed set Z
1: initialize $Z =$, seed index $i = 0$;
2: get network schema $S_G^{(1)}$ and $S_G^{(2)}$, get user set $U = U^{(1)} \cup U^{(2)}$;
3: **for** $v = 0$ to $|U|$ **do**
4: extract intra and inter network diffusion meta paths of v;
5: **end for**
6: calculate relations' diffusion strength $\phi_{(u,v)}$ and $\psi_{(u,v)}$;
7: define activation probability vector $P^{(1)}$, $P^{(2)}$ and calculate their initial value;
8: **while** $i < d$ **do**
9: **for** $u \in U \setminus Z$ **do**
10: using Monte Carlo method to estimate u's marginal gain $M_u = \sigma(Z \cup \{u\}) - \sigma(Z)$ based on users' activation probability;
11: **end for**
12: select $z = \arg\max_{u \in U \setminus Z} M_u$
13: $Z = Z \cup \{z\}$
14: update users' activation probability in $P^{(1)}$, $P^{(2)}$ and $i = i + 1$.
15: **end while**

8 Key Applications

The problem introduced in this article are all very important for many concrete real-world social network applications and services. Here, we list the key applications of these introduced works as follows:

- *Application of Network Alignment*: The *network alignment* framework introduced in this article can be applied to various types of existing real-world social networks to identify the common users. In addition, the model can also be applied to align other types of networks, e.g., email contact network, bibliographical co-operation network, message/telephone call network. It can even be used in the traditional entity resolution problem studied in database, and the biological PPI (protein-protein interaction) network alignment as well.
- *Application of Social Link Prediction*: The *link prediction* problem and method introduced in this article can be used to infer potential friendship connections to be formed among users, such that the network service provider can recommend the users to each other as potential friends. Besides recommending friends, it can also be used to recommend locations in location-based social networks, products in e-commerce sites and videos in online video sites, where information from different sources can be aggregated to improve the link prediction result.
- *Application of Community Detection*: With more information available about the entities, the *mutual community detection* framework introduced in this paper can also be applied to automatically categorize the products in e-commerce sites, tag the restaurants in location based sites. Meanwhile, the cross-network community detection problem and the proposed framework also provide another way for researchers to study the traditional multi-view and multi-source clustering problems.
- *Application of Cross-Network Information Diffusion*: By considering the shared anchor users' role in propagating information within and across networks, the

cross-network information diffusion model introduced in this paper can applied in real-world product promotions, election campaigns to propagate the information about products and ideas to activate more people.

9 Future Directions

There are several interesting directions for further research in the domain of *multiple aligned network studies*:

- *Multiple Aligned Social Sites*: Existing aligned network studies mainly focus on studying two aligned networks. Meanwhile, when it comes to *multiple aligned networks* (more than two), many of the studied problems will encounter many new challenges, e.g., the balance of information from different sites, constraints introduced by the multiple sources (e.g., on anchor links).
- *Large Scale Networks*: Most of the introduced methods and models work very well for small-sized social networks, but when it comes to the large scale networks they will suffer from the high time complexity problem a lot. Extending and generalize the existing models to the scalable version will be an interesting direction.
- *Domain Difference Problem*: Many of the existing cross-network studies tackle the domain difference problem in a very simple way, e.g., the meta path selection in *link prediction*, and meta path weighting in *community detection* and *information diffusion*. A more general and effective method to handle the domain difference problem is still an open problem so far.

10 Cross References

- Social Meta Path, Network Schema
- Intra-Network Meta Path, Anchor Meta Path, Inter-Network Meta Path
- Social Structure, Social Adjacency Matrix
- Social Attribute, INMP-Sim
- Positive Links, Unlabeled Links, reliable negative link
- PU Link Prediction
- Social Meta Path based Feature
- Meta Path Selection, Mutual Information
- Connection Probability, Formation Probability
- Multi-Network Link Prediction Framework
- Network Characteristic Preservation Clustering
- Cut, Normalized-Cut
- Discrepancy based Clustering of Multiple Aligned Networks
- Discrepancy, Normalized Discrepancy
- Multi-Aligned Multi-Relational Networks

- MMNs based LT model
- Logistic Function, Aggregation Function
- NP-Hard, Submodular, Monotone
- Greedy Algorithm

11 Acknowledgements

The past research works have been partially supported by NSF through grants III-1526499, IIS-0905215, CNS-1115234, DBI-0960443, and OISE-1129076, US Department of Army through grant W911NF-12-1-0066, Google Research Award, Huawei Grant, Pinnacle Lab at Singapore Management University, NSFC (61333014, 61321491), NSFC(61375069, 61403156) and 111 Program (B14020).

References

1. S. Bickel and T. Scheffer. Multi-view clustering. In *ICDM*, 2004.
2. X. Cai, F. Nie, and H. Huang. Multi-view k-means clustering on big data. In *IJCAI*, 2013.
3. W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD*, 2010.
4. W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *KDD*, 2009.
5. P. Domingos and M. Richardson. Mining the network value of customers. In *KDD*, 2001.
6. Y. Dong, J. Tang, S. Wu, J. Tian, N. Chawla, J. Rao, and H. Cao. Link prediction and recommendation across heterogeneous social networks. In *ICDM*, 2012.
7. C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *KDD*, 2008.
8. L. Getoor and C. P. Diehl. Link mining: A survey. *SIGKDD Explorations Newsletter*, 2005.
9. M. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM*, 2006.
10. M. A. Hasan and M. J. Zaki. A survey of link prediction in social networks. In *Social Network Data Analytics*. Springer, 2011.
11. S. Jin, J. Zhang, P. Yu, S. Yang, and A. Li. Synergistic partitioning in multiple large scale social networks. In *IEEE BigData*, 2014.
12. D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.
13. G. Klau. A new graph-based method for pairwise global network alignment. *BMC Bioinformatics*, 2009.
14. G. Klau. A new graph-based method for pairwise global network alignment. *BMC Bioinformatics*, 2009.
15. X. Kong, P. Yu, Y. Ding, and D. Wild. Meta path-based collective classification in heterogeneous information networks. In *CIKM*, 2012.
16. X. Kong, J. Zhang, and P. Yu. Inferring anchor links across multiple heterogeneous social networks. In *CIKM*, 2013.
17. D. Koutra, H. Tong, and D. Lubensky. Big-align: Fast bipartite graph alignment. In *ICDM'13*, 2013.
18. O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj. Topological network alignment uncovers biological function and phylogeny. *Journal of The Royal Society Interface*, 2010.

19. A. Kumar and H. Daumé. A co-training approach for multi-view spectral clustering. In *ICML*, 2011.
20. J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD*, 2007.
21. Y. Li, C. Shi, P. Yu, and Q. Chen. Hrank: A path based ranking method in heterogeneous information network. In F. Li, G. Li, S. Hwang, B. Yao, and Z. Zhang, editors, *Web-Age Information Management*. 2014.
22. C. Liao, K. Lu, M. Baym, R. Singh, and B. Berger. Isorankn: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 2009.
23. D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM*, 2003.
24. E. F. Lock and D. B. Dunson. Bayesian consensus clustering. *Bioinformatics*, 2013.
25. A. Loureno, S. R. Bul, N. Rebagliati, A. L. N. Fred, M. A. T. Figueiredo, and M. Pelillo. Probabilistic consensus clustering using evidence accumulation. *Machine Learning*, 2013.
26. C. Lu, H. Shuai, and P. Yu. Identifying your customers in social networks. In *CIKM*, 2014.
27. U. Luxburg. A tutorial on spectral clustering. *CoRR*, abs/0711.0189, 2007.
28. F. D. Malliaros and M. Vazirgiannis. Clustering and community detection in directed networks: A survey. *CoRR*, abs/1308.0971, 2013.
29. M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 2004.
30. W. Shao, J. Zhang, L. He, and P. Yu. Multi-source multi-view clustering via discrepancy penalty. In *IJCNN*, 2016.
31. J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 2000.
32. R. Singh, J. Xu, and B. Berger. Pairwise global alignment of protein interaction networks by matching neighborhood topology. In *RECOMB*, 2007.
33. R. Singh, J. Xu, and B. Berger. Pairwise global alignment of protein interaction networks by matching neighborhood topology. In *RECOMB*, 2007.
34. Y. Sun, C. Aggarwal, and J. Han. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. *VLDB*, 2012.
35. Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *ASONAM*, 2011.
36. Y. Sun, J. Han, C. Aggarwal, and N. Chawla. When will it happen?: relationship prediction in heterogeneous information networks. In *WSDM*, 2012.
37. Y. Sun, J. Han, X. Yan, P. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *VLDB*, 2011.
38. X. Yin, J. Han, and P. Yu. Crossclus: user-guided multi-relational clustering. *Data Mining and Knowledge Discovery*, 2007.
39. X. Yu, Y. Sun, B. Norick, T. Mao, and J. Han. User guided entity similarity search using meta-path selection in heterogeneous information networks. In *CIKM*, 2012.
40. Q. Zhan, J. Zhang, S. Wang, P. Yu, and J. Xie. Influence maximization across partially aligned heterogenous social networks. In *PAKDD*. 2015.
41. Q. Zhan, J. Zhang, P. Yu, and J. Xie. Discover tipping users for cross network influencing. In *IRI*, 2016.
42. J. Zhang, X. Kong, and P. Yu. Predicting social links for new users across aligned heterogeneous social networks. In *ICDM*, 2013.
43. J. Zhang, X. Kong, and P. Yu. Transferring heterogeneous links across location-based social networks. In *WSDM*, 2014.
44. J. Zhang, W. Shao, S. Wang, X. Kong, and P. Yu. Pna: Partial network alignment with generic stable matching. In *IRI*, 2015.
45. J. Zhang and P. Yu. Community detection for emerging networks. In *SDM*, 2015.
46. J. Zhang and P. Yu. Integrated anchor and social link predictions across partially aligned social networks. In *IJCAI*, 2015.
47. J. Zhang and P. Yu. Mcd: Mutual clustering across multiple social networks. In *IEEE BigData Congress*, 2015.

48. J. Zhang and P. Yu. Multiple anonymized social networks alignment. In *ICDM*, 2015.
49. J. Zhang and P. Yu. Pct: Partial co-alignment of social networks. In *WWW*, 2016.
50. J. Zhang, P. Yu, and Z. Zhou. Meta-path based multi-network collective link prediction. In *KDD*, 2014.