PNA: Partial Network Alignment with Generic Stable Matching

Jiawei Zhang*, Weixiang Shao†, Senzhang Wang‡, Xiangnan Kong*, Philip S. Yu‡
* University of Illinois at Chicago, Chicago, IL, USA
† Beihang University, Beijing, China
‡ Worcester Polytechnic Institute, Worcester, MA, USA

Abstract—To enjoy more social network services, users nowadays are usually involved in multiple online social networks simultaneously. The shared users between different networks are called anchor users, while the remaining unshared users are named as non-anchor users. Connections between accounts of anchor users in different networks are defined as anchor links and networks partially aligned by anchor links can be represented as partially aligned networks. In this paper, we want to predict anchor links between partially aligned social networks, which is formally defined as the partial network alignment problem. The partial network alignment problem is very difficult to solve because of the following two challenges: (1) the lack of general features for anchor links, and (2) the “one to at most one” constraint on anchor links. To address these two challenges, a new method PNA (Partial Network Aligner) is proposed in this paper. PNA (1) extracts various adjacency scores among users across networks based on a set of inter-network anchor meta paths, and (2) utilizes the generic stable matching to identify the non-anchor users to prune the redundant anchor links attached to them. Extensive experiments conducted on two real-world partially aligned social networks demonstrate that PNA can solve the partial network alignment problem very well and outperform all the other comparison methods with significant advantages.

Index Terms—Partial Network Alignment; Multiple Heterogeneous Social Networks; Data Mining

I. INTRODUCTION

In recent years, online social networks providing various featured services have become an essential part in our lives. To enjoy more social network services, users nowadays are usually involved in multiple online social networks simultaneously [13], [32], [33], [39] and there can be significant overlaps of users shared by different networks. As pointed out in [6], by the end of 2013, 42% of online adults are using multiple social sites at the same time. For example, 93% of Instagram users are involved in Facebook concurrently and 53% Twitter users are using Instagram as well [17]. Formally, the common users involved in different networks simultaneously are named as the "anchor users" [13], while the remaining unshared users are called the "non-anchor users" [39]. The connections between accounts of anchor users in different networks are defined as the "anchor links" [13] and networks partially aligned by anchor links can be represented as "partially aligned networks" [33].

Problem Studied: In this paper, we want to predict the anchor links across partially aligned networks, which is formally defined as the “partial network alignment” problem.

Partial network alignment problem is very important for social networks and can be the prerequisite for many real-world social applications, e.g., link prediction and recommendations [32], [33], [39], [37], community detection [12], [36], [38] and information diffusion [31]. Identifying accounts of anchor users across networks provides the opportunity to compose a more complete social graph with users’ information in all the networks they are involved in. Information in the complete social graph is helpful for a better understanding of users’ social behavior in online social networks [13], [38], [31]. In addition, via the predicted anchor links, cross-platform information exchange enables new social networks to start their services based on the rich data available in other developed networks. The information transferred from developed networks can help emerging networks [33], [36] to overcome the information shortage problem promisingly [32], [33], [36].

What’s more, the partial network alignment problem is a novel problem and different from existing link prediction works, like (1) traditional intra-network link prediction problems [23], [24], which mainly focus on predicting links in one single social network, (2) inter-network link transfer problems [33], which can predict links in one single network with information from multiple aligned networks, and (3) inferring anchor links across fully aligned networks [13], which aims at predicting anchor links across fully aligned networks.

The inferring anchor links across fully aligned networks problem [13] also studies the anchor link prediction problem. However, both the problem setting and method proposed to address the “network alignment” problem between two fully aligned networks in [13] are very ad hoc and have many disadvantages. First of all, the full alignment assumption of social networks proposed in [13] is too strong as fully aligned networks can hardly exist in the real world [39]. Secondly, the features extracted for anchor links in [13] are proposed for Foursquare and Twitter specifically, which can be hard to get generalized to other networks. Thirdly, the classification based link prediction algorithm used in [13] can suffer from the class imbalance problem [15], [18]. The problem will be more serious when dealing with partially aligned networks. Finally, the matching algorithm proposed in [13] is designed specially for fully aligned networks and maps all users (including both anchor and non-anchor users) from one network to another network via the predicted anchor links, which will introduce a large number of non-existing anchor links when applied in the partial network alignment problem.

Totally different from the “inferring anchor links across fully aligned networks” problem [13], we study a more general network alignment problem in this paper. Firstly, networks studied in this paper are partially aligned [39], which contain large number of anchor and non-anchor users [39] at the same time. Secondly, networks studied are not confined to Foursquare and Twitter social networks. A minor revision of the “partial network alignment” problem can be mapped to many other existing tough problems, e.g., large biology network alignment [1], entity resolution in database integration [2], ontology matching [7], and various types of entity matching in online social networks [19]. Thirdly, the class imbalance problem will be addressed via link sampling effectively in the paper. Finally, the constraint on anchor links is “one to at most one” (i.e., each user in one network can be mapped to at most one user in another network).
Across partially aligned networks, only anchor users can be connected by anchor links. Identifying the non-anchor users from networks and pruning all the predicted potential anchor links connected to them is a novel yet challenging problem. The “one-to-one” constraint on anchor links can distinguish the “partial network alignment” problem from most existing link prediction problems. For example, in traditional link prediction and link transfer problems [23], [24], [33], the constraint on links is “many-to-many”, while in the “anchor link inference” problem [13] across fully aligned networks, the constraint on anchor links is strict “one-to-one”.

To solve the “partial network alignment” problem, a new method, PNA (Partial Network Aligner), is proposed in this paper. PNA exploits the concept of anchor meta paths [39], [23] to extract a set of anchor adjacency scores among users across networks. In addition, PNA generalizes the traditional stable matching to support partially aligned network through self-matching and partial stable matching and introduces a novel matching method, generic stable matching, in this paper.

The rest of this paper is organized as follows. In Section II, we will give the definition of some important concepts and formulate the partial network alignment problem. PNA method will be introduces in Sections III-IV. Section V is about the experiments. Related works will be given in Section VI. Finally, we conclude the paper in Section VII.

II. PROBLEM FORMULATION

Before introducing the method PNA, we will first define some important concepts and formulate the partial network alignment problem in this section.

A. Terminology Definition

Definition 1 (Heterogeneous Social Networks): A heterogeneous social network can be represented as $G = (V, E)$, where $V = \bigcup_{i} V_{i}$ contains the sets about various kinds of nodes, while $E = \bigcup_{i} E_{i}$ is the set of different types of links among nodes in $V$.

Definition 2 (Aligned Heterogeneous Social Networks): Social networks that share common users are defined as the aligned heterogeneous social networks, which can be represented as $G = (G_{act}, A_{act})$, where $G_{act} = (G^{(1)}, G^{(2)}, \cdots, G^{(n)})$ is the set of $n$ different heterogeneous social networks and $A_{act} = (A^{(1,2)}, A^{(1,3)}, \cdots, A^{(n-1,n)})$ is the sets of undirected anchor links between networks in $G_{act}$.

In this paper, we will follow the definitions of “anchor links”, “anchor and non-anchor users”, as well as “full alignment, partial alignment” proposed in [39]. Besides anchor users, many other kinds of nodes can be shared between different networks, which are defined as the bridge nodes in this paper.

Definition 3 (Bridge Nodes): The bridge nodes shared between $G^{(i)}$ and $G^{(j)}$ can be represented as $B^{(i,j)} = \{v|(v \in (V^{(i)} \cup V^{(j)}))\wedge(v \in (V^{(i)} \cup V^{(j)}))\}$.

The social networks studied in this paper can be any partially aligned social networks and we use Foursquare, Twitter as an example to illustrate the studied problem and the proposed method. Users in both Foursquare and Twitter can make friends with other users, write posts, which can contain words, timestamps, and location checkins. In addition, users in Foursquare can also create lists of locations that they have visited/want to visit in the future. As a result, Foursquare and Twitter can be represented as heterogeneous social network $G = (V, E)$. In Twitter $V = U \cup P \cup W \cup T \cup L$ and in Foursquare $V = U \cup P \cup W \cup T \cup L$, where $U$, $P$, $W$, $T$, $I$, and $L$ are the nodes of users, posts, words, timestamps, lists, and locations. While in Twitter, the heterogeneous link set $E = E_{u,u} \cup E_{u,p} \cup E_{p,w} \cup E_{p,t} \cup E_{p,l}$ and in Foursquare $E = E_{u,u} \cup E_{u,p} \cup E_{p,w} \cup E_{p,t} \cup E_{p,l} \cup E_{u,i} \cup E_{i,j}$. The bridge nodes shared between Foursquare and Twitter include the common locations, common words and common timestamps.

B. Problem Statement

Definition 4 (Partial Network Alignment): For any two given partially aligned heterogeneous social networks, e.g., $G = ((G^{(i)}, G^{(j)}), (A^{(i,j)}))$, part of the known anchor links between $G^{(i)}$ and $G^{(j)}$ are represented as $A^{(i,j)}$. Let $U^{(i)}$, $U^{(j)}$ be the user sets of $G^{(i)}$ and $G^{(j)}$ respectively, the set of other potential anchor links between $G^{(i)}$ and $G^{(j)}$ can be represented as $L^{(i,j)} = \{(u^{(i)}, v^{(j)})|(u^{(i)} \in U^{(i)}) \wedge (v^{(j)} \in U^{(j)}) \wedge (u^{(i)} \neq v^{(j)})\}$. We solve the partial network alignment problem as a link classification problem, where existing and non-existing anchor links are labeled as “1” and “-1” respectively. In this paper, we aim at building a model $M$ with the existing anchor links $A^{(i,j)}$, which will be applied to predict potential anchor links in $L^{(i,j)}$. In model $M$, we want to determine both labels and existence probabilities of anchor links in $L^{(i,j)}$.

III. FEATURE EXTRACTION AND ANCHOR LINK PREDICTION

Supervised link prediction method has been widely used in research due to its excellent performance and the profound supervised learning theoretical basis. In supervised link prediction, links are labeled differently according to their physical meanings, e.g., existing vs non-existing [39], friends vs enemies [28], trust vs distrust [29], positive attitude vs negative attitude [30]. With information in the networks, a set of heterogeneous features can be extracted for links in the training set, which together with the labels are used to build the link prediction model $M$. In this section, we will introduce different anchor adjacency scores calculated for anchor links across partially aligned networks.

A. Traditional Intra-Network Meta Path

Traditional meta paths are mainly defined based on the social network schema of one single network [23], [25].

Definition 5 (Social Network Schema): For a given network $G$, its schema is defined as $S_{G} = (T_{G}, R_{G})$, where $T_{G}$ and $R_{G}$ are the sets of node types and link types in $G$ respectively.

Definition 6 (Meta Path): Based on the schema of network $G$, i.e., $S_{G} = (T_{G}, R_{G})$, the traditional intra-network meta path in $G$ is defined as $\Phi = T_{1} \rightarrow_{R_{1}} T_{2} \rightarrow_{R_{2}} \cdots \rightarrow_{R_{k-1}} T_{k}$, where $T_{i} \in T_{G}, i \in \{1, 2, \cdots, k\}$ and $R_{j} \in R_{G}, j \in \{1, 2, \cdots, k-1\}$ [23], [25].

For example, according to the networks introduced in Section II, we can define the network schema of Twitter as $S_{G} = \{User, Post, Word, Timestamp, List, Location\}, \{Follow,
Write, Create, Contain, At, Checkin}). Based on the schema, “User - Location - User” is a meta path of length 2 connecting user nodes in the network via location node and path “Alice - San Jose - Bob” is an instance of such meta path in the network, where Alice, Bob and San Jose are the users and location in the network.

B. Inter-Network Anchor Meta Path

Traditional Intra-network meta paths defined based on one single network cannot be applied to address the inter-network partial network alignment problem directly. To overcome such a problem, in this subsection, we will define the concept of anchor meta paths and introduce a set of inter-network anchor meta paths [39] across partially aligned networks.

Definition 7 (Aligned Social Network Schema): Given the partially aligned networks: $G = (G_{set}, A_{set})$, let $S_{G(i)} = (T_{G(i)}, R_{G(i)})$ be the schema of network $G(i) \in G_{set}$, the schema of partially aligned networks $G$ can be defined as $S_G = (\bigcup_i T_{G(i)}, \bigcup_i R_{G(i)}) \cup \{\text{Anchor}\}$, where \{Anchor\} is the anchor link type.

An example of the schema about two partially aligned social networks, e.g., $G(i)$ (e.g., Foursquare) and $G(j)$ (e.g., Twitter), is shown in Figure 1, where the schema of these two aligned networks are connected by the anchor link type and the green dashed circles are the shared bridge nodes between $G(i)$ and $G(j)$.

Definition 8 (Anchor Meta Path): Based on the aligned social network schema, anchor meta paths connecting users across $G$ is defined to be $\Psi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \ldots \xrightarrow{R_{k-1}} T_k$, where $T_1$ and $T_k$ are the “User” node type in two partially aligned social networks respectively. To differentiate the anchor link type from other link types in the anchor meta path, the direction of $R_i$ in $\Psi$ will be bidirectional if $R_i = \text{Anchor}$, $i \in \{1, 2, \ldots, k-1\}$, i.e., $T_i \xleftrightarrow{R_i} T_{i+1}$.

Via the instances of anchor meta paths, users across aligned social networks can be extensively connected to each other. In the two partially aligned social networks (e.g., $G = ((G(i), G(j), (A(i,j)))))$ studied in this paper, various anchor meta paths from $G(i)$ (i.e., Foursquare) and $G(j)$ (i.e., Twitter) can be defined as follows:

- **Common Out Neighbor Anchor Meta Path ($\Psi_1$):** $\text{User}(i) \xrightarrow{\text{follow}} \text{User}(i) \xrightarrow{\text{Anchor}} \text{User}(j) \xrightarrow{\text{follow}} \text{User}(j)$ or “$U(i) \rightarrow T(i) \leftarrow T(j)$” for short.
- **Common In Neighbor Anchor Meta Path ($\Psi_2$):** $\text{User}(i) \xleftarrow{\text{Anchor}} \text{User}(j) \xrightarrow{\text{follow}} \text{User}(j)$ or “$U(i) \leftarrow T(i) \rightarrow T(j)$”.
- **Common Out In Neighbor Anchor Meta Path ($\Psi_3$):** $\text{User}(i) \xrightarrow{\text{follow}} \text{User}(i) \xrightarrow{\text{Anchor}} \text{User}(j) \xrightarrow{\text{follow}} \text{User}(j)$ or “$U(i) \rightarrow U(i) \rightarrow T(j)$”.
- **Common In Out Neighbor Anchor Meta Path ($\Psi_4$):** $\text{User}(i) \xrightarrow{\text{Anchor}} \text{User}(i) \xrightarrow{\text{follow}} \text{User}(j) \xrightarrow{\text{follow}} \text{User}(j)$ or “$U(i) \rightarrow U(i) \rightarrow U(j)$”.

These above anchor meta paths are all defined based the “User” node type only across partially aligned social networks. Furthermore, there can exist many other anchor meta paths consisting user node type and other bridge node types from Foursquare to Twitter, e.g., Location, Word and Timestamp.

C. Anchor Adjacency Scores

Based on the above defined anchor meta paths, different kinds of anchor meta path based adjacency relationship can be extracted from the network. In this paper, we define the new concepts of anchor adjacency score to describe such relationships among users across partially aligned social networks.

Definition 9 (Anchor Meta Path Instance): Based on anchor meta path $\Psi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \ldots \xrightarrow{R_{k-1}} T_k$, path $\psi = n_1 - n_2 - \ldots - n_{k-1} - n_k$ is an instance of $\Psi$ if $n_1$ is an instance of node type $T_1$, $i \in \{1, 2, \ldots, k\}$ and $(n_i, n_{i+1})$ is an instance of link type $R_i$, $\forall i \in \{1, 2, \ldots, k-1\}$.

Definition 10 (Anchor Adjacency Score): The anchor adjacency score is quantified as the number of anchor meta path instances of various anchor meta paths connecting users across networks. The anchor adjacency score between $u(i) \in U(i)$ and $v(j) \in U(j)$ based on meta path $\Psi$ is defined as:

$$\text{score}_\Psi(u(i), v(j)) = \left|\left\{ \psi \in \Psi | \Psi(n_1) = T_1(n_1) \land \Psi(n_k) = T_k(n_k) \right\}\right|,$$

where path $\psi$ starts and ends with node types $T_1$ and $T_k$ respectively and $\psi \in \Psi$ denotes that $\psi$ is a path instance of meta path $\Psi$.

IV. ANCHOR LINK PRUNING WITH GENERIC STABLE MATCHING

In this section, we will introduce the anchor link pruning methods in details, which include (1) class-imbalance anchor link prediction, (2) candidate pre-pruning, (3) traditional stable matching, and (4) the generic stable matching method proposed in this paper, which generalizes the concept of traditional stable matching through both self matching and partial stable matching.

A. Class Imbalance Link Prediction

Based on the anchor adjacency scores calculated according to various anchor meta paths in previous section, various supervised link prediction models [13], [33], [39] can be built to infer the potential anchor links across networks. As proposed in [18], [15], conventional supervised link prediction methods [26], can suffer from the class imbalance problem a lot. To address the problem, two effective methods (down sampling [14] and over sampling [4]) are applied.

Down sampling methods aim at deleting the unreliable negative instances from the training set. In Figure 2, we show the distributions of training links in the feature space, where negative links can be
instances will not harm correct classifications but can lead to extra classification costs, which can be removed with the method proposed in [14].

Another method to overcome the class imbalance problem is to over sample the minority class. Many over sampling methods have been proposed, e.g., over sampling with replacement, over sampling with “synthetic” instances [4]. In this paper we apply over sampling with “synthetic” instances due to its effectiveness and wide usage in many scenarios [4].

For more detailed information about how to down sample/over sample the training instances, please refer to [34].

B. Candidate Pre-Pruning

Across two partially aligned social networks, users in a certain network can have a large number of potential anchor link candidates in the other network, which can lead to great time and space costs in predicting the anchor links. The problem can be even worse when the networks are of large scales, e.g., containing million even billion users, which can make the partial network alignment problem unsolvable. To shrink size of the candidate set, we propose to conduct candidate pre-pruning of links in the test set with users’ profile information (e.g., names and hometown).

As shown in Figure 3, in the given input test set, users are extensively connected with all their potential partners in other networks via anchor links. For each users, we propose to prune their potential candidates with cosine similarity scores calculated based on the usernames and hometowns. If the cosine similarity scores of two users’ username and hometown are both 0, then we will remove the anchor link between them from the test set.

C. Traditional Stable Matching

Meanwhile, as proposed in [13], the one-to-one constraint of anchor links across fully aligned social networks can be met by pruning extra potential anchor link candidates with traditional stable matching. In this subsection, we will introduce the concept of traditional stable matching briefly.

Given the user sets $U^{(1)}$ and $U^{(2)}$ of two partially aligned social networks $G^{(1)}$ and $G^{(2)}$, each user in $U^{(1)}$ (or $U^{(2)}$) has his preference over users in $U^{(2)}$ (or $U^{(1)}$). Term $v_j P_{ij}^u v_k$ is used to denote that $u_i \in U^{(1)}$ prefers $v_j$ to $v_k$, for simplicity, where $v_j, v_k \in U^{(2)}$ and $P_{ij}^u$ is the preference operator of $u_i \in U^{(1)}$. Similarly, we can use term $u_j P_{ji}^v u_k$ to denote that $v_j \in U^{(2)}$ prefers $u_i$ to $u_k$ in $U^{(1)}$ as well.

**Definition 11** (Matching): Mapping $\mu : U^{(1)} \cup U^{(2)} \rightarrow U^{(1)} \cup U^{(2)}$ is defined to be a matching iff $(1) \{\mu(u_i)\} = 1, \forall u_i \in U^{(1)}$ and $\mu(u_i) \in U^{(2)}$; $(2) \{\mu(v_j)\} = 1, \forall v_j \in U^{(2)}$ and $\mu(v_j) \in U^{(1)}$; $(3) \mu(u_i) = v_j$ iff $\mu(v_j) = u_i$.

**Definition 12** (Blocking Pair): A pair $(u_i, v_j)$ is a blocking pair of matching $\mu$ if $u_i$ and $v_j$ prefers each other to their mapped partner, i.e., $(\mu(u_i) \neq v_j) \land (\mu(v_j) \neq u_i)$ and $(v_j P_{ji}^u \mu(u_i)) \land (u_i P_{ij}^v \mu(v_j))$.

**Definition 13** (Stable Matching): Given a matching $\mu$, $\mu$ is stable if there is no blocking pair in the matching results.

As introduced in [13], the stable matching can be obtained with the Gale-Shapley algorithm proposed in [9].

D. Generic Stable Matching

Stable matching based method proposed in [13] can only work well in fully aligned social networks. However, in the real world, few social networks are fully aligned and lots of users in social networks are involved in one network only, i.e., non-anchor users, and they should not be connected by any anchor links. However, traditional stable matching method cannot identify these non-anchor users and remove the predicted potential anchor links connected with them. To overcome such a problem, we will introduce the generic stable matching to identify the non-anchor users and prune the anchor link results to meet the one-to-one constraint.

In PNA, we introduce a novel concept, self matching, which allows users to be mapped to themselves if they are discovered to be non-anchor users. In other words, we will identify the non-anchor users as those who are mapped to themselves in the final matching results.

**Definition 14** (Self Matching): For the given two partially aligned networks $G^{(1)}$ and $G^{(2)}$, user $u_i \in U^{(1)}$, can have his preference $P_{ii}^u$ over users in $U^{(2)} \cup \{u_i\}$ and $u_i$ preferring $u_i$ himself defines that $u_i$ is an non-anchor user and prefers to stay unconnected, which is formally defined as self matching.

Users in one social network will be matched with either partners in other social networks or themselves according to their preference lists (i.e., from high preference scores to low preference scores). Only partners that users prefer over themselves will be accepted finally, otherwise users will be matched with themselves instead.

**Definition 15** (Acceptable Partner): For a given matching $\mu : U^{(1)} \cup U^{(2)} \rightarrow U^{(1)} \cup U^{(2)}$, the mapped partner of $u_i \in U^{(1)}$, i.e., $\mu(u_i)$, is acceptable to $u_i$ iff $\mu(u_i) P_{ii}^u u_i$.

To cut off the partners with very low preference scores, we propose the partial matching strategy to obtain the promising partners, who
will participate in the matching finally.

**Definition 16** (Partial Matching Strategy): The partial matching strategy of user \( u_i \in U^{(1)} \), i.e., \( Q^{(1)}_{u_i} \), consists of the first \( K \) the acceptable partners in \( u_i \)'s preference list \( P_{u_i}^{(1)} \), which are in the same order as those in \( P_{u_i}^{(1)} \), and \( u_i \) in the \( (K + 1)\text{th} \) entry of \( Q^{(1)}_{u_i} \). Parameter \( K \) is called the partial matching rate in this paper.

An example is given in Figure 4, where to get the top 2 promising partners for the user, we place the user himself at the 3rd cell in the preference list. All the remaining potential partners will be cut off and only the top 3 users will participate in the final matching.

Based on the concepts of self matching and partial matching strategy, we define the concepts of partial stable matching and generic stable matching as follow.

**Definition 17** (Partial Stable Matching): For a given matching \( \mu, \nu \) is \( 1 \) rational if \( \mu(u_i) = v_j, \forall u_i \in U^{(1)} \) and \( \nu(v_j) = u_i, \forall v_j \in U^{(2)} \), \( 2 \) pairwise stable if there exist no blocking pairs in the matching results, and \( 3 \) stable if it is both rational and pairwise stable.

**Definition 18** (Generic Stable Matching): For a given matching \( \mu, \nu \) is a generic stable matching iff \( \mu \) is a self matching or \( \mu \) is a partial stable matching.

As example of generic stable matching is shown in the bottom two plots of Figure 3. Traditional stable matching can prune most non-existing anchor links and make sure the results can meet one-to-one constraint. However, it preserves the anchor links (Rebecca, Becky) and (Jonathan, Jon), which are connecting non-anchor users. In generic stable matching with parameter \( K = 1 \), users will be either connected with their most preferred partner or stay unconnected. Users “William” and “Wm” are matched as link (William, Wm) has the highest score. “Rebecca” and “Jonathan” will prefer to stay unconnected as their most preferred partners “Wm” is connected with “William” already. Furthermore, “Becky” and “Joh” will stay unconnected as their most preferred partner “Becca” and “Jonathan” prefer to stay unconnected. In this way, generic stable matching can further prune the non-existing anchor links (Rebecca, Becky) and (Jonathan, Jon).

The truncated generic stable matching results can be achieved with the Generic Gale-Shapley algorithm, whose pseudo-code is available in a more complete version of this paper [34].

V. Experiments

To demonstrate the effectiveness of PNA in predicting anchor links for partially aligned heterogeneous social networks, we conduct extensive experiments on two real-world heterogeneous social networks: Foursquare and Twitter. This section includes three parts: (1) dataset description, (2) experiment settings, and (3) experiment results.

A. Dataset Description

The datasets used in this paper include: Foursquare and Twitter, which were crawled during November 2012 [13], [32], [33], [39]. More detailed information about these two datasets is shown in Table I. The number of anchor links crawled between Foursquare and Twitter is 3,388 and 62.83% Foursquare users are anchor users.

B. Experiment Settings

In this part, we will talk about the experiment settings in details, which includes: (1) comparison methods, (2) evaluation methods, and (3) experiment setups.

1) Comparison Methods: The comparison methods used in the experiments can be divided into the following 4 categories:

Methods with Generic Stable Matching:
- PNAOGM: PNAOGM (PNA with Over sampling & Generic stable Matching) is the method proposed in this paper, which consists of two steps: (1) class imbalance link prediction with over sampling, and (2) candidate pruning with generic stable matching.
- PNAOMG: PNAOMG (PNA with Down sampling & Generic stable Matching) is another method proposed in this paper, which consists of two steps: (1) class imbalance link prediction with down sampling, and (2) candidate pruning with generic stable matching.

Methods with Traditional Stable Matching
- PNAOM: PNAOM (PNA with Over sampling & traditional stable Matching) is identical to PNAOMG except that in the second step, PNAOM applies the traditional stable matching [9], [13].
- PNAOMG: PNAOMG (PNA with Down sampling & traditional stable Matching) is identical to PNAOMG except that in the second step, PNAOMG applies the traditional stable matching [9], [13].

Class Imbalance Anchor Link Prediction:
- PNAO: PNAO (PNA with Over sampling) is the link prediction method with over sampling to overcome the class imbalance problem and has no matching step.
- PNAAD: PNAAD (PNA with Down sampling) is the link prediction method with down sampling to overcome the class imbalance problem and has no matching step.

Existing Network Anchoring Methods
- MNA: MNA (Multi-Network Anchoring) is a two-phase method proposed in [13] which includes: (1) supervised link prediction without addressing class imbalance problem; (2) traditional stable matching [9], [13].
- MNA_no: MNA_no (MNA without one-to-one constraint) is the first step of MNA proposed in [13] which can predict anchor links without addressing the class imbalance problem and has no matching step.

2) Evaluation Metrics: The output of different link prediction methods can be either predicted labels or confidence scores, which are evaluated by Accuracy, AUC, F1 in the experiments.
3) **Experiment Setup**: In the experiment, initially, a fully aligned network containing 3000 users in both Twitter and Foursquare is sampled from the datasets. All the existing anchor links are grouped into the positive link set and all the possible non-existing anchor links are used as the potential link set. Certain number of links are randomly sampled from the potential link set as the negative link set, which is controlled by parameter $\theta$. Parameter $\theta$ represents the ratio of $\#\text{negative}$ to $\#\text{positive}$; $\theta = 1$ denotes the class balance case, i.e., $\#\text{positive} = \#\text{negative} = 0$; $\theta = 0.3$ denotes that 0.3 times as large as that of the positive instance set, i.e., $\#\text{negative} \times 0.3 = \#\text{positive}$. In the experiment, $\theta$ is fixed as 0.3 for all the existing methods. The advantage of PNA is that sampling methods are used as the potential link set. Certain number of links are sampled from the datasets. All the existing methods are applied and the sampling rate is $\frac{\#\text{positive}}{\#\text{negative}} \times 0.3$. The fact that (1) PNA and PNAOMG can perform better than MNA (or PNAO) and PNAOMG and PNAADM can outperform PNAO and PNAADM in the partially aligned network settings, (2) PNAOMG can compare the performance of different link prediction methods; (3) $\eta$ parameter analysis.

1) **Analysis of Sampling Methods**: To examine whether sampling methods can improve the prediction performance on the imbalanced classification problem or not, we also compare PNAO, PNAO with MNA_no and the results are given in Figure 5, where we fix $\theta$ as 0.3 but change $\eta$ with values in $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and compare the AUC achieved by PNAO, PNAO and MNA_no. We can observe that the AUC values of all these three methods increase with the increase of $\eta$ but PNAO and PNAO perform consistently better than MNA_no, especially when $\eta$ is low. In Figures 5(a) and 5(b), we fix $\eta$ as 0.6 but change $\theta$ with values in $\{1, 2, 3, 4, 5, 10, 20, 30, 40, 50\}$ and compare the AUC of PNAO, PNAO and MNA_no. As shown in Figure 5(b), the performance of PNAO, MNA_no and MNA_no can all vary slightly with $\theta$ changing from 1 to 50 and PNAO, PNAO can achieve better performance than MNA_no consistently.

2) **Comparison of Different Link Prediction Methods**: Meanwhile, as generic stable matching based post pruning can only output the labels of potential anchor links in the test set, we also evaluate all these methods by comparing their Accuracy and F1 score. In Table II, we fix $\theta$ as 0.5 but change $\eta$ with values in $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. Table II has two parts. The upper part of Table II shows the Accuracy achieved by all the methods with various $\eta$, and the lower part shows the F1 score. Generally, the performance of all comparison methods rises as $\eta$ increases. In the upper part, methods PNAOMG and PNAADM can consistently perform better than all other comparison methods for different $\eta$. For example, when $\eta = 0.5$, the Accuracy achieved by PNAOMG is higher than PNAO by 3.45%, higher than MNA by 6.0%, higher than PNAO by 7.51% and higher than MNA_no by 7.75%; meanwhile, the Accuracy achieved by PNAADM is higher than PNAO and PNAADM can consistently perform better than all other comparison methods for different $\eta$. For example, when $\eta = 0.5$, the F1 score achieved by PNAOMG and PNAADM over other comparison methods are more obvious under the evaluation of $F1$ as in class imbalance settings, Accuracy is no longer an appropriate evaluation metric [3]. For example, when $\eta = 0.5$, the F1 score achieved by PNAOMG is about 13.25% higher than PNAO, 24% higher than MNA, 101.6% higher than PNAO and 165% higher than MNA_no; so is the case for method PNAADM. The experiment results show that PNAOMG and PNAADM can work well with datasets containing different ratio of anchor links across the networks. Similar results can be obtained from Table III, where we fix $\eta = 0.5$, $K$ as 5 but change $\theta$ with values in $\{1, 2, 3, 4, 5, 10, 20, 30, 40, 50\}$. It shows that PNAOMG and PNAADM can effectively address the class imbalance problem.

The fact that (1) PNAOMG can outperform PNAO (PNAADM outperforms PNAO) shows that generic stable matching can work well in dealing with partially aligned social networks; (2) PNAO can beat PNAO and PNAADM beats PNAO means that stable matching can achieve very good post-pruning results; (3) PNAO and PNAADM can perform better than MNA (or PNAO and PNAO can achieve better results than MNA_no) means that sampling methods can overcome the class imbalance problem very well.

3) **Analysis of Partial Matching Rate**: In the generic stable matching, only top $K$ anchor link candidates will be preserved. In this part, we will analyze the effects of parameter $K$ on the performance of PNAOMG and PNAADM. Figure 6 gives the results (both Accuracy and $F1$) of PNAOMG and PNAADM by setting parameter $K$ with values in $\{1, 2, 3, 4, 5, 10, 20, 30, 40, 50\}$.

In Figures 6(a)-6(b), parameters $\theta$ and $\eta$ are fixed as 5 and 0.4 respectively. From the results, we observe that both PNAOMG and PNAADM can perform very well when $K$ is small and the best is obtained at $K = 1$. It shows that the anchor link candidates with the highest confidence predicted by PNAO and PNAO are the optimal network alignment results when $\theta$ and $\eta$ are low. In Figures 6(c)-6(d), we set $\eta$ as 0.9 and $\theta$ as 50 (i.e., the networks contain more anchor links and the training/test sets become more imbalance), we find that the performance of both PNAOMG and PNAADM increases first and then decreases and finally stay stable as $K$ increases, which shows that the optimal anchor link candidates are those within the top $K$ candidate set rather than the one with the highest confidence as the training/test sets become more imbalance.

In addition, the partial matching strategy can shrink the preference lists of users a lot, which can lead to lower time cost as shown in Figure 7 especially for the smaller $K$ values which lead to better accuracy as shown in Figure 6.

Results in all these figures show that generic stable matching can effectively prune the redundant candidate links and significantly improve the prediction results.

VI. RELATED WORKS

Aligned social network studies have become a hot research topic in recent years. Kong et al. [13] are the first to propose the anchor link prediction problem in fully aligned social networks. Zhang et al. [32], [33], [39], [37] propose to predict links for new users and new networks by transferring heterogeneous information across aligned...
A comprehensive survey about link prediction problems across multiple social networks is available in [35]. In addition to link prediction problems, Jin and Zhang et al. [12], [36], [38] introduce the community detection problems across aligned networks and Zhan et al. [31] study the information diffusion across aligned social networks.

Meta path first proposed by Sun et al. [23] has become a powerful tool, which can be applied in either in link prediction problems [23], [24] or clustering problems [25], [22]. Sun et al. [23] propose to predict co-author relationship in heterogeneous bibliographic networks based on meta path. Sun et al. extend the link prediction model to relationship prediction model based on meta path in [24]. Sun et al. [25] propose to calculate the similarity scores among users based on meta path in bibliographical network. Sun et al. [22] also apply meta path in clustering problem of heterogeneous information networks with incomplete attributes.

Class imbalance problems in classification can be very common in real-world applications. Chawla et al. [4] propose a technique for over-sampling the minority class with generated new synthetic minority instances. Kubat et al. [14] propose to address the class imbalance problems with under sampling of the majority cases in the training set. A systematic study of the class imbalance problem is available in [11].

College admission problem [20] and stable marriage problem [10] have been studied for many years and lots of works have been done in the last century. In recent years, some new papers have come out in these areas. Sotomayor et al. [21] propose to analyze the stability of the equilibrium outcomes in the admission games induced by stable matching rules. Ma [16] analyzes the truncation in stable matching and the small core in nash equilibrium in college admission problems. Floren et al. [8] propose to study the almost stable matching by truncating the Gale-Shapley algorithm.

VII. CONCLUSION

In this paper, we study the partial network alignment problem across partially aligned social networks. To address the challenges of the studied problem, a new method named PNA is proposed in this paper. PNA can extract features for anchor links based on a set of anchor meta paths and overcome the class imbalance problem with over sampling and down sampling. PNA can effectively prune the non-existing anchor links with generic stable matching to ensure the results can meet the one – to – one constraint. Extensive experiments done on two real-world partially aligned networks show the superior performance of PNA in addressing the partial network alignment problem.
This work is supported in part by NSF through grants CNS-1115234, Google Research Award, the Pinnacle Lab at Singapore Management University, and Huawei grants.

REFERENCES


[17] Y. Sun, J. Han, C. Aggarwal, and N. Chawla. When will it happen?: Relationship prediction in heterogeneous information networks. In WSDM, 2012.


[33] Y. Sun, J. Han, C. Aggarwal, and N. Chawla. When will it happen?: Relationship prediction in heterogeneous information networks. In WSDM, 2012.
