Multiple Anonymized Social Networks Alignment

Jiawei Zhang
University of Illinois at Chicago
Chicago, IL, USA
jzhan9@uic.edu

Philip S. Yu
University of Illinois at Chicago
Chicago, IL, USA
Institute for Data Science
Tsinghua University, China
psyu@cs.uic.edu

***Abstract*—Users nowadays are normally involved in multiple (usually more than two) online social networks simultaneously to enjoy more social network services. Some of the networks that users are involved in can share common structures either due to the analogous network construction purposes or because of the similar social network features. However, the social network datasets available in research are usually pre-anonymized and accounts of the shared users in different networks are mostly isolated without any known connections. In this paper, we want to identify such connections between the shared users' accounts in multiple social networks (i.e., the anchor links), which is formally defined as the M-NASA (Multiple Anonymized Social Networks Alignment) problem. M-NASA is very challenging to address due to (1) the lack of known anchor links to build models, (2) the studied networks are anonymized, where no users' personal profile or attribute information is available, and (3) the "*transitivity law*" and the "*one-to-one property*" based constraints on anchor links. To resolve these challenges, a novel two-phase network alignment framework UMA (Unsupervised Multi-network Alignment) is proposed in this paper. Extensive experiments conducted on multiple real-world partially aligned social networks demonstrate that UMA can perform very well in solving the M-NASA problem.**

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

I. Introduction

As proposed in [12], people nowadays are normally involved in multiple (usually *more than two*) social networks simultaneously to enjoy more social network services. Many of these networks can share common structure information (e.g., friendship connections) due to either the analogous network establishing purpose or because of similar network features [33]. Meanwhile, social network data available for research are usually anonymized for privacy concerns [2], where users' personal profile and attribute information (e.g., names, hometown, gender and age) are either removed or replaced with meaningless unique identifiers, and the accounts of the shared users in these anonymized social networks are mostly isolated without any correspondence relationships. In this paper, we want to study the "*Multiple Anonymized Social Networks Alignment*" (M-NASA) problem to identify such correspondence relationships between the shared users' accounts across multiple anonymized social networks.

By following terminology definitions used in existing aligned networks studies [12], [36], social networks sharing common users are defined as "*partially aligned networks*", where the shared users are named as "*anchor users*" and the correspondence relationships between anchor users' accounts in different networks are called *anchor links*. The M-NASA problem studied in this paper aims at identifying the anchor links among multiple anonymized social networks. To help illustrate the M-NASA problem more clearly, we also give an example in Figure 1, which involves 3 different social networks (i.e., networks I, II and III). Users in these 3 networks are all anonymized and their names are replaced with randomly generated identifiers. Each pair of these 3 anonymized networks can actually share some common users, e.g., "David" participates in both networks I and II simultaneously, "Bob" is using networks I and III concurrently, and "Charles" is involved in all these 3 networks at the same time. Besides these shared anchor users, in these 3 partially aligned networks, some users are involved in one single network only (i.e., the non-anchor users [36]), e.g., "Alice" in network I, "Eva" in network II and "Frank" in network III. The M-NASA problem studied in this paper aims at discovering the anchor links (i.e., the dashed bi-directional red lines) connecting anchor users across these 3 social networks.

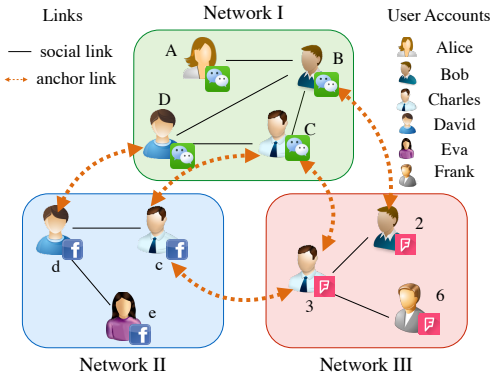


Fig. 1. An example of multiple anonymized partially aligned social networks.

The M-NASA problem is of great importance for online social networks, as it can be the prerequisite for various cross-site social network services, e.g., cross-network link transfer [30], [31], [36], inter-network community detection [9], [33], [35], and viral marketing across networks [29]. With the information transferred from developed social networks, link prediction models proposed in [36] can overcome the *cold-start problem* effectively; constrained by the anchor links, community detection across aligned networks can refine the

community structures of each social network mutually [9], [35]; via the anchor users, information can diffuse not only within but also across networks which will lead to broader impact and activate more users in viral marketing [29].

Besides its importance, the M-NASA problem is a novel problem and totally different from existing works, e.g., (1) *supervised anchor link inference across social networks* [12], which focuses on inferring the anchor links between *two* social networks with a supervised learning model; (2) *network matching* [11], [17], which explores various heuristics to match *two* networks based the known existence probabilities of potential correspondence relationships; (3) *entity resolution* [4], which aims at discovering multiple references to the same entity in *one single database* with a relational clustering algorithm; and (4) *cross-media user identification* [28], which matches users between *two* networks based on various node attribute information generated by users' social activities.

M-NASA differs from all these related works in various aspects: (1) M-NASA is a general multi-network alignment problem and can be applied to align either two [12] or more than two social networks; (2) M-NASA is an *unsupervised* network alignment problem and requires no known anchor links (which are extremely expensive to obtain in the real world); (3) no extra heuristics will be needed and used in the M-NASA problem; and (4) no information about the potential anchor links nor their existence probabilities is required; and (5) social networks studied in M-NASA are anonymized and involve structure information only but no attribute information.

Besides these easily distinguishable distinctions mentioned above, another significant difference of M-NASA from existing *two* network alignment problems is due to the "*transitivity law*" that anchor links follow. In traditional set theory [14], a relation $\mathcal{R}$ is defined to be a *transitive relation* in domain $\mathcal{X}$ iff $\forall a, b, c \in \mathcal{X}, (a, b) \in \mathcal{R} \wedge (b, c) \in \mathcal{R} \rightarrow (a, c) \in \mathcal{R}$. If we treat the union of user account sets of all these social networks as the target domain $\mathcal{X}$ and treat anchor links as the relation $\mathcal{R}$, then anchor links depict a "*transitive relation*" among users across networks. We can take the networks shown in Figure 1 as an example. Let u be a user involved in networks I, II and III simultaneously, whose accounts in these networks are u^I, u^{II} and u^{III} respectively. If anchor links (u^I, u^{II}) and (u^{II}, u^{III}) are identified in aligning networks (I, II) and networks (II, III) respectively (i.e., u^I, u^{II} and u^{III} are discovered to be the same user), then anchor link (u^I, u^{III}) should also exist in the alignment result of networks (I, III) as well. In the M-NASA problem, we need to guarantee the inferred anchor links can meet the *transitivity law*.

In addition to its importance and novelty, the M-NASA problem is very difficult to solve due to the following challenges:

- *unsupervised network alignment*: No existing anchor links are available between pairs of social networks in the M-NASA problem and inferring anchor links between social networks in an unsupervised manner is very challenging.
- *anonymized network alignment*: Networks studied in this paper are all pre-anonymized, where no attribute information indicating users' personal characteristics exists. It makes the M-NASA problem much tougher.
- *transitivity law preservation and utilization*: Anchor links among social networks follow the "transitivity law". How to (1) preserve such a property of anchor links, and (2) utilize such a property to improve the multiple networks partial alignment is still an open problem in this context so far.
- *one-to-one constraint on anchor links*: Anchor links have an inherent *one-to-one* constraint [12], i.e., each user can have at most one account in each social network, which will pose extra challenges on solving the M-NASA problem. (The case that users have multiple accounts in one network can be resolved with method introduced in [25], where these duplicated accounts can be aggregated in advance to form one unique virtual account and the constraint on anchor links connecting these virtual accounts will still be "one-to-one".)

To solve the M-NASA problem, a novel network alignment framework UMA (Unsupervised Multi-network Alignment) is proposed in this paper. UMA addresses the M-NASA problem with two steps: (1) unsupervised transitive anchor link inference across multi-networks, and (2) transitive multi-network matching to maintain the *one-to-one constraint*. In step (1), UMA infers sets of potential anchor links with unsupervised learning techniques by minimizing the *friendship inconsistency* and preserving the *alignment transitivity* property across networks. In step (2), UMA keeps the one-to-one constraint on anchor links by selecting those which can maximize the overall existence probabilities while maintaining the *matching transitivity* property at the same time. The above mentioned new concepts will be introduced in Section III.

The rest of this paper is organized as follows. In Section II, we define some important concepts and the M-NASA problem. Method UMA will be introduced in Section III and evaluated in Section IV. Finally, we introduce the related works in Section V and conclude this paper in Section VI.

II. Problem Formulation

In this section, we will follow the definitions of "*aligned networks*" and "*anchor links*" proposed in [36], which are introduced as follows.

Definition 1 (Anonymized Social Network): An anonymized social network can be represented as graph $G = (\mathcal{U}, \mathcal{E})$, where $\mathcal{U}$ denotes the set of users in the network and $\mathcal{E}$ represents the *social links* among users. Users' profile and attribute information in G has all been deleted to protect individuals' privacy.

Definition 2 (Multiple Aligned Social Networks): Multiple aligned social networks can be represented as $\mathcal{G} = ((G^{(1)}, G^{(2)}, \cdots, G^{(n)}), (\mathcal{A}^{(1,2)}, \mathcal{A}^{(1,3)}, \cdots, \mathcal{A}^{(n-1,n)}))$, where $G^{(i)}, i \in \{1, 2, \cdots, n\}$ represents an anonymized social network and $\mathcal{A}^{(i,j)}, i, j \in \{1, 2, \cdots, n\}$ denotes the set of undirected *anchor links* between networks $G^{(i)}$ and $G^{(j)}$.

Definition 3 (Anchor Links): Given two social networks $G^{(i)}$ and $G^{(j)}$, link $(u^{(i)}, v^{(j)})$ is an *anchor link* between $G^{(i)}$ and $G^{(j)}$ iff $(u^{(i)} \in \mathcal{U}^{(i)}) \wedge (v^{(j)} \in \mathcal{U}^{(j)}) \wedge (u^{(i)}$ and $v^{(j)}$ are accounts of the same user), where $\mathcal{U}^{(i)}$ and $\mathcal{U}^{(j)}$ are the user sets of $G^{(i)}$ and $G^{(j)}$ respectively.

Social networks studied in this paper are all partially aligned [36] and the formal definitions of the concepts like "anchor users", "non-anchor users", "full alignment", "partial alignment" are available in [36].

Based on the above definitions, the M-NASA problem can be formulated as follows:

The M-NASA Problem: Given the n isolated anonymized social networks $\{G^{(1)}, G^{(2)}, \cdots, G^{(n)}\}$, the M-NASA problem aims at discovering the anchor links among these n networks, i.e., the anchor link sets $\mathcal{A}^{(1,2)}, \mathcal{A}^{(1,3)}, \cdots, \mathcal{A}^{(n-1,n)}$. Networks $G^{(1)}, G^{(2)}, \cdots, G^{(n)}$ are partially aligned and the constraint on anchor links in $\mathcal{A}^{(1,2)}, \mathcal{A}^{(1,3)}, \cdots, \mathcal{A}^{(n-1,n)}$ is *one-to-one*, which also follow the *transitivity law*.

III. Proposed Method

Based on observation about the "transitivity property" of anchor links, in this section, we will introduce the UMA method to address the M-NASA problem: in Section 3.1, we formulate the unsupervised pairwise network alignment based on friendship connection information as an optimization problem; integrated multi-network alignment will be introduced in Section 3.2, where an extra constraint called *alignment transitivity* penalty is added to the objective function; the joint optimization function will be solved in Section 3.3 by relaxing its constraints, and the redundant non-existing anchor links introduced by such relaxation will be pruned with *transitive network matching* in Section 3.4.

A. Unsupervised Pairwise Network Alignment

Anchor links between any two given networks $G^{(i)}$ and $G^{(j)}$ actually define an *one-to-one* mapping (of users and social links) between $G^{(i)}$ and $G^{(j)}$. To evaluate the quality of different inferred mapping (i.e., the inferred anchor links), we introduce the concepts of cross-network *Friendship Consistency/Inconsistency* in this paper. The optimal inferred anchor links are those which can maximize the *Friendship Consistency* (or minimize the *Friendship Inconsistency*) across networks.

For any anonymized social network $G = (\mathcal{U}, \mathcal{E})$, the social connections among users in it can be represented with the *social adjacency matrix*.

Definition 4 (Social Adjacency Matrix): Given network $G = (\mathcal{U}, \mathcal{E})$, its *social adjacency matrix* can be represented with binary matrix $\mathbf{S} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{U}|}$ and entry $\mathbf{S}(l,m) = 1$ iff the corresponding social link $(u_l, u_m) \in \mathcal{E}$, where u_l and u_m are users in G.

Based on the above definition, given two partially aligned social networks $G^{(i)} = (\mathcal{U}^{(i)}, \mathcal{E}^{(i)})$ and $G^{(j)} = (\mathcal{U}^{(j)}, \mathcal{E}^{(j)})$, we can represent their corresponding *social adjacency* matrices to be $\mathbf{S}^{(i)} \in \mathbb{R}^{|\mathcal{U}^{(i)}|\times|\mathcal{U}^{(i)}|}$ and $\mathbf{S}^{(j)} \in \mathbb{R}^{|\mathcal{U}^{(j)}|\times|\mathcal{U}^{(j)}|}$ respectively.

Meanwhile, let $\mathcal{A}^{(i,j)}$ be the set of undirected anchor links to be inferred connecting networks $G^{(i)}$ and $G^{(j)}$, based on which, we can construct the corresponding *binary transitional matrix* $\mathbf{T}^{(i,j)}$ between networks $G^{(i)}$ and $G^{(j)}$, where users corresponding to rows and columns of $\mathbf{T}^{(i,j)}$ are of the same order as those of $\mathbf{S}^{(i)}$ and $\mathbf{S}^{(j)}$ respectively.

Definition 5 (Binary Transitional Matrix): Given anchor link set $\mathcal{A}^{(i,j)} \subset \mathcal{U}^{(i)} \times \mathcal{U}^{(j)}$ between networks $G^{(i)}$ and $G^{(j)}$, the *binary transitional matrix* from $G^{(i)}$ to $G^{(j)}$ can be represented as $\mathbf{T}^{(i,j)} \in \{0,1\}^{|\mathcal{U}^{(i)}|\times|\mathcal{U}^{(j)}|}$, where $\mathbf{T}^{(i,j)}(l,m) = 1$ iff link $(u_l^{(i)}, u_m^{(j)}) \in \mathcal{A}^{(i,j)}$, $u_l^{(i)} \in \mathcal{U}^{(i)}$, $u_m^{(j)} \in \mathcal{U}^{(j)}$.

The *binary transitional matrix* from $G^{(j)}$ to $G^{(i)}$ can be defined in a similar way, which can be represented as $\mathbf{T}^{(j,i)} \in \{0,1\}^{|\mathcal{U}^{(j)}|\times|\mathcal{U}^{(i)}|}$, where $(\mathbf{T}^{(i,j)})^\top = \mathbf{T}^{(j,i)}$ as the anchor links between $G^{(i)}$ and $G^{(j)}$ are undirected. Considering that anchor links have an inherent *one-to-one* constraint, each row and each column of the *binary transitional matrices* $\mathbf{T}^{(i,j)}$ and $\mathbf{T}^{(j,i)}$ should have at most one entry filled with 1, which will constrain the inference space of potential *binary transitional matrices* $\mathbf{T}^{(i,j)}$ and $\mathbf{T}^{(j,i)}$ greatly.

Binary transitional matrix $\mathbf{T}^{(i,j)}$ defines a mapping of users from network $G^{(i)}$ to $G^{(j)}$, i.e., $\mathbf{T}^{(i,j)} : \mathcal{U}^{(i)} \to \mathcal{U}^{(j)}$. Besides the user nodes, the social links in network $G^{(i)}$ can also be projected to network $G^{(j)}$ via the binary transitional matrices $\mathbf{T}^{(i,j)}$ and $\mathbf{T}^{(j,i)}$: the *social adjacency matrix* $\mathbf{S}^{(i)}$ being mapped from $G^{(i)}$ to $G^{(j)}$ can be represented as $\mathbf{T}^{(j,i)}\mathbf{S}^{(i)}\mathbf{T}^{(i,j)}$ (i.e., $(\mathbf{T}^{(i,j)})^\top\mathbf{S}^{(i)}\mathbf{T}^{(i,j)}$). Furthermore, considering social networks $G^{(i)}$ and $G^{(j)}$ share significant community structure overlaps, the friendship connections mapped from $G^{(i)}$ to $G^{(j)}$ (i.e., $(\mathbf{T}^{(i,j)})^\top\mathbf{S}^{(i)}\mathbf{T}^{(i,j)}$) should be consistent with those in $G^{(j)}$ (i.e., $\mathbf{S}^{(j)}$), which can be quantified as the following cross-network *friendship consistency* formally [13].

Definition 6 (Friendship Consistency/Inconsistency): The *friendship consistency* between network $G^{(i)}$ and $G^{(j)}$ introduced by the cross-network mapping $\mathbf{T}^{(i,j)}$ is defined as number of shared social links between those mapped from $G^{(i)}$ and the social links in $G^{(j)}$ originally.

Meanwhile, we can define the *friendship inconsistency* as the number of non-shared social links between those mapped from $G^{(i)}$ and those in $G^{(j)}$. Based on the inferred *anchor transitional matrix* $\mathbf{T}^{(i,j)}$, the introduced *friendship inconsistency* between matrices $(\mathbf{T}^{(i,j)})^\top\mathbf{S}^{(i)}\mathbf{T}^{(i,j)}$ and $\mathbf{S}^{(j)}$ can be represented as:

$$\left\|(\mathbf{T}^{(i,j)})^\top\mathbf{S}^{(i)}\mathbf{T}^{(i,j)} - \mathbf{S}^{(j)}\right\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. And the optimal *binary transitional matrix* $\bar{\mathbf{T}}^{(i,j)}$, which can lead to the minimum *friendship inconsistency* can be represented as

$$\begin{aligned}
\bar{\mathbf{T}}^{(i,j)} = \arg\min_{\mathbf{T}^{(i,j)}} & \left\|(\mathbf{T}^{(i,j)})^\top\mathbf{S}^{(i)}\mathbf{T}^{(i,j)} - \mathbf{S}^{(j)}\right\|_F^2 \\
s.t. \quad & \mathbf{T}^{(i,j)} \in \{0,1\}^{|\mathcal{U}^{(i)}|\times|\mathcal{U}^{(j)}|}, \\
& \mathbf{T}^{(i,j)}\mathbf{1}^{|\mathcal{U}^{(j)}|\times 1} \preccurlyeq \mathbf{1}^{|\mathcal{U}^{(i)}|\times 1}, \\
& (\mathbf{T}^{(i,j)})^\top\mathbf{1}^{|\mathcal{U}^{(i)}|\times 1} \preccurlyeq \mathbf{1}^{|\mathcal{U}^{(j)}|\times 1},
\end{aligned}$$

where the last two equations are added to maintain the *one-to-one* constraint on anchor links and $\mathbf{X} \preccurlyeq \mathbf{Y}$ iff $\mathbf{X}$ is of the same dimensions as $\mathbf{Y}$ and every entry in $\mathbf{X}$ is no greater than the corresponding entry in $\mathbf{Y}$.

B. Transitive Integrate Network Alignment

Isolated network alignment can work well in addressing the alignment problem of two social networks. However, in the M-NASA problem studied in this paper, multiple social networks (more than two) social networks are to be aligned simultaneously. Besides minimizing the *friendship inconsistency* between each pair of networks, the *transitivity* property of anchor links also needs to be preserved in the transitional matrices inference.

The *transitivity* property should holds for the alignment of any n networks, where the minimum of n is 3. To help illustrate the *transitivity property* more clearly, we will use 3 network alignment as an example to introduce the M-NASA problem, which can be easily generalized to the case of n networks alignment. Let $G^{(i)}$, $G^{(j)}$ and $G^{(k)}$ be 3 social networks to be aligned concurrently. To accommodate the alignment results and preserve the *transitivity* property, we introduce the following *alignment transitivity penalty*:

Definition 7 (Alignment Transitivity Penalty): Let $\mathbf{T}^{(i,j)}$, $\mathbf{T}^{(j,k)}$ and $\mathbf{T}^{(i,k)}$ be the inferred binary transitional matrices from $G^{(i)}$ to $G^{(j)}$, from $G^{(j)}$ to $G^{(k)}$ and from $G^{(i)}$ to $G^{(k)}$ respectively among these 3 networks. The *alignment transitivity penalty* $C(\{G^{(i)}, G^{(j)}, G^{(k)}\})$ introduced by the inferred transitional matrices can be quantified as the number of inconsistent social links being mapped from $G^{(i)}$ to $G^{(k)}$ via two different alignment paths $G^{(i)} \to G^{(j)} \to G^{(k)}$ and $G^{(i)} \to G^{(k)}$, i.e.,

$$C(\{G^{(i)}, G^{(j)}, G^{(k)}\}) = \left\| (\mathbf{T}^{(j,k)})^\top (\mathbf{T}^{(i,j)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} - (\mathbf{T}^{(i,k)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,k)} \right\|_F^2.$$

Alignment transitivity penalty is a general penalty concept and can be applied to n networks $\{G^{(1)}, G^{(2)}, \cdots, G^{(n)}\}$, $n \geq 3$ as well, which can be defined as the summation of penalty introduced by any three networks in the set, i.e.,

$$C(\{G^{(1)}, G^{(2)}, \cdots, G^{(n)}\}) = \sum_{\forall \{G^{(i)}, G^{(j)}, G^{(k)}\} \subset \{G^{(1)}, G^{(2)}, \cdots, G^{(n)}\}} C(\{G^{(i)}, G^{(j)}, G^{(k)}\}).$$

The optimal *binary transitional matrices* $\bar{\mathbf{T}}^{(i,j)}$, $\bar{\mathbf{T}}^{(j,k)}$ and $\bar{\mathbf{T}}^{(k,i)}$ which can minimize friendship inconsistency and the *alignment transitivity penalty* at the same time can be represented to be

$$\begin{aligned}
&\bar{\mathbf{T}}^{(i,j)}, \bar{\mathbf{T}}^{(j,k)}, \bar{\mathbf{T}}^{(k,i)} \\
&= \arg\min_{\mathbf{T}^{(i,j)}, \mathbf{T}^{(j,k)}, \mathbf{T}^{(k,i)}} \left\| (\mathbf{T}^{(i,j)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} - \mathbf{S}^{(j)} \right\|_F^2 \\
&+ \left\| (\mathbf{T}^{(j,k)})^\top \mathbf{S}^{(j)} \mathbf{T}^{(j,k)} - \mathbf{S}^{(k)} \right\|_F^2 + \left\| (\mathbf{T}^{(k,i)})^\top \mathbf{S}^{(k)} \mathbf{T}^{(k,i)} - \mathbf{S}^{(i)} \right\|_F^2 \\
&+ \alpha \cdot \left\| (\mathbf{T}^{(j,k)})^\top (\mathbf{T}^{(i,j)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} - \mathbf{T}^{(k,i)} \mathbf{S}^{(i)} (\mathbf{T}^{(k,i)})^\top \right\|_F^2 \\
&s.t.\ \mathbf{T}^{(i,j)} \in \{0,1\}^{|\mathcal{U}^{(i)}| \times |\mathcal{U}^{(j)}|}, \mathbf{T}^{(j,k)} \in \{0,1\}^{|\mathcal{U}^{(j)}| \times |\mathcal{U}^{(k)}|} \\
&\quad \mathbf{T}^{(k,i)} \in \{0,1\}^{|\mathcal{U}^{(k)}| \times |\mathcal{U}^{(i)}|} \\
&\quad \mathbf{T}^{(i,j)} \mathbf{1}^{|\mathcal{U}^{(j)}| \times 1} \preccurlyeq \mathbf{1}^{|\mathcal{U}^{(i)}| \times 1}, (\mathbf{T}^{(i,j)})^\top \mathbf{1}^{|\mathcal{U}^{(i)}| \times 1} \preccurlyeq \mathbf{1}^{|\mathcal{U}^{(j)}| \times 1}, \\
&\quad \mathbf{T}^{(j,k)} \mathbf{1}^{|\mathcal{U}^{(k)}| \times 1} \preccurlyeq \mathbf{1}^{|\mathcal{U}^{(j)}| \times 1}, (\mathbf{T}^{(j,k)})^\top \mathbf{1}^{|\mathcal{U}^{(j)}| \times 1} \preccurlyeq \mathbf{1}^{|\mathcal{U}^{(k)}| \times 1}, \\
&\quad \mathbf{T}^{(k,i)} \mathbf{1}^{|\mathcal{U}^{(i)}| \times 1} \preccurlyeq \mathbf{1}^{|\mathcal{U}^{(k)}| \times 1}, (\mathbf{T}^{(k,i)})^\top \mathbf{1}^{|\mathcal{U}^{(k)}| \times 1} \preccurlyeq \mathbf{1}^{|\mathcal{U}^{(i)}| \times 1},
\end{aligned}$$

where parameter α denotes the weight of the alignment transitivity penalty term, which is set as 1 by default in this paper.

C. Relaxation of the Optimization Problem

The above objective function aims at obtaining the *hard* mappings among users across different networks and entries in all these *transitional matrices* are binary, which can lead to a fatal drawback: *hard assignment* can be neither possible nor realistic for networks with star structures as proposed in [13] and the hard subgraph isomorphism [15] is NP-hard.

To overcome such a problem, we propose to relax the binary constraint of entries in transitional matrices to allow them to be real values within range $[0, 1]$. Each entry in the transitional matrix represents a probability, denoting the confidence of certain user-user mapping across networks. Such a relaxation can make the *one-to-one* constraint no longer hold (which will be addressed with transitive network matching in the next subsection) as multiple entries in rows/columns of the transitional matrix can have non-zero values. To limit the existence of non-zero entries in the transitional matrices, we replace the one-to-one constraint, e.g.,

$$\mathbf{T}^{(k,i)} \mathbf{1}^{|\mathcal{U}^{(i)}| \times 1} \preccurlyeq \mathbf{1}^{|\mathcal{U}^{(k)}| \times 1}, (\mathbf{T}^{(k,i)})^\top \mathbf{1}^{|\mathcal{U}^{(k)}| \times 1} \preccurlyeq \mathbf{1}^{|\mathcal{U}^{(i)}| \times 1}$$

with *sparsity constraints*

$$\left\| \mathbf{T}^{(k,i)} \right\|_0 \leq t$$

instead, where term $\|\mathbf{T}\|_0$ denotes the L_0 norm of matrix $\mathbf{T}$, i.e., the number of non-zero entries in $\mathbf{T}$, and t is a small positive number to limit the non-zero entries in the matrix (i.e., the sparsity). Furthermore, in this paper, we propose to add term $\|\mathbf{T}\|_0$ to the minimization objective function, as it can be hard to determine the value of t in the constraint.

Based on the above relaxations, we can obtain the new objective function (available in the Appendix), which involves 3 variables $\mathbf{T}^{(i,j)}$, $\mathbf{T}^{(j,k)}$ and $\mathbf{T}^{(k,i)}$ simultaneously, obtaining the joint optimal solution for which at the same time is very hard and time consuming. We propose to address the above objective function by fixing two variables and updating the other variable alternatively with gradient descent method [1]. As proposed in [13], if during the alternating updating steps,

the entries of the transitional matrices become invalid (i.e., values less than 0 or greater than 1), we apply the projection technique introduced in [13] to project (1) negative entries to 0, and (2) entries greater than 1 to 1 instead. With these processes, the updating equations of matrices $\mathbf{T}^{(i,j)}, \mathbf{T}^{(j,k)}, \mathbf{T}^{(k,i)}$ at step $t+1$ are given as follows

$$\mathbf{T}^{(i,j)}(t+1) = \mathbf{T}^{(i,j)}(t) - \eta^{(i,j)} \frac{\partial \mathcal{L}\left(\mathbf{T}^{(i,j)}(t), \mathbf{T}^{(j,k)}(t), \mathbf{T}^{(k,i)}(t), \beta, \gamma, \theta\right)}{\partial \mathbf{T}^{(i,j)}},$$

$$\mathbf{T}^{(j,k)}(t+1) = \mathbf{T}^{(j,k)}(t) - \eta^{(j,k)} \frac{\partial \mathcal{L}\left(\mathbf{T}^{(i,j)}(t+1), \mathbf{T}^{(j,k)}(t), \mathbf{T}^{(k,i)}(t), \beta, \gamma, \theta\right)}{\partial \mathbf{T}^{(j,k)}},$$

$$\mathbf{T}^{(k,i)}(t+1) = \mathbf{T}^{(k,i)}(t) - \eta^{(k,i)} \frac{\partial \mathcal{L}\left(\mathbf{T}^{(i,j)}(t+1), \mathbf{T}^{(j,k)}(t+1), \mathbf{T}^{(k,i)}(t), \beta, \gamma, \theta\right)}{\partial \mathbf{T}^{(k,i)}}.$$

Such an iteratively updating process will stop when all *transitional matrices* converge. In the updating equations, $\eta^{(i,j)}$, $\eta^{(j,k)}$ and $\eta^{(k,i)}$ are the gradient descent steps in updating $\mathbf{T}^{(i,j)}$, $\mathbf{T}^{(j,k)}$ and $\mathbf{T}^{(k,i)}$ respectively. The Lagrangian function of the objective function is available in the Appendix.

Meanwhile, considering that $\|\cdot\|_0$ is not differentiable because of its discrete values [27], we will replace the $\|\cdot\|_0$ with the $\|\cdot\|_1$ instead (i.e., the sum of absolute values of all entries). Furthermore, as all the negative entries will be projected to 0, the L_1 norm of transitional matrix $\mathbf{T}$ can be represented as $\left\|\mathbf{T}^{(k,i)}\right\|_1 = \mathbf{1}^\top \mathbf{T}^{(k,i)} \mathbf{1}$ (i.e., the sum of all entries in the matrix). In addition, the Frobenius norm $\|\mathbf{X}\|_F^2$ can be represented with trace $\mathrm{Tr}(\mathbf{X}\mathbf{X}^\top)$. The partial derivatives of function $\mathcal{L}$ with regard to $\mathbf{T}^{(i,j)}$, $\mathbf{T}^{(j,k)}$, and $\mathbf{T}^{(k,i)}$ are given in the Appendix.

D. Transitive Network Matching

The constraint relaxation in previous section violates the *one-to-one* property on anchor links seriously. To resolve such a problem, in this section, we will apply the *transitive network matching* to prune the introduced redundant anchor links. The matching results (i.e., selected anchor links) need to meet both the *one-to-one* constraint and *transitivity property*.

Given two networks $G^{(i)}$ and $G^{(j)}$, each potential anchor link, e.g., $(u_l^{(i)}, u_m^{(j)})$, between $G^{(i)}$ and $G^{(j)}$ is associated with a binary variable $x_{l,m}^{(i,j)} \in \{0,1\}$ to denote whether anchor link $(u_l^{(i)}, u_m^{(j)})$ is selected or not in the matching, where

$$x_{l,m}^{(i,j)} = \begin{cases} 1 & \text{if } (u_l^{(i)}, u_m^{(j)}) \text{ is selected,} \\ 0, & \text{otherwise.} \end{cases}$$

For each user in network $G^{(i)}$, e.g., $u_l^{(i)} \in \mathcal{U}^{(i)}$, at most one potential anchor link attached to him will be selected in the final alignment result with another network, e.g., $G^{(j)}$ (or $G^{(k)}$). So, based on the introduced binary variables, the *one-to-one* constraint on anchor links between networks $G^{(i)}$ and $G^{(j)}$ as well as networks $G^{(i)}$ and $G^{(k)}$ can be represented as follows:

$$\sum_{u_m^{(j)} \in \mathcal{U}^{(j)}} x_{l,m}^{(i,j)} \le 1, \sum_{u_o^{(k)} \in \mathcal{U}^{(k)}} x_{l,o}^{(i,k)} \le 1, \forall u_l^{(i)} \in \mathcal{U}^{(i)}.$$

Similarly, we can also define the binary variables $x_{m,o}^{(j,k)}, x_{o,l}^{(k,i)} \in \{0,1\}$ and the corresponding *one-to-one* constraints for potential anchor links $(u_m^{(j)}, u_o^{(k)})$ and $(u_o^{(k)}, u_l^{(i)})$ between networks $G^{(j)}$, $G^{(k)}$ and between networks $G^{(k)}$, $G^{(i)}$ respectively to represent whether these links are selected or not.

Besides the *one-to-one* constraint, the finally selected anchor links should also follow the *transitivity law*.

According to the definition of "transitivity law" in Section I, if anchor links $(u_l^{(i)}, u_m^{(j)})$ and $(u_m^{(j)}, u_o^{(k)})$ are selected $\forall l \in \{1, 2, \cdots, |\mathcal{U}^{(i)}|\}, m \in \{1, 2, \cdots, |\mathcal{U}^{(j)}|\}, o \in \{1, 2, \cdots, \mathcal{U}^{(k)}|\}$ in matching networks $G^{(i)}$, $G^{(j)}$ and networks $G^{(j)}$, $G^{(k)}$, then anchor link $(u_o^{(k)}, u_l^{(i)})$ should be selected as well in the matching of networks $G^{(k)}$, $G^{(i)}$, i.e., $x_{o,l}^{(k,i)} = 1$. In other words, in 3 networks matching, the case that only two variables in $\{x_{l,m}^{(i,j)}, x_{m,o}^{(j,k)}, x_{o,l}^{(k,i)}\}$ are assigned with value 1 while the remaining one is 0 cannot hold in the final matching results, i.e.,

$$x_{l,m}^{(i,j)} + x_{m,o}^{(j,k)} + x_{o,l}^{(k,i)} \neq 2, \forall l \in \{1, 2, \cdots, |\mathcal{U}^{(i)}|\}, \\ \forall m \in \{1, 2, \cdots, |\mathcal{U}^{(j)}|\}, \forall o \in \{1, 2, \cdots, \mathcal{U}^{(k)}|,$$

which is called the *matching transitivity constraint*. The *matching transitivity constraint* can be easily generalized to the case of matching$n$ $(n \ge 3)$ networks.

Definition 8 (Matching Transitivity Constraint): Let $\mathcal{G} = \{G^{(1)}, G^{(2)}, \cdots, G^{(n)}\}$ be a set of n networks, the *matching transitivity constraint* (MTC) for matching these n networks in $\mathcal{G}$ can be defined recursively as follows:

$$\mathrm{MTC}(\mathcal{G}) = \{\sum x_{\mathcal{G}} \neq |\mathcal{G}| - 1\} \cup \{\bigcup_{\mathcal{G}' \subset \mathcal{G}, |\mathcal{G}'| = |\mathcal{G}| - 1} \mathrm{MTC}(\mathcal{G}')\},$$

where $\sum x_{\mathcal{G}} = x_{l,m}^{(1,2)} + x_{m,o}^{(2,3)} + \cdots + x_{p,l}^{(n,1)}, \forall l \in \{1, 2, \cdots, \mathcal{U}^{(1)}\}, \forall m \in \{1, 2, \cdots, \mathcal{U}^{(2)}\}, \cdots, \forall p \in \{1, 2, \cdots, \mathcal{U}^{(n)}\}$ represents the transitivity constraint involving all these n networks.

The final selected anchor links should be those with high confidence scores in the inferred *transitional matrices* but also can meet the *one-to-one matching* constraint and *matching transitivity* constraint simultaneously. We formulate the *transitive network matching* as the following optimization problem:

$$\max_{\mathbf{x}^{(i,j)},\mathbf{x}^{(j,k)},\mathbf{x}^{(k,i)}} \sum_{l,m} x_{l,m}^{(i,j)} \mathbf{T}^{(i,j)}(l,m) + \sum_{l,m} x_{l,m}^{(i,j)} \mathbf{T}^{(i,j)}(l,m) + \sum_{l,m} x_{l,m}^{(i,j)} \mathbf{T}^{(i,j)}(l,m),$$

$$s.t. \sum_{u_m^{(j)} \in \mathcal{U}^{(j)}} x_{l,m}^{(i,j)} \leq 1, \sum_{u_o^{(k)} \in \mathcal{U}^{(k)}} x_{l,o}^{(i,k)} \leq 1, \forall u_l^{(i)} \in \mathcal{U}^{(i)},$$

$$\sum_{u_l^{(i)} \in \mathcal{U}^{(i)}} x_{m,l}^{(j,i)} \leq 1, \sum_{u_o^{(k)} \in \mathcal{U}^{(k)}} x_{m,o}^{(j,k)} \leq 1, \forall u_m^{(j)} \in \mathcal{U}^{(j)},$$

$$\sum_{u_l^{(i)} \in \mathcal{U}^{(i)}} x_{o,l}^{(k,i)} \leq 1, \sum_{u_m^{(j)} \in \mathcal{U}^{(j)}} x_{o,m}^{(k,j)} \leq 1, \forall u_o^{(k)} \in \mathcal{U}^{(k)},$$

$$x_{l,m}^{(i,j)} + x_{m,o}^{(j,k)} + x_{o,l}^{(k,i)} \neq 2, \forall l \in \{1, 2, \cdots, |\mathcal{U}^{(i)}|\},$$

$$\forall m \in \{1, 2, \cdots, |\mathcal{U}^{(j)}|\}, \forall o \in \{1, 2, \cdots, \mathcal{U}^{(k)}|,$$

$$x_{l,m}^{(i,j)} \in \{0, 1\}, \forall u_l^{(i)} \in \mathcal{U}^{(i)}, u_m^{(j)} \in \mathcal{U}^{(j)}.$$

$$x_{m,o}^{(j,k)} \in \{0, 1\}, \forall u_m^{(j)} \in \mathcal{U}^{(j)}, u_o^{(k)} \in \mathcal{U}^{(k)}.$$

$$x_{o,l}^{(k,i)} \in \{0, 1\}, \forall u_o^{(k)} \in \mathcal{U}^{(k)}, u_l^{(i)} \in \mathcal{U}^{(i)}.$$

In the above objective function, the matching transitivity constraint $x_{l,m}^{(i,j)} + x_{m,o}^{(j,k)} + x_{o,l}^{(k,i)} \neq 2$ is actually non-convex, which can be another challenge in addressing the function. In this paper, we propose to (1) remove the matching transitivity constraint from the objective function, and (2) apply the matching transitivity constraint to post-process the solution (obtained from the objective function without the constraint).

The objective function (with the matching transitivity constraint removed) can be solved with open source optimization toolkit, e.g., Scipy.Optimization[1] and GLPK[2], and we will not describe how to solve in details due to the limited space. Among all the obtained solutions, we can check all the links whose corresponding variables meeting $x_{l,m}^{(i,j)} + x_{m,o}^{(j,k)} + x_{o,l}^{(k,i)} = 2$ and assign the variable with value 0 with 1 instead. For example, for 3 given variables $x_{l,m}^{(i,j)}$, $x_{m,o}^{(j,k)}$ and $x_{o,l}^{(k,i)}$, if $x_{l,m}^{(i,j)} = x_{m,o}^{(j,k)} = 1$ but $x_{o,l}^{(k,i)} = 0$, we will assign $x_{o,l}^{(k,i)}$ with new value 1 and $x_{o,x}^{(k,i)} = 0, \forall x \neq l$, $x_{x,l}^{(k,i)} = 0, \forall x \neq o$ to preserve the matching transitivity constraint.

IV. Experiments

To examine the effectiveness of UMA in addressing the M-NASA problem, extensive experiments on real-world multiple partially aligned social networks will be done in this section. Next, we will introduce the dataset used in the experiments in Section IV-A and give brief descriptions about the experiment settings in Section IV-B. Experiment results and detailed analysis will be given in Sections IV-C and IV-D.

[1]http://docs.scipy.org/doc/scipy/reference/optimize.html

[2]http://www.gnu.org/software/glpk/

A. Dataset Description

Nowadays, Question-and-Answer (Q&A) websites are becoming a new platform for people to share knowledge, where individuals can conveniently post their questions online and get first-hand replies very quickly. A large number of Q&A sites have sprung out overnight, e.g., Stack Overflow[3], Super User[4], Programmers[5], Quora[6]. Stack Overflow, Super User and Programmers are all Q&A sites constructed for exchanging knowledge about computer science and share large number of common users, which are used as the partially aligned networks $G^{(i)}$, $G^{(j)}$ and $G^{(k)}$ respectively in the experiments.

We crawled the multiple partially aligned Q&A networks during November 2014-January 2015 and the complete information of $10,000$ users in Stack Overflow, Super User and Programmers Q&A sites respectively. The anchor links (i.e., the ground truth) between pairs of these Q&A networks are obtained by crawling their homepages in these sites respectively, where users' IDs in all these networks they participate in are listed. For example, at site[7], we can have access to all the Q&A sites IDs that Jon Skeet owns, which can be used to extract the ground truth anchor links across networks. Among these 3 networks, the number of shared anchor users (1) between Stack Overflow and Super User is $3,677$, (2) between Stack Overflow and Programmers is $2,626$, (3) between Super User and Programmers is $1,953$. Users in Q&A sites can answer questions which are of their interests. Considering that users don't have social links in these Q&A sites, we will create social connections among users if they have every answered the same question in the past. Answering common questions in Q&A sites denotes that they may share common interests as well as common expertise in certain areas.

B. Experiment Settings

In the experiments, anchor links between users across networks are used for validation only and are not involved in building models. Considering that the network alignment method introduced in this paper is based on the social link information only, isolated users with no social connections in each network are sampled and removed. Based on the social links among users, we infer the optimal transitional matrices between pairs of networks by minimizing the *friendship inconsistency* as well as the alignment transitivity penalty. Alternative updating method is used to solve the joint objective function, where the transitional matrices are initialized with method introduced in [13]. All users in each network are partitioned into 10 bins according to their social degrees, where initial anchor links are assumed to exist between users belonging to the corresponding bins between pairs of networks, e.g., users in bin 1 of Stack Overflow and those in bin 1 of Programmers. The initial values of entries corresponding to these anchor links in transitional matrices are calculated

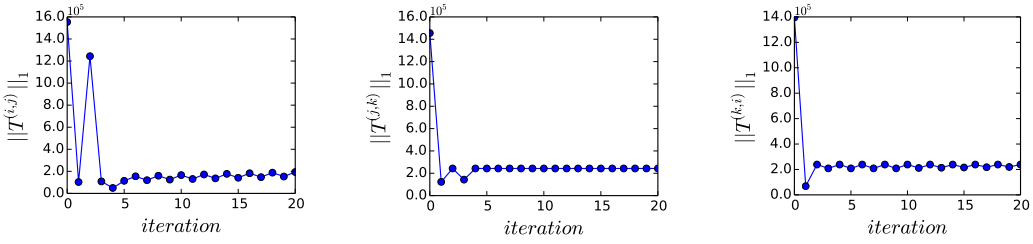
[3]http://stackoverflow.com

[4]http://superuser.com

[5]http://programmers.stackexchange.com

[6]http://www.quora.com

[7]http://stackexchange.com/users/11683/jon-skeet?tab=accounts



(a) Matrix $\mathbf{T}^{(i,j)}$ (b) Matrix $\mathbf{T}^{(j,k)}$ (c) Matrix $\mathbf{T}^{(k,i)}$

Fig. 2. L_1 norm of transitional matrices at each iteration.

with the *relative degree distance* based on their social degrees, e.g., $rdd(u_l^{(i)}, u_m^{(j)}) = \left(1 + \frac{|deg(u_l^{(i)}) - deg(u_m^{(j)})|}{(deg(u_l^{(i)}) + deg(u_m^{(j)}))/2}\right)^{-1}$, where $deg(u)$ denotes the social degree of user u in the networks. Based on the inferred transitional matrices, anchor links with the highest scores but can meet the *one-to-one* constraint and *transitivity law* are selected with the method introduced in Section III-D, which can output both the confidence scores and their inferred labels.

Comparison Methods: Considering that social networks studied in this paper (1) contain only social link information, and (2) no known anchor links exist between networks, therefore, neither inter-network user resolution method MOBIUS [28] built with various user attribute information nor supervised network alignment method MNA [12] can be applied to address the M-NASA problem. To show the advantages of UMA, we compare UMA with many other baseline methods, including both state-of-art network alignment methods as well as extended traditional methods, which are listed as follows.

- *Unsupervised Multi-network Alignment*: Method UMA introduced in this paper can align multiple partially networks concurrently, which include two steps: (1) transitive network alignment, and (2) transitive network matching. Anchor links inferred by UMA can maintain both *one-to-one* constraint and *transitivity property*.
- *Integrated Network Alignment* (INA): To show that transitive network matching can improve the alignment results, we introduce another method named INA, which is identical to the first step of UMA but without the matching step. Anchor links inferred by INA cannot maintain the *one-to-one* constraint nor *transitivity law* property.
- *Pairwise Network Alignment*: BIG-ALIGN is a state-of-art unsupervised network alignment method proposed in [13] for aligning pairwise networks. The output of BIG-ALIGN cannot maintain the *one-to-one* constraint nor *transitivity property* of anchor links.
- *Pairwise Alignment + Pairwise Matching*: We also extend BIG-ALIGN [13] and introduce another baseline method BIG-ALIGN-PM, which can further prune the redundant non-existing anchor links with pairwise network stable matching proposed in [12] to guarantee the inferred anchor links can meet the *one-to-one* constraint.
- *Relative Degree Distance (RDD) based Alignment*: The transitional matrix initialization method RDD [13] is compared as another baseline methods, which calculate the confidence scores of potential anchor links with the degree information of users.
- *Relative PageRank based Alignment*: For completeness, we also extend the traditional PageRank method and propose a new method RPR to infer potential anchor links. For a potential anchor link $(u_l^{(i)}, u_m^{(j)})$, RPR calculates the reciprocal of the relative pagerank scores between $u_l^{(i)}$, $u_m^{(j)}$ as its existence confidence, i.e., $|pagerank(u_l^{(i)}) - pagerank(u_m^{(j)})|^{-1}$.

Evaluation Metrics:

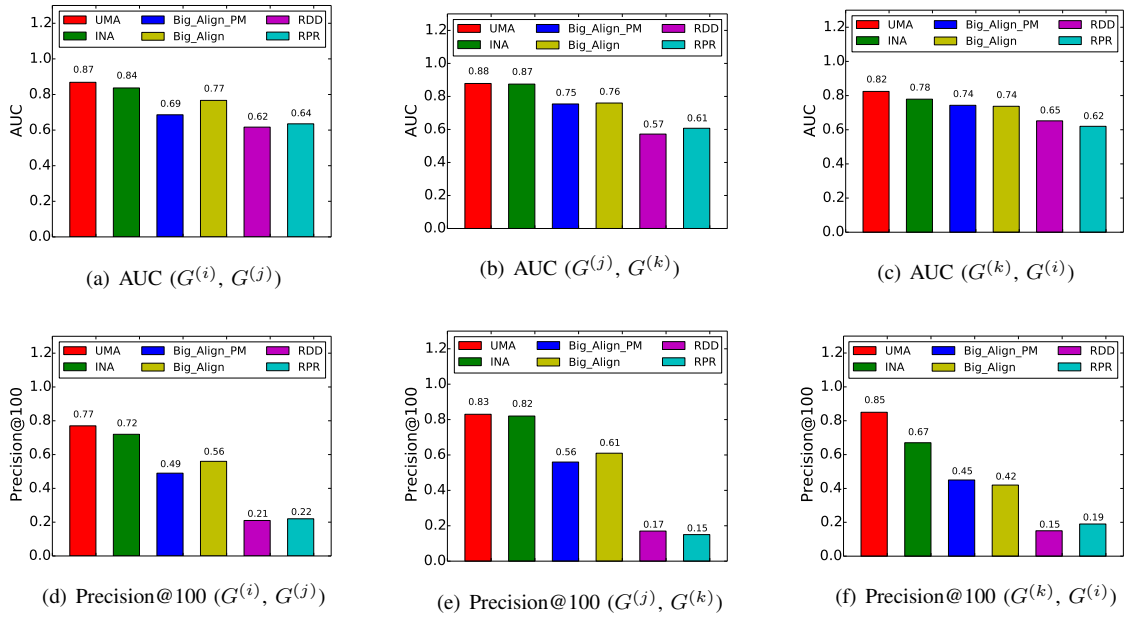
To evaluate the performance of different comparison methods, various commonly used evaluation metrics are applied. All these comparison methods (in INA, the selected anchor links are assigned with scores 1, while those not selected are assigned with scores 0) can output confidence scores of potential anchor links, which are evaluated by metrics AUC and Precision@100. UMA and INA can both output the predicted labels of potential anchor links, which are also evaluated by metrics Accuracy, Precision, Recall and F1.

C. Convergence Analysis

To solve the objective function in Section III-C, alternative updating method is applied to infer the optimal transitional matrices across networks. To demonstrate that the matrix updating equation can converge within a limited iterations, we calculate the L_1 norms (i.e., the sum of all entries' absolute value) of transitional matrices $\mathbf{T}^{(i,j)}$, $\mathbf{T}^{(j,k)}$ and $\mathbf{T}^{(k,i)}$ at each iteration, which are available in Figure 2. As shown in the plots, after a few iterations (about 5 iterations), the L_1 norm of these transitional matrices will converge quickly with minor fluctuations around certain values, which demonstrates that the derived equation updating can converge very well in updating the transitional matrices.

D. Experiment Results

The experiment results of all these comparison methods are available in Figures 3-4 respectively, where performance of all these comparison methods in Figure 3 are evaluated by AUC and Precision@100, while those of UMA and BIG-ALIGN-PM in Figure 4 are evaluated by Precision, Recall, F1 and Accuracy respectively.



(a) AUC $(G^{(i)}, G^{(j)})$ (b) AUC $(G^{(j)}, G^{(k)})$ (c) AUC $(G^{(k)}, G^{(i)})$

(d) Precision@100 $(G^{(i)}, G^{(j)})$ (e) Precision@100 $(G^{(j)}, G^{(k)})$ (f) Precision@100 $(G^{(k)}, G^{(i)})$

Fig. 3. Performance comparison of different methods evaluated by AUC and Precision@100.

In Figure 3, we show the alignment results achieved by all the 6 comparison methods between network pairs $(G^{(i)}, G^{(j)})$, $(G^{(j)}, G^{(k)})$ and $(G^{(k)}, G^{(i)})$. As shown in the plots, UMA performs much better than all the other comparison methods with great advantages in predicting the anchor links between all these networks pairs. For instance, in Figure 3(a), the AUC obtained by UMA is 0.87, which is about 4% larger than INA and over 13% larger than the other comparison methods; in Figure 3(f), the Precision@100 achieved by UMA is 0.85, which is over 25% higher than that of INA, almost the double of that gained by BIG-ALIGN and BIG-ALIGN-PM, and even 4-5 times of that obtained by RDD and RPR.

By comparing UMA and INA, method UMA consisting of transitive integrated network alignment and transitive network matching performs better, which demonstrates the effectiveness of the transitive network matching step in pruning redundant non-existing anchor links.

Compared with the isolated pairwise network alignment method BIG-ALIGN, the fact that INA achieves better performance justifies that aligning multiple networks simultaneously by incorporating the alignment transitivity penalty into the objective function can identify better anchor links than pairwise isolated network alignment.

By comparing BIG-ALIGN-PM and BIG-ALIGN, the pairwise network matching step can help improve the prediction results of anchor links between networks $(G^{(k)}, G^{(i)})$ but has no positive effects (even has negative effects) on the anchor links between other network pairs, e.g., network pairs $(G^{(i)}, G^{(j)})$ and $(G^{(j)}, G^{(k)})$. However, the effective of the transitive network matching method applied in UMA has been proved in the comparison of UMA and INA. It may show that transitive network matching exploiting the transitivity law performs much better than the pairwise network matching method.

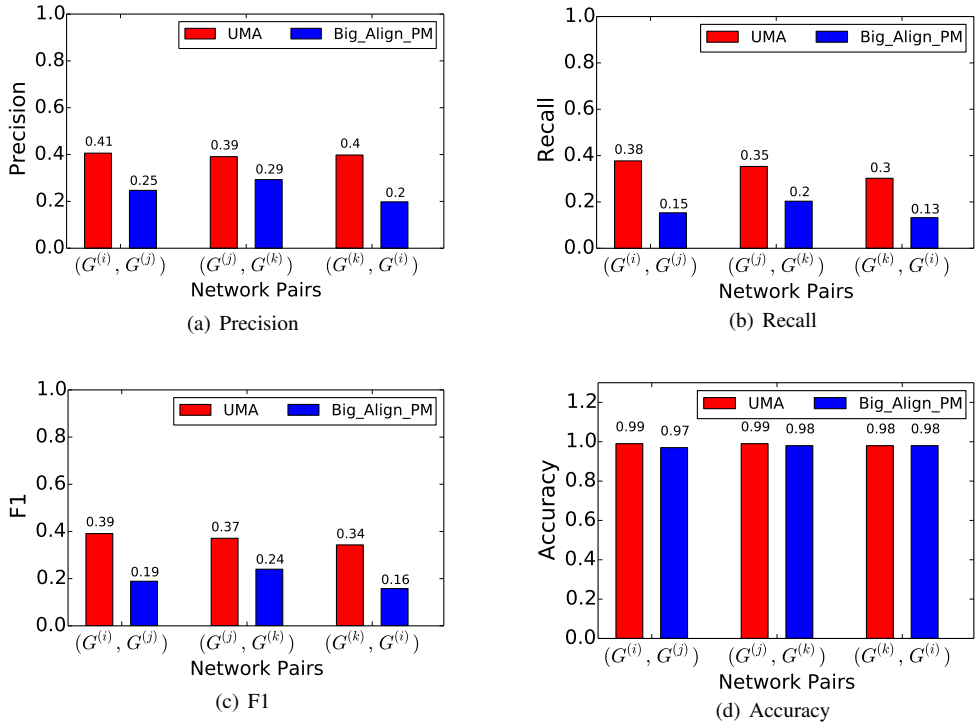
For completeness, we also compare UMA with extensions of traditional methods RDD and RPR and the advantages of UMA over these methods are very obvious.

Furthermore, in Figure 4, we also show the results obtained by UMA and BIG-ALIGN-PM between each network pair. The Precision, Recall and F1 scores achieved by UMA in Figures 4(a)-4(c) are all much higher than those obtained by BIG-ALIGN-PM, which demonstrates the effective and advantages of UMA over BIG-ALIGN. However, Accuracy scores of UMA and BIG-ALIGN-PM in Figure 4(d) are both very high (around 100%) with negligible gaps, which may be due to the class imbalance issues [5].

V. RELATED WORKS

Graph alignment is an important research problem and dozens of papers have been published on this topic in the past decades. Depending on specific disciplines, the studied graphs can be social networks in data mining [12] protein-protein interaction (PPI) networks and gene regulatory networks in bioinformatics [10], [21], [16], [22], chemical compound in chemistry [24], data schemas in data warehouse [18], ontology in web semantics [7], graph matching in combinatorial mathematics [17], as well as graphs in computer vision and pattern recognition [6], [3].

In bioinformatics, the network alignment problem aims at predicting the best mapping between two biological networks based on the similarity of the molecules and their interaction patterns. By studying the cross-species variations of biological networks, network alignment problem can be applied to predict conserved functional modules [20] and infer the functions of proteins [19]. Graemlin [8] conducts pairwise network



(a) Precision (b) Recall (c) F1 (d) Accuracy

Fig. 4. Performance comparison of UMA and BIG-ALIGN-PM evaluated by Precision, Recall, F1 and Accuracy.

alignment by maximizing an objective function based on a set of learned parameters. Some works have been done on aligning multiple network in bioinformatics. IsoRank proposed in [23] can align multiple networks greedily based on the pairwise node similarity scores calculated with spectral graph theory. IsoRankN [16] further extends IsoRank by exploiting a spectral clustering scheme.

In recent years, with rapid development of online social networks, researchers' attention starts to shift to the alignment of social networks. Enlightened by the homogeneous network alignment method in [26], Koutra et al. [13] propose to align two bipartite graphs with a fast alignment algorithm. Zafarani et al. [28] propose to match users across social networks based on various node attributes, e.g., username, typing patterns and language patterns etc. Kong et al. formulate the heterogeneous social network alignment problem as an anchor link prediction problem. A two-step supervised method MNA is proposed in [12] to infer potential anchor links across networks with heterogeneous information in the networks. However, social networks in the real world are mostly partially aligned actually and lots of users are not anchor users. Zhang et al. have proposed the partial network alignment methods based on supervised learning setting and PU learning setting in [32] and [34] respectively.

VI. CONCLUSION

In this paper, we have studied the *multiple anonymized social network alignment* (M-NASA) problem to infer the anchor links across multiple anonymized online social networks simultaneously. An effective two-step multiple network alignment framework UMA has been proposed to address the M-NASA problem. The anchor links to be inferred follow both *transitivity law* and *one-to-one* property, under the constraint of which, UMA matches multiple anonymized networks by minimizing the *friendship inconsistency* and selects anchor links which can lead to the maximum confidence scores across multiple anonymized social networks.

VII. ACKNOWLEDGMENT

This work is supported in part by NSF through grants III-1526499, CNS-1115234, and OISE-1129076, Google Research Award, and the Pinnacle Lab at Singapore Management University.

REFERENCES

[1] M. Avriel. *Nonlinear Programming: Analysis and Methods*. Prentice-Hall, Englewood Cliffs, NJ, 1976.
[2] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. In *WWW*, 2007.
[3] M. Bayati, M. Gerritsen, D. Gleich, A. Saberi, and Y. Wang. Algorithms for large, sparse network alignment problems. In *ICDM*, 2009.
[4] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. *TKDD*, 2007.
[5] N. Chawla. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*. 2005.
[6] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *IJPRAI*, 2004.
[7] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Ontology matching: A machine learning approach. In *Handbook on Ontologies*. 2004.
[8] J. Flannick, A. Novak, B. Srinivasan, H. McAdams, and S. Batzoglou. Graemlin: general and robust alignment of multiple large interaction networks. *Genome research*, 2006.

[9] S. Jin, J. Zhang, P. Yu, S. Yang, and A. Li. Synergistic partitioning in multiple large scale social networks. In *IEEE BigData*, 2014.
[10] M. Kalaev, V. Bafna, and R. Sharan. Fast and accurate alignment of multiple protein networks. In *RECOMB*. 2008.
[11] A. Khan, D. Gleich, A. Pothen, and M. Halappanavar. A multithreaded algorithm for network alignment via approximate matching. In *SC*, 2012.
[12] X. Kong, J. Zhang, and P. Yu. Inferring anchor links across multiple heterogeneous social networks. In *CIKM*, 2013.
[13] D. Koutra, H. Tong, and D. Lubensky. Big-align: Fast bipartite graph alignment. In *ICDM*, 2013.
[14] K. Kunen. *Set Theory*. ELSEVIER SCIENCE PUBLISHERS, 1980.
[15] J. Lee, W. Han, R. Kasperovics, and J. Lee. An in-depth comparison of subgraph isomorphism algorithms in graph databases. *VLDB*, 2012.
[16] C. Liao, K. Lu, M. Baym, R. Singh, and B. Berger. Isorankn: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 2009.
[17] F. Manne and M. Halappanavar. New effective multithreaded matching algorithms. In *IPDPS*, 2014.
[18] S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *ICDE*, 2002.
[19] D. Park, R. Singh, M. Baym, C. Liao, and B. Berger. Isobase: a database of functionally related proteins across ppi networks. *Nucleic Acids Research*, 2011.
[20] R. Sharan, S. Suthram, R. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. 2005.
[21] Y. Shih and S. Parthasarathy. Scalable global alignment for multiple biological networks. *Bioinformatics*, 2012.
[22] R. Singh, J. Xu, and B. Berger. Pairwise global alignment of protein interaction networks by matching neighborhood topology. In *RECOMB*, 2007.
[23] R. Singh, J. Xu, and B. Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 2008.
[24] A. Smalter, J. Huan, and G. Lushington. Gpm: A graph pattern matching kernel with diffusion for chemical compound classification. In *IEEE BIBE*, 2008.
[25] M. Tsikerdekis and S. Zeadally. Multiple account identity deception detection in social media using nonverbal behavior. *IEEE TIFS*, 2014.
[26] S. Umeyama. An eigendecomposition approach to weighted graph matching problems. *IEEE TPAMI*, 1988.
[27] D. Wipf and B. Rao. L0-norm minimization for basis selection. In *NIPS*, 2005.
[28] R. Zafarani and H. Liu. Connecting users across social media sites: A behavioral-modeling approach. In *KDD*, 2013.
[29] Q. Zhan, J. Zhang, S. Wang, P. Yu, and J. Xie. Influence maximization across partially aligned heterogenous social networks. In *PAKDD*, 2015.
[30] J. Zhang, X. Kong, and P. Yu. Predicting social links for new users across aligned heterogeneous social networks. In *ICDM*, 2013.
[31] J. Zhang, X. Kong, and P. Yu. Transfer heterogeneous links across location-based social networks. In *WSDM*, 2014.
[32] J. Zhang, W. Shao, S. Wang, X. Kong, and P. Yu. Partial network alignment with anchor meta path and truncated generalized stable matching. In *IRI*, 2015.
[33] J. Zhang and P. Yu. Community detection for emerging networks. In *SDM*, 2015.
[34] J. Zhang and P. Yu. Integrated anchor and social link predictions across social networks. In *IJCAI*, 2015.
[35] J. Zhang and P. Yu. Mcd: Mutual clustering across multiple heterogeneous networks. In *IEEE BigData Congress*, 2015.
[36] J. Zhang, P. Yu, and Z. Zhou. Meta-path based multi-network collective link prediction. In *KDD*, 2014.

VIII. Appendix: New Objective Function

Based on the above relaxations used in Section III-C, the new objective function can be represented as

$$
\begin{aligned}
&\bar{\mathbf{T}}^{(i,j)}, \bar{\mathbf{T}}^{(j,k)}, \bar{\mathbf{T}}^{(k,i)} \\
&= \arg\min_{\mathbf{T}^{(i,j)}, \mathbf{T}^{(j,k)}, \mathbf{T}^{(k,i)}} \left\| (\mathbf{T}^{(i,j)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} - \mathbf{S}^{(j)} \right\|_F^2 \\
&+ \left\| (\mathbf{T}^{(j,k)})^\top \mathbf{S}^{(j)} \mathbf{T}^{(j,k)} - \mathbf{S}^{(k)} \right\|_F^2 + \left\| (\mathbf{T}^{(k,i)})^\top \mathbf{S}^{(k)} \mathbf{T}^{(k,i)} - \mathbf{S}^{(i)} \right\|_F^2 \\
&+ \alpha \cdot \left\| (\mathbf{T}^{(j,k)})^\top (\mathbf{T}^{(i,j)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} - \mathbf{T}^{(k,i)} \mathbf{S}^{(i)} (\mathbf{T}^{(k,i)})^\top \right\|_F^2 \\
&+ \beta \cdot \left\| \mathbf{T}^{(i,j)} \right\|_0 + \gamma \cdot \left\| \mathbf{T}^{(j,k)} \right\|_0 + \theta \cdot \left\| \mathbf{T}^{(k,i)} \right\|_0 \\
&s.t.\ \mathbf{0}^{|\mathcal{U}^{(i)}| \times |\mathcal{U}^{(j)}|} \preccurlyeq \mathbf{T}^{(i,j)} \preccurlyeq \mathbf{1}^{|\mathcal{U}^{(i)}| \times |\mathcal{U}^{(j)}|}, \\
&\quad \mathbf{0}^{|\mathcal{U}^{(j)}| \times |\mathcal{U}^{(k)}|} \preccurlyeq \mathbf{T}^{(j,k)} \preccurlyeq \mathbf{1}^{|\mathcal{U}^{(j)}| \times |\mathcal{U}^{(k)}|}, \\
&\quad \mathbf{0}^{|\mathcal{U}^{(k)}| \times |\mathcal{U}^{(i)}|} \preccurlyeq \mathbf{T}^{(k,i)} \preccurlyeq \mathbf{1}^{|\mathcal{U}^{(k)}| \times |\mathcal{U}^{(i)}|}.
\end{aligned}
$$

The Lagrangian function of the objective function can be represented as

$$
\begin{aligned}
&\mathcal{L}(\mathbf{T}^{(i,j)}, \mathbf{T}^{(j,k)}, \mathbf{T}^{(k,i)}, \beta, \gamma, \theta) = \left\| (\mathbf{T}^{(i,j)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} - \mathbf{S}^{(j)} \right\|_F^2 \\
&+ \left\| (\mathbf{T}^{(j,k)})^\top \mathbf{S}^{(j)} \mathbf{T}^{(j,k)} - \mathbf{S}^{(k)} \right\|_F^2 + \left\| (\mathbf{T}^{(k,i)})^\top \mathbf{S}^{(k)} \mathbf{T}^{(k,i)} - \mathbf{S}^{(i)} \right\|_F^2 \\
&+ \alpha \cdot \left\| (\mathbf{T}^{(j,k)})^\top (\mathbf{T}^{(i,j)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} - \mathbf{T}^{(k,i)} \mathbf{S}^{(i)} (\mathbf{T}^{(k,i)})^\top \right\|_F^2 \\
&+ \beta \cdot \left\| \mathbf{T}^{(i,j)} \right\|_0 + \gamma \cdot \left\| \mathbf{T}^{(j,k)} \right\|_0 + \theta \cdot \left\| \mathbf{T}^{(k,i)} \right\|_0.
\end{aligned}
$$

The partial derivatives of function $\mathcal{L}$ with regard to $\mathbf{T}^{(i,j)}$, $\mathbf{T}^{(j,k)}$, and $\mathbf{T}^{(k,i)}$ will be:

$$
\begin{aligned}
(1)\ &\frac{\partial \mathcal{L}\left(\mathbf{T}^{(i,j)}, \mathbf{T}^{(j,k)}, \mathbf{T}^{(k,i)}, \beta, \gamma, \theta\right)}{\partial \mathbf{T}^{(i,j)}} \\
&= 2 \cdot \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} (\mathbf{T}^{(i,j)})^\top (\mathbf{S}^{(i)})^\top \mathbf{T}^{(i,j)} \\
&+ 2 \cdot (\mathbf{S}^{(i)})^\top \mathbf{T}^{(i,j)} (\mathbf{T}^{(i,j)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} \\
&+ 2\alpha \cdot \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} (\mathbf{T}^{(j,k)})^\top (\mathbf{T}^{(i,j)})^\top (\mathbf{S}^{(i)})^\top \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} (\mathbf{T}^{(j,k)})^\top \\
&+ 2\alpha \cdot (\mathbf{S}^{(i)})^\top \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} (\mathbf{T}^{(j,k)})^\top (\mathbf{T}^{(i,j)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} (\mathbf{T}^{(j,k)})^\top \\
&- 2 \cdot \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} (\mathbf{S}^{(j)})^\top - 2 \cdot (\mathbf{S}^{(i)})^\top \mathbf{T}^{(i,j)} \mathbf{S}^{(j)} \\
&- 2\alpha \cdot (\mathbf{S}^{(i)})^\top \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} \mathbf{T}^{(k,i)} \mathbf{S}^{(i)} (\mathbf{T}^{(k,i)})^\top (\mathbf{T}^{(j,k)})^\top \\
&- 2\alpha \cdot \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} \mathbf{T}^{(k,i)} (\mathbf{S}^{(i)})^\top (\mathbf{T}^{(k,i)})^\top (\mathbf{T}^{(j,k)})^\top - \beta \cdot \mathbf{1}\mathbf{1}^\top. \\
(2)\ &\frac{\partial \mathcal{L}\left(\mathbf{T}^{(i,j)}, \mathbf{T}^{(j,k)}, \mathbf{T}^{(k,i)}, \beta, \gamma, \theta\right)}{\partial \mathbf{T}^{(j,k)}} \\
&= 2 \cdot \mathbf{S}^{(j)} \mathbf{T}^{(j,k)} (\mathbf{T}^{(j,k)})^\top (\mathbf{S}^{(j)})^\top \mathbf{T}^{(j,k)} \\
&+ 2 \cdot (\mathbf{S}^{(j)})^\top \mathbf{T}^{(j,k)} (\mathbf{T}^{(j,k)})^\top \mathbf{S}^{(j)} \mathbf{T}^{(j,k)} \\
&+ 2\alpha \cdot (\mathbf{T}^{(i,j)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} (\mathbf{T}^{(j,k)})^\top (\mathbf{T}^{(i,j)})^\top (\mathbf{S}^{(i)})^\top \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} \\
&+ 2\alpha \cdot (\mathbf{T}^{(i,j)})^\top (\mathbf{S}^{(i)})^\top \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} (\mathbf{T}^{(j,k)})^\top (\mathbf{T}^{(i,j)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} \\
&- 2 \cdot \mathbf{S}^{(j)} \mathbf{T}^{(j,k)} (\mathbf{S}^{(k)})^\top - 2 \cdot (\mathbf{S}^{(j)})^\top \mathbf{T}^{(j,k)} \mathbf{S}^{(k)} \\
&- 2\alpha \cdot (\mathbf{T}^{(i,j)})^\top (\mathbf{S}^{(i)})^\top \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} \mathbf{T}^{(k,i)} \mathbf{S}^{(i)} (\mathbf{T}^{(k,i)})^\top \\
&- 2\alpha \cdot (\mathbf{T}^{(i,j)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} \mathbf{T}^{(k,i)} (\mathbf{S}^{(i)})^\top (\mathbf{T}^{(k,i)})^\top - \gamma \cdot \mathbf{1}\mathbf{1}^\top. \\
(3)\ &\frac{\partial \mathcal{L}\left(\mathbf{T}^{(i,j)}, \mathbf{T}^{(j,k)}, \mathbf{T}^{(k,i)}, \beta, \gamma, \theta\right)}{\partial \mathbf{T}^{(k,i)}} \\
&= 2 \cdot \mathbf{S}^{(k)} \mathbf{T}^{(k,i)} (\mathbf{T}^{(k,i)})^\top (\mathbf{S}^{(k)})^\top \mathbf{T}^{(k,i)} \\
&+ 2 \cdot (\mathbf{S}^{(k)})^\top \mathbf{T}^{(k,i)} (\mathbf{T}^{(k,i)})^\top \mathbf{S}^{(k)} \mathbf{T}^{(k,i)} \\
&+ 2\alpha \mathbf{T}^{(k,i)} (\mathbf{S}^{(i)})^\top (\mathbf{T}^{(k,i)})^\top \mathbf{T}^{(k,i)} \mathbf{S}^{(i)} \\
&+ 2\alpha \mathbf{T}^{(k,i)} \mathbf{S}^{(i)} (\mathbf{T}^{(k,i)})^\top \mathbf{T}^{(k,i)} (\mathbf{S}^{(i)})^\top \\
&- 2 \cdot \mathbf{S}^{(k)} \mathbf{T}^{(k,i)} (\mathbf{S}^{(i)})^\top - 2 \cdot (\mathbf{S}^{(k)})^\top \mathbf{T}^{(k,i)} \mathbf{S}^{(i)} \\
&- 2\alpha \cdot (\mathbf{T}^{(j,k)})^\top (\mathbf{T}^{(i,j)})^\top (\mathbf{S}^{(i)})^\top \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} \mathbf{T}^{(k,i)} \mathbf{S}^{(i)} \\
&- 2\alpha \cdot (\mathbf{T}^{(j,k)})^\top (\mathbf{T}^{(i,j)})^\top \mathbf{S}^{(i)} \mathbf{T}^{(i,j)} \mathbf{T}^{(j,k)} \mathbf{T}^{(k,i)} (\mathbf{S}^{(i)})^\top - \theta \cdot \mathbf{1}\mathbf{1}^\top.
\end{aligned}
$$