Social Network De-anonymization with Overlapping Communities: Analysis, Algorithm and Experiments

Xinyu Wu[1], Zhongzhao Hu[2], Xinzhe Fu[2], Luoyi Fu[2], Xinbing Wang[1,2], Songwu Lu[3]
[1,2]Dept. of {Electronic Engineering, Computer Science}, Shanghai Jiao Tong University, China.
[3]Dept. of Computer Science, University of California, Los Angeles, U.S.A.
Email:[1]{wuxinyu,xwang8}@sjtu.edu.cn, [2]{hzz5611577,fxz0114,yiluofu}@sjtu.edu.cn, [3]slu@cs.ucla.edu.

***Abstract*—The advent of social networks poses severe threats on user privacy as adversaries can de-anonymize users' identities by mapping them to correlated cross-domain networks. Without ground-truth mapping, prior literature proposes various cost functions in hope of measuring the quality of mappings. However, there is generally a lacking of rationale behind the cost functions, whose minimizer also remains algorithmically unknown.**

We jointly tackle above concerns under a more practical social network model parameterized by *overlapping communities*, which, neglected by prior art, can serve as side information for de-anonymization. Regarding the unavailability of ground-truth mapping to adversaries, by virtue of the Minimum Mean Square Error (MMSE), our first contribution is a well-justified cost function minimizing the expected number of mismatched users over all possible true mappings. While proving the NP-hardness of minimizing MMSE, we validly transform it into the weighted-edge matching problem (WEMP), which, as disclosed theoretically, resolves the tension between optimality and complexity: (i) WEMP asymptotically returns a negligible mapping error in large network size under mild conditions facilitated by higher overlapping strength; (ii) WEMP can be algorithmically characterized via the convex-concave based de-anonymization algorithm (CBDA), perfectly finding the optimum of WEMP. Extensive experiments further confirm the effectiveness of CBDA under overlapping communities, in terms of averagely 90% re-identified users in the rare true cross-domain co-author networks when communities overlap densely, and roughly 70% enhanced re-identification ratio compared to non-overlapping cases.

I. Introduction

With the mounting popularity of social networks, the privacy of users has been under great concern, as information of users in social networks is often released to public for wide usage in academy or advertisement [8]. Although users can be anonymized by removing personal identifiers such as names and family addresses, it is not sufficient for privacy protection since adversaries may re-identify these users by correlated side information, for example the cross domain networks where the identities of these users are unveiled [8].

Such user identification process in social networks resorting to auxiliary information is called *Social Network De-anonymization*. Initially proposed by Narayanan and Shimatikov [2], this fundamental issue has then gained increasing attention, leading to a large body of subsequent works [3]–[9]. Particularly, this family of works embarked on de-anonymization under a common framework, as will also be the framework of interest in our setting. To elaborate, in the framework there is an underlying network G which characterizes the relationship among users. Then there are two networks observed in reality, named as published network G_1 and auxiliary network G_2, whose node sets are identical and edges are independently sampled from G with probability s_1 and s_2 respectively. *The aim of de-anonymization is to discover the correct mapping between V_1 and V_2, which corresponds the same user in two networks, with the network structure as the only side information available to the adversaries.*

Regardless of the considerable efforts paid to de-anonymization, there is still a severe lacking of a comprehensive understanding about the conditions under which the adversaries can perfectly de-anonymize user identities. It can be accounted for from three aspects. (i) Analytically, despite a variety of existing work [3], [4] that proposed several cost functions in measuring the quality of mappings, the theoretical devise of those costs functions lacks sufficient rationale behind. (ii) Algorithmically, previous works [3], [4] failed to provide any algorithm to demonstrate that the optimal solution of proposed cost functions can indeed be effectively obtained. (iii) Experimentally, due to the destitution of real cross-domain datasets, state-of-the-art research [6], [7] simply evaluated the performance of proposed algorithms on synthetic datasets or real cross-domain networks formed by artificial sampling, falling short of reproducing the genuine social networks.

The above limitations motivate us to shed light on de-anonymization problem by jointly incorporating analytical, algorithmic and experimental aspects under the common framework noted earlier. As far as we know, the only work that shares the closest correlation with us belongs to Fu et. al. [22], [23], who investigated this problem on social networks with non-overlapping communities and derived their cost function from the Maximum A Posterior (MAP) manner. However, we remark that the assumption of disjoint communities fails to reflect the real situation where a user belongs to multiple communities, as observed in massive real situations. For example, in social networks of scientific collaborators [9], actors and political blogospheres [18], users might belong to several research groups with different research topics, movies and political parties respectively. Furthermore, while MAP enables adversaries to find the correct mapping with the highest probability, it relies heavily on a prerequisite, i.e., a hypothetically true mapping between the given published and auxiliary networks. However, once the MAP estimation fails to exactly match this "true" mapping, then the mapping

error becomes unpredictable, with the probability that the estimation deviates largely from the real ground-truth. For the first concern, by adopting the overlapping stochastic block model (OSBM), we allow the communities to overlap arbitrarily, which can well capture a majority of real social networks. For the second concern, we derive our cost function based on Minimum Mean Square Error (MMSE), which minimizes the expected number of mismatched users by incorporating all the possible true mappings between the given published and auxiliary networks. This incorporation, from an average perspective, keeps the estimation of MMSE from significant deviation from any possible hypothetic true mapping.

Hereinafter we unfold our main contributions in analytical, algorithmic and experimental aspects respectively as follows:

1. Analytically, we are the first to derive cost function based on MMSE, which justifiably ensures the minimum expected mapping error between our estimation and the ground-truth mapping. Then we demonstrate the NP-hardness of solving MMSE, whose intractability stems mainly from the calculation of all $n!$ possible mappings (n is the total number of users). To cope with the hardness, we simplify MMSE by transforming it into a weighted-edge matching problem (WEMP), with mapping error negatively related to weights.

2. Algorithmically, in terms of solving WEMP, we theoretically reveal that WEMP alleviates the tension between optimality and complexity: Solving WEMP ensures optimality since its optimum, in large network size, negligibly deviates from the ground-truth mapping under mild conditions where on average a user belongs to asymptotically non-constant communities. Meanwhile it reduces complexity since perfectly deriving its optimum only entails a convex-concave based de-anonymization algorithm (CBDA) with polynomial time. The proposed CBDA serves as one of the very few attempts to address the algorithmic characterization, that has long remained open, of de-anonymization without pre-identification.

3. Experimentally, we validate our theoretical findings that minimizing WEMP indeed incurs negligible mapping error in large social networks based on real datasets. Interestingly, we also observe significant benefits that community overlapping effect brings to the performance of CBDA: (i) in notable true cross-domain co-author networks with dense overlapping communities, CBDA can correctly re-identify 90% nodes on average; (ii) the overlapping communities bring about an enhancement of around 70% re-identification ratio compared with non-overlapping cases.

Unlike de-anonymization with pre-identified seed nodes, to which a family of work pays endeavor, no prior knowledge of such seeds complicates this problem, thus leaving many aspects largely unexplored. Meanwhile, theoretical results on such seedless cases in prior art is short of experimental verification. Our work is, as far as we are concerned, the initial devotion to theoretically dissecting seedless cases with overlapping communities, under real cross-domain networks with more than 3000 nodes. With novel exploitations of structural information, future design of more efficient mechanisms will be expected to further dilute the limitation of network size.

II. Related Works

Narayanan and Shimatikove [2] formulated social network de-anonymization problem initially and proposed a generic algorithm based on some pre-identified (seeded) nodes. Predicated on this seminal paper, amounts of work zoomed in on de-anonymization with seed nodes or not. For seeded networks, Yartseva et al. [12], Kazemi et al. [13] and Fabiana et al. [14] studied de-anonymization under Erdős-Rényi graph, while Korula and Lattenzi [15] shed light on it under preferential attachment model. For seedless networks, Pedarsani and Grossglauer [3] are precursors studying this problem under Erdős-Rényi graph. Kazemi et al. [4] considered the partial overlapping of nodes in two networks. Onaran et. al. [8] justified a cost function based on Maximum A Posterior (MAP) and Fu et. al. [22], [23] algorithmically solved it.

For the clustering effect, Chiasserini et al. [16] studied clustering under seeded de-anonymization problem and pointed that the clustering reduces seeded nodes while crippling algorithmic robustness. Onaran et al. [8] modeled clustering as communities and Fu et al. [22], [23] showed that the community enhance re-identification accuracy in seedless cases. However, as far as we know, no existing work has focused on overlapping communities, an omnipresent case in large-scale social networks.

III. Models and Definitions

In this section, we will introduce the fundamental model and some related definitions. Before we start, we list some basic notations frequently used in our later analysis.

A. Preliminary Notations

Definition 1. ***(Expectation Over Matrix)*** *Given a random matrix variable* $\mathbf{A}$ *and a function* $f(\mathbf{A})$*, the expectation of* $f(\mathbf{A})$ *over matrix* $\mathbf{A}$ *is denoted as* $\boldsymbol{E}_{\mathbf{A}}(f(\mathbf{A}))$*.*

Definition 2. ***(Frobenius Norm)*** *Given an* $m \times n$ *matrix* $\mathbf{X}$*, the Frobenius norm of* $\mathbf{X}$ *is* $||\mathbf{X}||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}(\mathbf{X}_{ij}^2)}$*, where* $\mathbf{X}_{ij}$ *is the element at the* i_{th} *row and* j_{th} *column of* $\mathbf{X}$*.*

Definition 3. ***(Hadamard Product)*** *Given two* $n \times n$ *matrices* $\mathbf{Y}$ *and* $\mathbf{Z}$*, The Hadamard Product between* $\mathbf{Y}$ *and* $\mathbf{Z}$ *is defined as* $\forall i, j \in \{1, 2, ..., n\}, (\mathbf{Y} \circ \mathbf{Z})_{ij} = \mathbf{Y}_{ij}\mathbf{Z}_{ij}$*.*

B. Social Network Models

The social network model considered in this paper is composed of three parts, i.e., the underlying network G, the published network G_1 and the auxiliary network G_2. G_1 and G_2 can be viewed as the incomplete observations of G, which represents the underneath relationship among all users. For instance, in reality G may characterize the true underlying relationship among a group of people, while G_1 might represent the online network in Facebook of this group of people and G_2 might represent the communication records in the cell phones of them, both of which are observable.

1) Underlying Social Network: Let $G = (V, E, \mathbf{U})$, where V is the node set, E is the edge set and $\mathbf{U}$ is the adjacent matrix. We regard G as undirected with $|V| = n$ nodes. To reflect the property of overlapping communities, we suppose G is generated based on the overlapping stochastic block model (OSBM) [18], whose idea can be interpreted as follows:

Suppose there are Q communities in G, where each community $q \in Q$ contains a subset of nodes. For a generic node i, we introduce a latent Q-dimensional column vector $\boldsymbol{C_i}$, in which all elements are independent boolean variables $C_{iq} \in \{0, 1\}$, with C_{iq} being the q_{th} row in $\boldsymbol{C_i}$. $C_{iq} = 1$ means that node i is in community q and $C_{iq} = 0$ otherwise. Thus $\boldsymbol{C_i}$ can be seen as drawn from the Bernoulli distribution: $\boldsymbol{C_i} \sim \prod_{q=1}^{Q} (p_q)^{C_{iq}} (1 - p_q)^{1 - C_{iq}}$, where p_q is the probability of any node in G belonging to community q. We call $\boldsymbol{C_i}$ the *community representation* of node i, since $\boldsymbol{C_i}$ shows to which communities node i belongs exactly.

In OSBM, the probability of edge existence between nodes i and j in G relies on $\boldsymbol{C_i}$ and $\boldsymbol{C_j}$. Hence we denote $Pr\{(i, j) \in E\} = p_{\boldsymbol{C_i C_j}}$, where $p_{\boldsymbol{C_i C_j}}$ is pre-defined depending on the number of communities nodes i and j co-exist in, which is easy to obtain in real de-anonymization.

2) Published Network and Auxiliary Network: We let $G_1(V_1, E_1, \mathbf{A})$ denote the published network, whose node labeling is identical with the underlying graph G and edges are independently sampled from G with probability s_1. In contrast, an auxiliary network, denoted by $G_2(V_2, E_2, \mathbf{B})$, does not necessarily share the same node labeling as G, and the edges are independently sampled from G with probability s_2. $\mathbf{A}$ and $\mathbf{B}$ respectively represent the adjacency matrix of G_1 and G_2. In correspondence to real situations, G_1 characterizes the anonymized network where users' identities are unavailable for privacy concern. On the contrary, G_2 characterizes an un-anonymized network where users' identities are all available.

Adversaries can leverage G_2 to identify nodes in G_1 based on the edge relationship and community information: (i) For edge relationship, adversaries can harness the *degree similarity* that a node of high degree in G_1 should be inclined to match a node of high degree in G_2; (ii) For community information, adversaries can exploit the *community representation similarity* that nodes in G_1 and G_2 with the same community representation should be matched with higher probability.

For the edge set E_k $(k \in \{1, 2\})$ of either network, $Pr\{(i, j) \in E_k\} = s_k$ if $(i, j) \in E_k$ and $Pr\{(i, j) \in E_k\} = 0$ otherwise. For the node sets V_1 and V_2, we assume same number of nodes in G, G_1 and G_2, i.e., $|V| = |V_1| = |V_2| = n$ for convenience. Note that it is easy to extend to the situation where $|V_1| \neq |V_2|$ as shown in Section III-C.

Furthermore, we should clarify that we render each node pair (i, j) a weight w_{ij}, which, quantified in Section III-C, is the cost of mistakenly matching the node pair (i, j) and is contingent on $p_{\boldsymbol{C_i C_j}}$, s_1 and s_2. As we will show in Section IV-A, w_{ij} is negatively proportional to the number of communities nodes i and j co-exist in, evincing the cost reduction arose from higher *overlapping strength* of communities.

Remark: In fact G, G_1 and G_2 are all random variables. We directly use G, G_1, G_2 as notations for the realizations of these random variables with no loss of clearance. Moreover, we set $\boldsymbol{\theta} = \{\{p_{\boldsymbol{C_i C_j}} | 1 \leq i, j \leq n\}, s_1, s_2\}$ as the parameter set incorporating all pre-defined parameters in the model.

C. Social Network De-anonymization

The goal of social network de-anonymization problem is to find a mapping $\pi : V_1 \mapsto V_2$, which corresponds nodes on behalf of the same user in G_1 and G_2. We can equivalently express this mapping by forming a permutation matrix $\mathbf{\Pi} \in \{0, 1\}^{n \times n}$, where $\mathbf{\Pi}(i, j) = 1$ if $\pi(i) = j$ and $\mathbf{\Pi}(i, j) = 0$ otherwise (If $|V_1| \neq |V_2|$, then $\mathbf{\Pi}$ is a non-square matrix which does not affect our analysis and algorithm design). We denote $\mathbf{\Pi_0}(\pi_0)$ as the true permutation matrix (mapping) between G_1 and G_2. We do not have any prior knowledge of $\mathbf{\Pi_0}$ and access to the underlying graph G. We formally define the social network de-anonymization problem in Definition 4 along with an illustrative instance in Fig. 1.

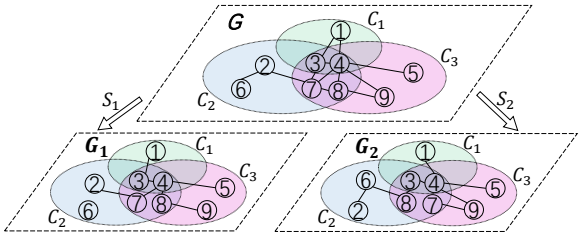


Fig. 1: An example of G, G_1 and G_2. The edges of $G_{1(2)}$ are sampled independently from G with probability $s_{1(2)}$. C_1, C_2, C_3 denote 3 different communities in OSBM. The true mapping $\pi_0 = \{(1, 1), (2, 6), (3, 3), (4, 4), (5, 5), (6, 2), (7, 8), (8, 7), (9, 9)\}$.

Definition 4. ***(Social Network De-anonymization Problem)*** *Given the published network G_1, the auxiliary network G_2, parameter set $\boldsymbol{\theta}$, social network de-anonymization problem aims to construct the true mapping π_0 between V_1 and V_2.*

However, our estimated permutation, $\hat{\mathbf{\Pi}}$, may deviate from the ground-truth $\mathbf{\Pi_0}$. To quantify this difference, we introduce a metric called "node mapping error (NME)" as follows.

Definition 5. ***(Node Mapping Error)*** *Given the estimated $\tilde{\mathbf{\Pi}}$ and ground-truth $\mathbf{\Pi_0}$, the node mapping error (NME) between $\hat{\mathbf{\Pi}}$ and $\mathbf{\Pi_0}$ is defined as $d(\hat{\mathbf{\Pi}}, \mathbf{\Pi_0}) = \frac{1}{2} ||\hat{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2$.*

Obviously $d(\hat{\mathbf{\Pi}}, \mathbf{\Pi_0})$ equals to 0 if and only if two permutations are identical, and if k nodes are mapped mistakenly, then NME equals to k, showing that NME is well-defined. Thus the goal of de-anonymization is to minimize NME.

Moreover, since adversaries is uncertain about the true mapping between the given G_1 and G_2, $\mathbf{\Pi_0}$ can be viewed as a random variable whose probability distribution is conditioned on G_1 and G_2 in adversaries' perspectives. Naturally adversaries prefer an estimation of $\mathbf{\Pi_0}$ keeping from severe NME on average. To this end, we consider selecting $\hat{\mathbf{\Pi}}$ in the light of "Minimum Mean Square Error (MMSE)" criterion, which, formally presented in Definition 6, is the minimizer of the expected NME in the form of mean square.

Definition 6. ***(The MMSE Estimator)*** *Given G_1, G_2 and $\boldsymbol{\theta}$, the MMSE estimator is an estimation of $\mathbf{\Pi_0}$ minimizing the*

TABLE I: **Notions and Definitions**

Notation	Definition
G	Underlying social network
G_1, G_2	Published and auxiliary networks
V, V_1, V_2	Vertex sets of graphs G, G_1 and G_2
E, E_1, E_2	Edge sets of graphs G, G_1, G_2
s_1, s_2	Edge sampling probabilities of graphs G_1, G_2
n	Total number of nodes
w_{ij}	The weight of node pair (i, j)
$\boldsymbol{C_i}$	Community representation of node i
$p_{\boldsymbol{C_i C_j}}$	Probability of edge existence between node i and j in G
$\boldsymbol{\theta}$	Parameter set
$\mathbf{W}$	The weight matrix
$\mathbf{U}, \mathbf{A}, \mathbf{B}$	Adjacency matrices of G, G_1, G_2
$\mathbf{\Pi_0}(\pi_0)$	True permutation matrix (True mapping) between V_1 and V_2
$\mathbf{\Pi}(\pi)$	A permutation matrix (A mapping) between V_1 and V_2
$\hat{\mathbf{\Pi}}(\hat{\pi})$	The MMSE estimator (the corresponding mapping)
$\tilde{\mathbf{\Pi}}(\tilde{\pi})$	The minimizer of WEMP (the corresponding mapping)
$\mathbf{\Pi^n}$	The set of $n \times n$ permutation matrices.

number of mistakenly matched nodes in expectation, which is

$$\begin{aligned}\hat{\mathbf{\Pi}} &= \arg\min_{\mathbf{\Pi}\in\mathbf{\Pi^n}} \mathbf{E}_{\mathbf{\Pi_0}}\{d(\mathbf{\Pi}, \mathbf{\Pi_0})\} \\ &= \arg\min_{\mathbf{\Pi}\in\mathbf{\Pi^n}} \sum_{\mathbf{\Pi_0}\in\mathbf{\Pi^n}} ||\mathbf{\Pi} - \mathbf{\Pi_0}||_F^2 Pr(\mathbf{\Pi_0}|G_1, G_2, \boldsymbol{\theta}),\end{aligned} \tag{1}$$

where $\mathbf{\Pi^n}$ is the set of $n \times n$ permutation matrices.

Remark: Recall that prior effort [8] has leveraged Maximum A Posterior (MAP), which provides the solution with the highest probability being exactly identical to the true permutation. MMSE and MAP characterize different aspects of minimizing NME. As far as we know, no previous work has learned de-anonymization under MMSE, which, however, is also of great significance as MAP in reducing NME.

The main notations in our work are summarized in Table 1.

IV. Analytical Aspect

In this section, we rewrite the MMSE problem in an equivalent but more explicit form, which is then proved to be NP-hard. To facilitate the problem analysis, we give an approximation to the MMSE problem and discuss its validity.

A. Reforming MMSE Estimator

Note that $Pr(\mathbf{\Pi_0}|G_1, G_2, \boldsymbol{\theta})$ in Eqn. (1) needs to be expressed more explicitly. Inspired by the derivation in [8], we have the following theorem reforming MMSE estimator.

Theorem 1. *Given G_1, G_2 and $\boldsymbol{\theta}$, the MMSE estimator can be equivalently reformed as*

$$\hat{\mathbf{\Pi}} = \arg\max_{\mathbf{\Pi}\in\mathbf{\Pi^n}} \sum_{\mathbf{\Pi_0}\in\mathbf{\Pi^n}} ||\mathbf{\Pi} - \mathbf{\Pi_0}||_F^2 ||\mathbf{W} \circ (\mathbf{\Pi_0 A} - \mathbf{B\Pi_0})||_F^2, \tag{2}$$

where $\mathbf{W}$ satisfies that $\mathbf{W}(i, j) = \sqrt{w_{ij}}$ and $w_{ij} = \log\left(\frac{1-p_{\boldsymbol{C_i C_j}}(s_1+s_2-s_1s_2)}{p_{\boldsymbol{C_i C_j}}(1-s_1)(1-s_2)}\right)$ is weight between nodes i and j.

Proof: Here we present an sketch of our proof, which is similar to Appendix D in [22]. Define $\mathcal{G}_{\mathbf{\Pi}}$ as the set of all realizations of the underlying network which is in consistency with the given G_1, G_2 and $\mathbf{\Pi}$. Then the MMSE estimator can be written as

$$\hat{\mathbf{\Pi}} = \arg\min_{\mathbf{\Pi}\in\mathbf{\Pi^n}} \sum_{\mathbf{\Pi_0}\in\mathbf{\Pi^n}} ||\mathbf{\Pi} - \mathbf{\Pi_0}||_F^2 \sum_{G\in\mathcal{G}_{\mathbf{\Pi}}} Pr(G, \mathbf{\Pi_0}|G_1, G_2, \boldsymbol{\theta}). \tag{3}$$

Focusing on $Pr(G, \mathbf{\Pi_0}|G_1, G_2, \boldsymbol{\theta})$ in Eqn. (2). By Bayesian's formula along with the independency of the sampling process of G_1 and G_2, we obtain $Pr(G, \mathbf{\Pi_0}|G_1, G_2, \boldsymbol{\theta}) = (Pr(G)Pr(G_1|G)Pr(G_2|G, \mathbf{\Pi_0}))/Pr(G_1, G_2)$. Then we explicitly express $Pr(G)$, $Pr(G_1|G)$, $Pr(G_2|G, \mathbf{\Pi_0})$ by virtue of defining a graph G_π^* with the least number of edges among all graphs in $G_{\mathbf{\Pi}}$. Ultimately we transform Eqn. (3) into

$$\hat{\mathbf{\Pi}} = \arg\max_{\mathbf{\Pi}\in\mathbf{\Pi^n}} \sum_{\mathbf{\Pi_0}\in\mathbf{\Pi}^n} ||\mathbf{\Pi} - \mathbf{\Pi_0}||_F^2 \sum_{G\in\mathcal{G}_{\mathbf{\Pi}}} ||\mathbf{W} \circ (\mathbf{\Pi_0 A} - \mathbf{B\Pi_0})||_F^2,$$

where $\mathbf{W}$ is the symmetric weight matrix such that $\mathbf{W}(i, j) = \sqrt{w_{ij}} = \mathbf{W}(j, i)$ and $w_{ij} = \log\left(\frac{1-p_{\boldsymbol{C_i C_j}}(s_1+s_2-s_1s_2)}{p_{\boldsymbol{C_i C_j}}(1-s_1)(1-s_2)}\right)$ is weight between nodes i and j. Proved. ■

Remark: To simplify the form of $||\mathbf{W} \circ (\mathbf{\Pi_0 A} - \mathbf{B\Pi_0})||_F^2$, we set $\mathbf{\Pi_0}\hat{\mathbf{A}} = \mathbf{W} \circ \mathbf{\Pi_0 A}$, and $\hat{\mathbf{B}}\mathbf{\Pi_0} = \mathbf{W} \circ \mathbf{B\Pi_0}$. Therefore we can rewrite the MMSE estimator as

$$\hat{\mathbf{\Pi}} = \arg\max_{\mathbf{\Pi}\in\mathbf{\Pi^n}} \sum_{\mathbf{\Pi_0}\in\mathbf{\Pi^n}} ||\mathbf{\Pi} - \mathbf{\Pi_0}||_F^2 ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2, \tag{4}$$

We discuss MMSE estimator in the form of Eqn. (4) which will be the object of interests in our subsequent analysis throughout the paper. In Section V-A, we will discuss when $\mathbf{W} \circ \mathbf{A} = \hat{\mathbf{A}}$ and $\mathbf{W} \circ \mathbf{B} = \hat{\mathbf{B}}$.

B. NP-hardness of Solving the MMSE Estimator

We prove the NP-hardness of MMSE problem, showing no polynomial time approximation algorithm for it.

Proposition 1. *Solving the MMSE estimator is an NP-hard problem. There is no polynomial time approximation algorithm with any multiplicative approximation guarantee unless P=NP.*

Proof: Due to the limitation of space, we provide an outline of our proof. Generally, the main idea to demonstrate the NP-hardness of MMSE problem is that: *We reduce the* 1*-median problem*[1] *to MMSE problem, and demonstrate that when the size of* 1*-median problem is identical to MMSE problem (which is n! since we need to calculate all $\mathbf{\Pi_0} \in \mathbf{\Pi^n}$), then the lower bound of time complexity is larger than polynomial.*

Reduction from 1-median problem: We construct a clique with $n!$ nodes, each node i representing a possible $\mathbf{\Pi_0}(i) \in \mathbf{\Pi^n}$. We modify Eqn. (4) equivalently into $\hat{\mathbf{\Pi}} = \arg\min_{\mathbf{\Pi}\in\mathbf{\Pi}^n} \sum_{\mathbf{\Pi_0}\in\mathbf{\Pi^n}} (4n - ||\mathbf{\Pi} - \mathbf{\Pi_0}||_F^2) ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2$ where we set $D(i, j) = 4n - ||\mathbf{\Pi_0}(i) - \mathbf{\Pi_0}(j)||_F^2$ and $\omega(i) = ||\mathbf{\Pi_0}(i)\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}(i)||_F^2$ (Note that $\mathbf{\Pi_0}(i)$ is a node in the clique). Both $D(i, j)$ and $\omega(i)$ meet the requirements in 1-median problem , thus we complete the reduction.

The Lower Bound for 1-Median Problem: For a 1-median problem with $n!$ nodes, obviously we need to calculate at least $\lceil \frac{n!}{2} \rceil$ edges to form an edge set such that the endpoints of all edges inside cover all nodes in the graph, or else at least one node will not be calculated for any edge connecting it, which we can not judge if it is the node we intend to find. Since the

[1]The 1-median problem [20] refers to that: Given a connected undirected graph $G = (V, E)$ in which no isolated vertices exist and each node v is endowed with a nonnegative weight $\omega(v)$, find the vertex v^* which minimizes weighted sum. $H(v^*) = \sum_{v\in V} \omega(v) \cdot D(v, v^*)$ where $D(v, v^*)$ means the shortest path length between nodes v and v^*. Note that 1-median itself is not NP-hard if the problem size is O(n).

size of our input, a matrix, is n^2, the complexity turns out to be $\Omega((\sqrt{n}/2)!) = \Omega(\sqrt{n}!)$, exceeding polynomial. ■

C. Approximation of the MMSE estimator

MMSE involves all the $n!$ possible true mappings, leading to fairly prohibitive computational cost. To tackle the difficulty, we validly transform the MMSE problem into a weighted-edge matching problem (WEMP), which ensures tractability by eliminating the need for calculating all $n!$ possible cases. Definition 7 formally formulates WEMP.

Definition 7. ***(Weighted-Edge Matching Problem)*** *Given $G_1(V_1, E_1, \mathbf{A})$, $G_2(V_2, E_2, \mathbf{B})$ and weight matrix $\mathbf{W}$, the weighted-edge matching problem is to find $\tilde{\mathbf{\Pi}} = \arg\min_{\mathbf{\Pi}\in\mathbf{\Pi^n}} ||\mathbf{\Pi}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi}||_F^2$.*

The idea of this transformation is that: For any fixed $\mathbf{\Pi}$, we define a set $S_k(\mathbf{\Pi}), 0 \le k \le n$, any element of which is $\mathbf{\Pi_0}$ such that $d(\mathbf{\Pi}, \mathbf{\Pi_0}) = k$. Obviously $S_0(\mathbf{\Pi}) = \{\mathbf{\Pi}\}$, $S_1(\mathbf{\Pi}) = \emptyset$. Then we can reform MMSE problem as

$$\hat{\mathbf{\Pi}} = \arg\max_{\mathbf{\Pi}\in\mathbf{\Pi^n}} \sum_{k=0}^{n} k \left(\sum_{\mathbf{\Pi_0}\in S_k(\mathbf{\Pi})} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 \right). \quad (5)$$

From Eqn. (5) we transform MMSE problem to WEMP by three steps: (i) Analyzing the error caused by $\mathbf{\Pi_0} \in S_2(\mathbf{\Pi})$, the simplest case where only two nodes are mapped erroneously; (ii) Extending the basic analysis on $\mathbf{\Pi_0} \in S_2(\mathbf{\Pi})$ to $\mathbf{\Pi_0} \in S_k(\mathbf{\Pi}), \forall k > 2$; (iii) Finding that WEMP coincides MMSE estimator under Sequence Inequality [19]. It is proved to be valid under average case. Meanwhile, for a specific network, we also show that the validity of this transformation can be ensured. Due to the space limitation we omit the proof here.

V. Algorithmic Aspect

In this section, we show that WEMP is of significant advantages in seedless de-anonymization since it resolves the tension between *optimality* and *complexity*. For optimality, We prove the good performance of solving WEMP that the result makes the node mapping error (NME) negligible in large social networks under mild conditions, facilitated by higher overlapping strength; For complexity, the optimal mapping of WEMP, $\tilde{\mathbf{\Pi}}$, can be perfectly sought algorithmically by our convex-concave based de-anonymization algorithm (CBDA).

A. Optimality: WEMP Returns Negligible NME

Recall that our aim is to minimize NME in expectation, thus a natural question arises: *how much NME $\tilde{\mathbf{\Pi}}$ may cause for any probable real permutation matrix $\mathbf{\Pi_0}$?* The answer reflects the ability of solving WEMP in enhancing mapping accuracy. To answer it, we demonstrate that under mild conditions, the *relative NME*, defined as $\frac{||\tilde{\mathbf{\Pi}}-\mathbf{\Pi_0}||_F^2}{||\mathbf{\Pi_0}||_F^2}$, vanishes to 0 as $n \to \infty$. This implies that under large network size, NME caused by $\tilde{\mathbf{\Pi}}$ is negligible compared with $|V| = n$. Furthermore, we surprisingly find that the conditions are facilitated under higher overlapping strength, explicitly delineating benefits brought by overlapping communities in NME reduction. Theorem 2 formally presents our result mentioned above. Before that, we give Lemma 1.

Lemma 1. *[22]Suppose the permutation matrix $\mathbf{\Pi}$ keeps invariant of the community representation of all the nodes, i.e., $\forall \mathbf{\Pi} \in \mathbf{\Pi^n}$ such that $\mathbf{\Pi}(i,j) = 1$, $\boldsymbol{C_i} = \boldsymbol{C_j}$, then $\hat{\mathbf{A}} = \mathbf{W}\circ\mathbf{A}$, $\hat{\mathbf{B}} = \mathbf{W}\circ\mathbf{B}$ and $||\mathbf{W}\circ(\mathbf{\Pi A \Pi^T} - \mathbf{B})||_F = ||\mathbf{\Pi}\hat{\mathbf{A}}\mathbf{\Pi^T} - \hat{\mathbf{B}}||_F$.*

Remark: Note that there are no differences in form between $||\mathbf{\Pi_1}\hat{\mathbf{A}}\mathbf{\Pi_1^T} - \hat{\mathbf{B}}||_F$ and $||\hat{\mathbf{A}} - \mathbf{\Pi_2}\hat{\mathbf{B}}\mathbf{\Pi_2^T}||_F$ since we can simply set $\mathbf{\Pi_2} = \mathbf{\Pi_1^T}$. Therefore, we do not distinguish the forms $||\mathbf{\Pi}\hat{\mathbf{A}}\mathbf{\Pi^T} - \hat{\mathbf{B}}||_F$ and $||\hat{\mathbf{A}} - \mathbf{\Pi}\hat{\mathbf{B}}\mathbf{\Pi^T}||_F$ anymore.

Theorem 2. *Given $G_1(V_1, E_1, \mathbf{A})$, $G_2(V_2, E_2, \mathbf{B})$, $\boldsymbol{\theta}$ and $\mathbf{W}$. Set $\tilde{p}_{\boldsymbol{C_i}\boldsymbol{C_j}} = w_{ij} p_{\boldsymbol{C_i}\boldsymbol{C_j}}$ and*

$$\begin{aligned} K &= \min_{s,t,j}\{(\tilde{p}_{\boldsymbol{C_s}\boldsymbol{C_j}} + \tilde{p}_{\boldsymbol{C_t}\boldsymbol{C_j}})\min\{s_1, s_2\}\}, \\ L &= \max_{s,t,j}\{[(\tilde{p}_{\boldsymbol{C_s}\boldsymbol{C_j}} + \tilde{p}_{\boldsymbol{C_t}\boldsymbol{C_j}})\max\{s_1, s_2\}]^2\}. \end{aligned} \quad (6)$$

If

- *(i) $\frac{L}{K} = o(1)$;*
- *(ii) $\frac{||\hat{\mathbf{A}}-\mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0^T}||_F^2}{||\hat{\mathbf{A}}-\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}||_F^2} = \Omega(1)$;*
- *(iii) $||\hat{\mathbf{A}} - \mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0^T}||_F^2 = o(Kn^2)$;*
- *(iv) $\mathbf{\Pi_0}$ and $\tilde{\mathbf{\Pi}}$ keep invariant of community representations,*

then as $n \to \infty$, $\frac{||\tilde{\mathbf{\Pi}}-\mathbf{\Pi_0}||_F^2}{||\mathbf{\Pi_0}||_F^2} \to 0$.

Proof: The proof has four steps: (i) Upper bounding $||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F$ by $||(\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}||_F$; (ii) Finding the relationship between $||(\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}||_F$ and $\mathbf{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}}\hat{\mathbf{A}})$.; (iii) Upper bounding $\mathbf{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}}\hat{\mathbf{A}})$; (iv) Upper bounding $\frac{||\mathbf{\Pi_0}-\tilde{\mathbf{\Pi}}||_F^2}{||\mathbf{\Pi_0}||_F^2}$.

1. Upper bounding $||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F$ by $||(\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}||_F$:

For the i_{th} row of $(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})$, we set $\pi_0(i) = s$ and $\tilde{\pi}(i) = t$. If $s = t$, then the i_{th} row of $(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}$ is a zero vector; else the i_{th} row of $(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}$ is $(\hat{\mathbf{B}}_{s1} - \hat{\mathbf{B}}_{t1}, \hat{\mathbf{B}}_{s2} - \hat{\mathbf{B}}_{t2}, \cdots, \hat{\mathbf{B}}_{sn} - \hat{\mathbf{B}}_{tn})$. For an element, $([(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}]_{ij})^2 = (\hat{\mathbf{B}}_{sj} - \hat{\mathbf{B}}_{tj})^2 = (\sqrt{w_{sj}}\mathbf{B}_{sj} - \sqrt{w_{tj}}\mathbf{B}_{tj})^2$. Taking the expectation on both sides, we can derive that $\mathbf{E}[(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}]_{ij}^2 = \mathbf{E}(\hat{\mathbf{B}}_{sj} - \hat{\mathbf{B}}_{tj})^2 = (\tilde{p}_{\boldsymbol{C_s}\boldsymbol{C_j}} + \tilde{p}_{\boldsymbol{C_t}\boldsymbol{C_j}} - 2\sqrt{w_{sj}w_{tj}}p_{\boldsymbol{C_s}\boldsymbol{C_j}}p_{\boldsymbol{C_t}\boldsymbol{C_j}}s_2)s_2 \sim (\tilde{p}_{\boldsymbol{C_s}\boldsymbol{C_j}} + \tilde{p}_{\boldsymbol{C_t}\boldsymbol{C_j}})s_2$. where $\mathbf{E_B}$ means taking expectation on every element in $\mathbf{B}$. By summing up all the columns, we have $\mathbf{E}\sum_{j=1}^{n}[(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}]_{ij}^2 = \sum_{j=1}^{n}(\tilde{p}_{\boldsymbol{C_s}\boldsymbol{C_j}} + \tilde{p}_{\boldsymbol{C_t}\boldsymbol{C_j}})s_2$. By summing up all the rows and columns,

$$\begin{aligned} ||(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}||_F^2 &= \mathbf{E}\sum_{i=1}^{n}\sum_{j=1}^{n}[(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}]_{ij}^2 \\ &= \sum_{i=1}^{n}\mathbb{1}\{\pi_0(i) \neq \tilde{\pi}(i)\}\sum_{j=1}^{n}(\tilde{p}_{\boldsymbol{C_s}\boldsymbol{C_j}} + \tilde{p}_{\boldsymbol{C_t}\boldsymbol{C_j}})s_2 \\ &\ge \sum_{i=1}^{n} n\mathbb{1}\{\pi_0(i) \neq \tilde{\pi}(i)\}\min_j(\tilde{p}_{\boldsymbol{C_s}\boldsymbol{C_j}} + \tilde{p}_{\boldsymbol{C_t}\boldsymbol{C_j}})s_2, \end{aligned}$$

Note that $||(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})||_F^2 = 2\sum_{i=1}^{n}\mathbf{1}\{\pi_0(i) \neq \tilde{\pi}(i)\}$. Setting $K = \min_{s,t,j}(\tilde{p}_{\boldsymbol{C_s}\boldsymbol{C_j}} + \tilde{p}_{\boldsymbol{C_t}\boldsymbol{C_j}})s_2$, we have

$$||\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}}||_F^2 \le \frac{2}{nK}||(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}||_F^2. \quad (7)$$

Similarly we can replace $\hat{\mathbf{B}}$ by $\hat{\mathbf{A}}$ and change s_2 to s_1 in K.

2. $||(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}||_F$ and $\mathbf{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}})\hat{\mathbf{A}})$:

Note that $||(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}||_F \le ||(\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi_0^T} - \hat{\mathbf{A}}) - (\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}} -$

$\hat{\mathbf{A}})||_F \leq ||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}} - \hat{\mathbf{A}}||_F + ||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}} - \hat{\mathbf{A}}||_F$. Then $||(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}||_F^2 \leq 2(||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}} - \hat{\mathbf{A}}||_F^2 + ||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}} - \hat{\mathbf{A}}||_F^2)$. For the term $||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}} - \hat{\mathbf{A}}||_F^2$,

$$\begin{aligned} &||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}} - \hat{\mathbf{A}}||_F^2 = ||\hat{\mathbf{A}}||_F^2 + ||\hat{\mathbf{B}}||_F^2 - 2\mathbf{tr}(\mathbf{\Pi_0}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}\hat{\mathbf{A}}) \\ &= \frac{1}{2}(||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}} - \hat{\mathbf{A}}||_F^2 + ||\mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}} - \hat{\mathbf{A}}||_F^2) \\ &\quad + \mathbf{tr}(\mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}}\hat{\mathbf{A}}) + \mathbf{tr}(\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}\hat{\mathbf{A}}) - 2\mathbf{tr}(\mathbf{\Pi_0}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}\hat{\mathbf{A}}) \\ &\leq ||\mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}} - \hat{\mathbf{A}}||_F^2 + \mathbf{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}}\hat{\mathbf{A}}), \end{aligned} \tag{8}$$

3. Upper Bound of $\mathbf{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}}\hat{\mathbf{A}})$:

Set $\mathbf{Z} = (\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}}\hat{\mathbf{A}}$. For simplicity, we define $\mathbf{Y} = (\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}$ and $\mathbf{X} = (\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}}\hat{\mathbf{A}}$, thus $\mathbf{Z} = \mathbf{YX}$. We focus on $\mathbf{tr}(\mathbf{Z})$. It is easy to verify that for any node i, when $\tilde{\mathbf{\Pi}}$ and $\mathbf{\Pi_0}$ map it to the same node, then $\mathbf{Z}_{ii} = 0$. If not, for node i we assume that $\tilde{\mathbf{\Pi}}$ maps it to s and $\mathbf{\Pi_0}$ maps it to t, where $s \neq t$. We can obtain the i_{th} row of $\mathbf{Y}$ as $\mathbf{Y_{i\cdot}} = (\hat{\mathbf{B}}_{s1} - \hat{\mathbf{B}}_{t1}, \hat{\mathbf{B}}_{s2} - \hat{\mathbf{B}}_{t2}, \cdots, \hat{\mathbf{B}}_{sn} - \hat{\mathbf{B}}_{tn})$. Similarly, we can obtain the i_{th} column of $\mathbf{X}$ as $\mathbf{X_{\cdot i}} = (\hat{\mathbf{A}}_{p_1 1} - \hat{\mathbf{A}}_{q_1 1}, \hat{\mathbf{A}}_{p_2 2} - \hat{\mathbf{A}}_{q_2 2}, \cdots, \hat{\mathbf{A}}_{p_n n} - \hat{\mathbf{A}}_{q_n n})^{\mathbf{T}}$, where $p_i(q_i)$ means the row index of the $1(-1)$ in the i_{th} column of $\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}$. If $\pi_0(j) = \tilde{\pi}(j)$, we simply set $\mathbf{X}_{ji} = 0$. Therefore $\mathbf{Z}_{ii}$, an element on diagonal of $\mathbf{Z}$, satisfies

$$\begin{aligned} |\mathbf{Z}_{ii}| &= |\langle \mathbf{Y_{i\cdot}}, \mathbf{X_{\cdot i}} \rangle| \leq ||\mathbf{Y_{i\cdot}}||_F ||\mathbf{X_{\cdot i}}||_F \\ &\leq n \max_k |\hat{\mathbf{B}}_{sk} - \hat{\mathbf{B}}_{tk}| \max_\ell |\hat{\mathbf{A}}_{p_\ell \ell} - \hat{\mathbf{A}}_{q_\ell \ell}|. \end{aligned} \tag{9}$$

then $|\mathbf{Z}_{ii}| \leq n$. Taking the expectation of $\mathbf{A}$ and $\mathbf{B}$ on both sides of Inequality (9), we can obtain that

$$\begin{aligned} \mathbf{E}_{\mathbf{A},\mathbf{B}}|\mathbf{Z}_{ii}| &= \mathbf{E}_{\mathbf{A},\mathbf{B}}(\max_{s,t,k} |\hat{\mathbf{B}}_{sk} - \hat{\mathbf{B}}_{tk}| \max_{p,q,\ell} |\hat{\mathbf{A}}_{p_\ell \ell} - \hat{\mathbf{A}}_{q_\ell \ell}|) \\ &\leq \max_{p,q,\ell} \{[(\tilde{p}_{\boldsymbol{C_s C_j}} + \tilde{p}_{\boldsymbol{C_t C_j}}) \max\{s_1, s_2\}]^2\} = L, \end{aligned}$$

which is based on Jensen's Inequality. Hence

$$|\mathbf{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}})\hat{\mathbf{A}})| \leq n \max_i |\langle \mathbf{Y_{i\cdot}}, \mathbf{X_{\cdot i}} \rangle| \leq n^2 L. \tag{10}$$

4. Upper Bound of $\frac{||\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}}||_F^2}{||\mathbf{\Pi_0}||_F^2}$:

From Inequalities (7), (8) and (10), we can obtain $||\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}}||_F^2 \leq \frac{2}{nK}||(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}||_F^2 \leq \frac{8}{nK}||\mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}} - \hat{\mathbf{A}}||_F^2 + \frac{4nL}{K}$. Since condition 2 holds, there exists a constant $\tilde{c} \geq 1$ such that $||\hat{\mathbf{A}} - \tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}||_F \leq \tilde{c}||\hat{\mathbf{A}} - \mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}}||_F$. Therefore since $||\mathbf{\Pi_0}||_F^2 = 2n$ and conditions 1 and 3, when $n \to \infty$,

$$\frac{||\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}}||_F^2}{||\mathbf{\Pi_0}||_F^2} \leq \frac{4\tilde{c}}{n^2 K}||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}} - \hat{\mathbf{A}}||_F^2 + \frac{2L}{K} \to 0.$$

This completes our proof. ■

Overlapping Communities Benefits De-anonymization: Now we show that overlapping communities positively impact on reducing relative NME through facilitating conditions in Theorem 2, specifically condition 3. For convenience, we assume $s = s_1 = s_2$. When π_0 keeps invariant of community representations, then on average condition 3 can be written as $2\sum_{1 \leq i < j \leq n} p_{\boldsymbol{C_i C_j}} s \log\left(\frac{1 - p_{\boldsymbol{C_i C_j}}(2s - s^2)}{p_{\boldsymbol{C_i C_j}}(1-s)^2}\right) = o(Kn^2)$. To characterize the global situation in the networks, we define an *average probability* $\hat{p}$ such that $\sum_{1 \leq i < j \leq n} p_{\boldsymbol{C_i C_j}} s \log\left(\frac{1 - p_{\boldsymbol{C_i C_j}}(2s - s^2)}{p_{\boldsymbol{C_i C_j}}(1-s)^2}\right) = \frac{n(n-1)}{2} \log\left(\frac{1 - \hat{p}(2s - s^2)}{\hat{p}(1-s)^2}\right) \hat{p}s$, where $\hat{p}$ is positively correlated to the overlapping strength of the whole networks. Taking the derivative of $\hat{p}$ over $\log\left(\frac{1 - \hat{p}(2s - s^2)}{\hat{p}(1-s)^2}\right) \hat{p}s$, we find that $\frac{d(\log\left(\frac{1 - \hat{p}(2s - s^2)}{\hat{p}(1-s)^2}\right)\hat{p}s)}{d\hat{p}} = \log\left(\frac{1 - \hat{p}(2s - s^2)}{\hat{p}(1-s)^2}\right) s - \frac{1}{1 - \hat{p}(2s - s^2)}$, and it is easy to verify that $\frac{d\left(\log\left(\frac{1 - \hat{p}(2s - s^2)}{\hat{p}(1-s)^2}\right)\hat{p}s\right)}{d\hat{p}}$ is a decreasing function in terms of $\hat{p}$. Now focus on $\frac{d\left(\log\left(\frac{1 - \hat{p}(2s - s^2)}{\hat{p}(1-s)^2}\right)\hat{p}s\right)}{d\hat{p}}$. If we consider dense communities such that $\hat{p} = 1 - o(1)$, which means that $\hat{p}$ asymptotically approaches 1 (shown to be right under the Overlapping Stochastic Block Model(OSBM) below), then we can derive $\log\left(\frac{1 - \hat{p}(2s - s^2)}{\hat{p}(1-s)^2}\right) \hat{p}s = \log\left(1 + \frac{1 - \hat{p}}{\hat{p}(1-s)^2}\right) \hat{p}s \sim \frac{1 - \hat{p}}{(1-s)^2} s = o(1)$, where $s = \Omega(1)$. Therefore if $\hat{p}$ is asymptotically close to 1 as the overlapping strength enhances, then the order of $||\hat{\mathbf{A}} - \mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}}||_F^2$ turns smaller, which is more prone to satisfy $||\hat{\mathbf{A}} - \mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}}||_F^2 = o(Kn^2)$.

Taking a vivid example of the proposed OSBM [18] in which $p_{\boldsymbol{C_i C_j}} = \frac{1}{1 + ae^{-x}}$, where a is an adjustable parameter and x is the number of overlapping communities. We find that $\min_{i,j} p_{\boldsymbol{C_i C_j}} = \frac{1}{1+a}$ is a constant if $a = \Omega(1)$, and can be arbitrarily close to 1 when x is large enough. So if $s = o(1)$ and $\hat{p} = 1 - o(1)$, which means that the overlapping strength is very large, then $\hat{p}\log(\frac{1 - \hat{p}(2s - s^2)}{\hat{p}(1-s)^2}) = \hat{p}\log(1 + \frac{1 - \hat{p}}{\hat{p}(1-s)^2}) \approx \frac{1 - \hat{p}}{(1-s)^2} = o(1) = o(\min_{i,j} p_{\boldsymbol{C_i C_j}})$, thus condition (iii) holds. Meanwhile $s = o(1)$ makes condition (i) hold as well.

B. Complexity: WEMP can be Algorithmically Solved

Upon proving the good performance of solving WEMP in large-scale networks, now we algorithmically demonstrate that WEMP reduces the complexity of the MMSE problem since the optimal mapping of WEMP can be perfectly found by the convex-concave based de-anonymization algorithm (CBDA).

1) The Constraints of WEMP: We state WEMP as a constrained optimization problem: The objective function is $\|\hat{\mathbf{A}} - \mathbf{\Pi}\hat{\mathbf{B}}\mathbf{\Pi}^{\mathbf{T}}\|_{\mathrm{F}}^2$, with four constraints: (i) $\forall i \in V_1$, $\sum_i \mathbf{\Pi}_{ij} = 1$; (ii) $\forall j \in V_2$, $\sum_j \mathbf{\Pi}_{ij} = 1$; (iii) $\forall i, j$, $\mathbf{\Pi}_{ij} \in \{0, 1\}$ and (iv) $\forall i \in V_1, \boldsymbol{C_i} = \boldsymbol{C_{\pi(i)}}$. Constraints (i), (ii) and (iii) are the attributes of permutation matrices. Note that we append constraint (iv) in that our estimated mapping should keep the community representations as $\mathbf{\Pi_0}$, the true permutation we intend. To change it in the form of $\mathbf{\Pi}$ as constraints (i), (ii) and (iii), we define "Community Representation Matrix" to characterize the community representations of all the nodes.

Definition 8. ***(Community Representation Matrix)*** *Given a graph G with n nodes and m communities, the community representation matrix of G is an $n \times m$ matrix $\mathbf{M}$ which is composed of 0s and 1s, and $\forall i \in \{1, 2, \cdots, n\}$, the i_{th} row of $\mathbf{M}$ is the community representation of node i in G.*

Note that the community representation matrices for G, G_1 and G_2 are identical, hence constraint (iv) can be rewritten as $||\mathbf{\Pi M} - \mathbf{M}||_F^2 = 0$. We equivalently embed this constraint into the objective function as $F_0(\mathbf{\Pi}) = ||(\hat{\mathbf{A}} - \mathbf{\Pi}\hat{\mathbf{B}}\mathbf{\Pi}^{\mathbf{T}})||_{\mathrm{F}}^2 + \mu||\mathbf{\Pi M} - \mathbf{M}||_F^2$, where μ is a large enough parameter such that

when $F_0(\mathbf{\Pi})$ reaches its minimum, $||\mathbf{\Pi}\mathbf{M}-\mathbf{M}||_F^2$ is exactly 0, ensuring the invariance of community representations.

2) Problem Relaxation and Idea of Algorithm Design:

Problem Relaxation: WEMP is an integer program problem which cannot be solved efficiently. We relax the original feasible region of WEMP Ω_0 into Ω, which are respectively

$$\Omega_0 = \{\mathbf{\Pi}_{ij} \in \{0,1\} | \forall i,j, \textstyle\sum_i \mathbf{\Pi}_{ij} = 1 \ , \sum_j \mathbf{\Pi}_{ij} = 1\};$$

$$\Omega = \{\mathbf{\Pi}_{ij} \in [0,1] | \forall i,j, \textstyle\sum_i \mathbf{\Pi}_{ij} = 1 \ , \sum_j \mathbf{\Pi}_{ij} = 1\}.$$

After this relaxation the problem becomes tractable. However, a natural question arises: *How to obtain the solution of the original unrelaxed problem from that of the relaxed problem*?

Idea of Convex-Concave Relaxation Method: Note that the minimizer of a concave function must be at the boundary of the feasible region, coinciding that Ω_0, the original feasible set, is just the boundary of Ω. Therefore, a natural idea emerges: *We can modify the convex relaxed problem into a concave problem gradually.* Thus we apply the convex-concave optimization method (CCOM), whose concept is pioneeringly proposed in [21] to solve pattern matching problems: For $F_0(\mathbf{\Pi})$, we find its convex and concave relaxed version respectively $F_1(\mathbf{\Pi})$ and $F_2(\mathbf{\Pi})$. Then we obtain a new objective function as $F(\mathbf{\Pi}) = (1-\alpha)F_1(\mathbf{\Pi}) + \alpha F_2(\mathbf{\Pi})$. We modify α gradually from 0 to 1 with interval $\Delta\alpha$, each time solving the new $F(\mathbf{\Pi})$ initialized by the optimizer last time. $F(\mathbf{\Pi})$ becomes more concave, with its optimum closer to Ω_0 where $\tilde{\mathbf{\Pi}}$ lies.

3) Implementation of CCOM and Algorithm Design: The way to obtain $F_1(\mathbf{\Pi})$ and $F_2(\mathbf{\Pi})$ in [21] is rather complex. We provide a simple way to get them by Lemma 2.

Lemma 2. *A proper way to get F_1 and F_2 is $F_1(\mathbf{\Pi}) = F_0(\mathbf{\Pi}) + \frac{\lambda_{min}}{2}(n - ||\mathbf{\Pi}||_F^2)$; $F_2(\mathbf{\Pi}) = F_0(\mathbf{\Pi}) + \frac{\lambda_{max}}{2}(n - ||\mathbf{\Pi}||_F^2)$, where λ_{min} (λ_{max}) is the smallest (largest) eigenvalue of the Hessian matrix of $F_0(\mathbf{\Pi})$.*

Proof: First we verify that $F_1(\mathbf{\Pi})$ is a convex function. One of the sufficient and necessary condition for a function whose variable is matrix is convex is that the Hessian matrix of this function is positive semi-definite. The Hessian matrix of $F(\mathbf{\Pi})$ can be obtained by taking the second derivative over $\mathbf{\Pi}$ on $F(\mathbf{\Pi})$, we denote it as $\nabla^2 F(\mathbf{\Pi})$. Therefore we can obtain the Hessian matrix of $F_1(\mathbf{\Pi})$ by $\nabla^2 F_1(\mathbf{\Pi}) = \nabla^2 F_0(\mathbf{\Pi}) - \lambda_{min}\mathbf{I}$. where $\mathbf{I}$ is the identity matrix. Note that λ_{min} is the minimum eigenvalue of $\nabla^2 F_0(\mathbf{\Pi})$, therefore all the eigenvalues of $\nabla^2 F_0(\mathbf{\Pi}) - \lambda_{min}\mathbf{I}$ are equal to or larger than 0. Hence $\nabla^2 F_1(\mathbf{\Pi})$ is a nonnegative definite matrix and $F_1(\mathbf{\Pi})$ is a convex function. Similarly we can verify that $F_2(\mathbf{\Pi})$ is a concave function. ■

Lemma 2 presents a simple way to implement CCOM, by which we form our new objective function in CCOM as $F(\mathbf{\Pi}) = (1-\alpha)F_1(\mathbf{\Pi}) + \alpha F_2(\mathbf{\Pi}) = F_0(\mathbf{\Pi}) + 2\xi(n - ||\mathbf{\Pi}||_F^2)$, where $\xi = (1-\alpha)\lambda_{min} + \alpha\lambda_{max}$. Then we propose Algorithm 1, called *Convex-concave Based De-anonymization Algorithm (CBDA)*, as our main algorithm for WEMP.

Note that $F_0(\mathbf{\Pi})$ itself is convex in our problem, thus we can set ξ from 0 to an arbitrarily large number, which obviates the great complexity to calculate eigenvalues of Hessian matrices.

CBDA consists of an outer loop (lines 3 to 10) and an inner loop (lines 4 to 8). The outer loop modifies ξ in CCOM. The inner loop finds the minimizer of $F(\mathbf{\Pi})$, whose main idea resembles descending algorithms: In line 5, we obtain descending direction by minimizing $\mathbf{tr}(\nabla_{\mathbf{\Pi_k}} F(\mathbf{\Pi_k})^T \mathbf{X}^\perp)$, dangling the highest probability to find a descending direction characterized by $\mathbf{tr}(\nabla_{\mathbf{\Pi_k}} F(\mathbf{\Pi_k})^T \mathbf{X}^\perp) < 0$. In line 6 we search for step length γ_k contributing most to lowering $F(\mathbf{\Pi})$ on this descending direction. Line 7 is the update of estimation.

Algorithm 1: Convex-concave Based De-anonymization Algorithm (CBDA)

Input: Adjacent matrices $\mathbf{A}$ and $\mathbf{B}$; Community assignment matrix $\mathbf{M}$; Weight controlling parameter μ; Adjustable parameters δ, $\Delta\xi$.
Output: Estimated permutation matrix $\tilde{\mathbf{\Pi}}$.

1: Form the objective function $F_0(\mathbf{\Pi})$ and $F(\mathbf{\Pi})$.
2: $\xi \leftarrow 0$, $k \leftarrow 1$, Initialize $\mathbf{\Pi_1}$. Set ξ_m, the upper limit of ξ.
3: **while** $\xi < \xi_m$ and $\mathbf{\Pi_k} \notin \Omega_0$ **do**
4: **while** $k = 1$ or $|F(\mathbf{\Pi_{k+1}}) - F(\mathbf{\Pi_k})| \geq \delta$ **do**
5: $\mathbf{X}^\perp \leftarrow \arg\min_{\mathbf{X}^\perp} \mathbf{tr}(\nabla_{\mathbf{\Pi_k}} F(\mathbf{\Pi_k})^T \mathbf{X}^\perp)$, where $\mathbf{X}^\perp \in \Omega$.
6: $\gamma_k \leftarrow \arg\min_\gamma F(\mathbf{\Pi_k} + \gamma(\mathbf{X}^\perp - \mathbf{\Pi_k}))$, where $\gamma_k \in [0,1]$.
7: $\mathbf{\Pi_{k+1}} \leftarrow \mathbf{\Pi_k} + \gamma_k(\mathbf{X}^\perp - \mathbf{\Pi_k})$, $k \leftarrow k+1$.
8: **end while**
9: $\xi \leftarrow \xi + \Delta\xi$.
10: **end while**

4) Time Complexity and Convergence Analysis:

Time Complexity: The inner loop is similar to the Frank-Wolfe algorithm, with $O(n^6)$ in a round (since the input is an $n \times n$ matrix). If the maximum number of inner loops as T, thus the whole algorithm has a complexity of $O\big(\frac{n^6 T\xi}{\Delta\xi}\big)$. As far as we know, a dearth of algorithmic analysis of seedless de-anonymization exists except for [22], [23], with their proposed algorithm sharing identical complexity of $O(n^6)$ with ours.

Convergence: The inner loop is similar to the Frank-Wolfe algorithm, and if the step size of outer loop $\Delta\xi$ is small enough, then it is convergent.

VI. Experimental Aspect

In this section, we utilize three datasets, especially the rare true cross-domain co-author networks, to validate our theoretical results and performance of CBDA. Before presenting empirical results, we first introduce our experimental setup.

A. Experimental Setup

1) Main Parameters: We list the main parameters in validating the performance of our CDBA in Table II, where η is the ratio between the number of communities and nodes. η reflects more communities in networks with larger size.

TABLE II: **Main Experimental Parameters**

Notation	Definition	Range
N	Number of Nodes	{500, 1000, 1500, 2000}
s	Sampling Probability ($s_1 = s_2 = s$)	0.3-0.9
η	Community Ratio	{0.05, 0.1}
OL/NOL	Overlapping or Non-Overlapping	{OL, NOL}

2) Experimental Datasets: (i) Synthetic Networks: We generate networks by setting the community representation of every node independently and randomly deciding the edge existence in node pair (i,j) based on OSBM [18]. (ii) Sampled Real Social Networks: The underlying social network G is extracted from LiveJournal [17], while G_1

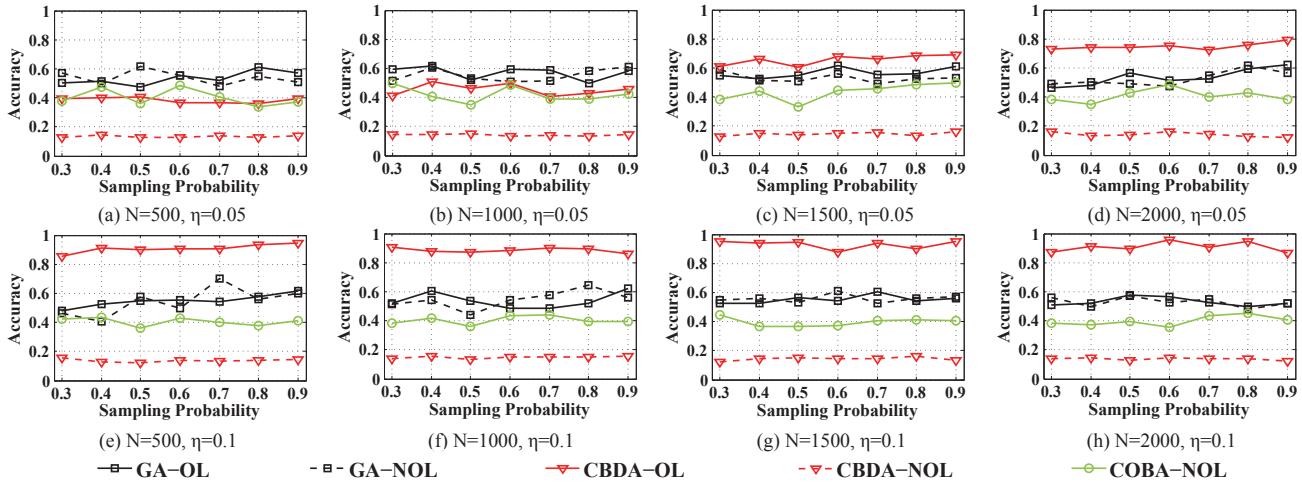


Fig. 2: Experiments on Synthetic Networks.

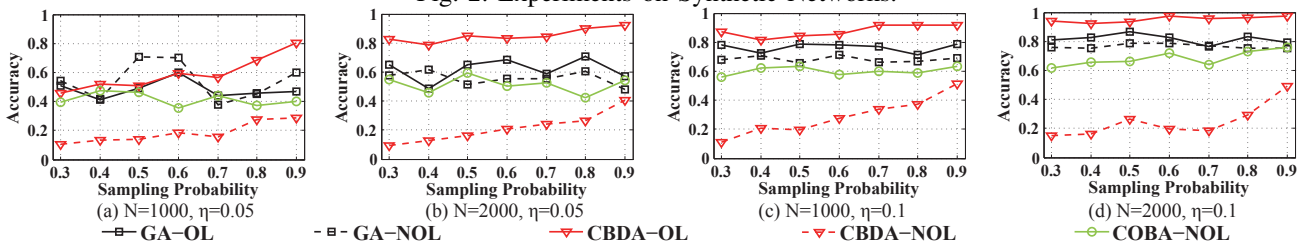


Fig. 3: Experiments on Sampled Real Social Networks.

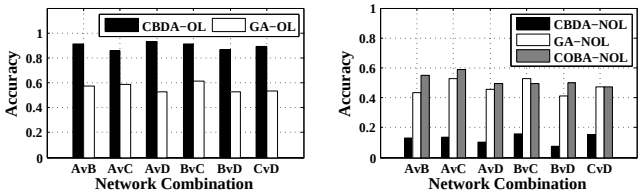
and G_2 are sampled from G with the same probability s. (iii) Cross-Domain Co-author Networks: The co-author networks are from the Microsoft Academic Graph (MAG) [11]. We extract 4 networks belonging to different sub-areas in the field of computer science, with the same group of authors, each of whom has a unique 8-bit hexadecimal ID enabling us to construct the true mapping between two networks as the one mapping nodes with same ID. Each network can be viewed as G_1 or G_2, thus there are $C_4^2 = 6$ combinations. (Table III) Note that we can assign w_{ij} on all these 3 datasets since the prior knowledge is just M, which can be generated or known from the real networks.

TABLE III: **Datasets in Basic Experiments**

Dataset	Synthetic	Sampled Real Social	Cross-Domain Co-author
Source	OSBM	LiveJournal [1]	MAG [11]
Num. of Nodes	$500 \sim 2000$	$500 \sim 2000$	3176
Num. of Communities	$25 \sim 1000$	$25 \sim 1000$	89

3) Algorithms for Comparison and Performance Metric: We exclude algorithms for seeded de-anonymization and select algorithms suitable for seedless cases related to our main point: showing the impact of overlapping communities on reducing NME, though other algorithms might outperform ours. We select two algorithms for comparison: (i) the Genetic Algorithm (GA), an epitome of heuristic algorithms; (ii) the Convex Optimization-Based Algorithm (COBA) in [22], [23], assigning a node to a unique community, which primarily suits non-overlapping cases. The performance metric is *accuracy*, the proportion of correctly mapped nodes.

4) Supplementary Experiments: To make our results more solid, based on sampled real social networks we study (i) the effect of η, reflecting overlapping strength, on the accuracy; (ii) the priority of our cost function with $\mathbf{W}$ derived from



(a) Overlapping (b) Non-Overlapping

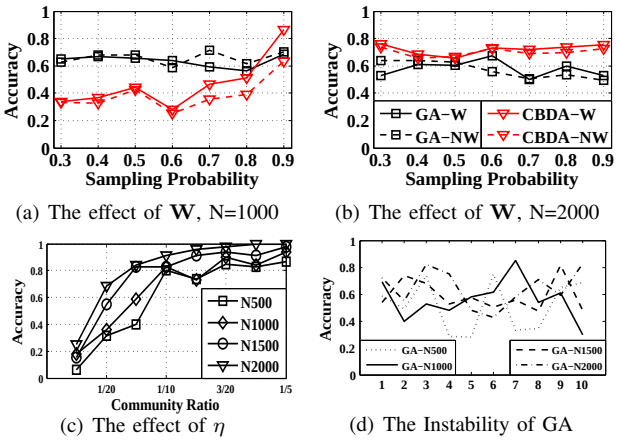
Fig. 4: Experiments on Cross-Domain Co-author Networks

MMSE makes for higher accuracy, comparing with the cost function without $\mathbf{W}$ in [3]; (iii) the instability of GA revealing its practical limitation *and thus in validation on* 3 *datasets we take the average preformance of GA* 10 *times as its accuracy.*

B. Experiment Results

1) Synthetic Networks: Fig. 2 illustrates the results on synthetic networks. We observe that the average performance of GA fluctuates from 40% to 60%, due to its heuristic search on local minimum without community information involved. The accuracy of CBDA rises up under larger network size N, in line with Theorem 2 that the relative NME shrinks as N mounts. However, in non-overlapping cases CBDA is inferior to COBA, since, COBA tackles non-overlapping property explicitly by assigning a node to a unique community while our CBDA utilizes it implicitly in optimizing $F(\mathbf{\Pi})$. Under denser communities ($\eta = 0.1$), our proposed CBDA always performs best in overlapping cases, with accuracy mildly swinging around 90%, propped by the facilitation of overlapping communities discussed at the end of Section V-A.

2) Sampled Real Social Networks: The results under sampled real social networks are plotted in Fig. 3. CBDA still



(a) The effect of $\mathbf{W}$, N=1000 (b) The effect of $\mathbf{W}$, N=2000

(c) The effect of η (d) The Instability of GA

Fig. 5: Supplementary Experiments

performs well under denser overlapping communities and larger network size, with the peak around 95% and the highest average level around 90% when $N = 2000$ and $\eta = 0.1$. Surprisingly, the performance of our CBDA is better than synthetic networks, which further undergirds the high practical applicability of CBDA. Additionally, the rising tendency of accuracy for larger N serves as a foil to the practicality of Theorem 2.

3) Cross-Domain Co-author Networks: Fig. 4 illustrates results on co-author networks. In non-overlapping cases, our CBDA does not perform well as GA and COBA, while in overlapping cases CBDA reaches accuracy around 90%, outstripping GA whose accuracy is averagely 60%. This phenomenon upgrades the significance of CBDA in de-anonymization in overlapping cases since the dataset is entirely realistic. Moreover, since overlapping cases are much more quotidian in real social networks, CBDA has wider usage than GA and COBA.

4) The Effect of η: The results are shown in Fig. 5(c). With larger η, CBDA works more accurately, accounted for by the facilitation of overlapping communities (Section V-A), which evinces its fitness for networks with high overlapping strength.

5) The Effect of Appending $\mathbf{W}$*:* As Fig. 5(a) and Fig. 5(b) show, CBDA works better appending $\mathbf{W}$ derived by MMSE, since the non-weighted cost function, adopted in [3], fails to distinguish nodes belonging to different number of communities. It shows the superiority of cost functions derived with rationale, as we claim in Section IV. Under larger network size, however, the difference becomes fainter since the impact of distinguishing a single node by w_{ij} is weaker than the benefits brought by large size shown in Theorem 2.

6) The Instability of GA: We disclose the instability of GA in Fig. 5(d). We run GA 10 times under sampled real social networks with different sizes. The performance of GA fluctuates violently, bewildering adversaries in the quality of a specific estimation, which inhibits the usage of GA in practice.

VII. Conclusion

We tackle seedless de-anonymization under a more practical social network model parameterized by *overlapping communities* than existing work. By MMSE, we derive a well-justified cost function minimizing the expected number of mismatched users. While showing the NP-hardness of minimizing MMSE, we validly transform it into WEMP which resolves the tension between optimality and complexity: (i) WEMP asymptotically returns a negligible mapping error under mild conditions facilitated by higher overlapping strength; (ii) WEMP can be algorithmically solved via CBDA, which exactly finds the optimum of WEMP. Extensive experiments further confirm the effectiveness of CBDA under overlapping communities.

VIII. Acknowledgement

This work was supported by NSF China (No. 61532012, 61325012, 61521062, 61602303 and 91438115).

References

[1] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford Large Network Dataset Collection", http://snap.stanford.edu/data, 2014.
[2] A. Narayanan and V. Shmatikov, "De-anonymizing social networks", in *IEEE Symposium on Security and Privacy*, pp. 173-187, 2009.
[3] P. Pedarsani and M. Grossglauser, "On the privacy of anonymized networks" in *Proc. ACM SIGKDD*, pp. 1235-1243, 2011.
[4] E. Kazemi, L. Yartseva and M. Grossglauser, "When can two unlabeled networks be aligned under partial overlap?", in *IEEE 53rd Annual Allerton Conference on Communication, Control, and Computing*, pp. 33-42, 2015.
[5] D. Cullina and N. Kiyavash, "Improved achievability and converse bounds for Erdős-Rényi graph matching", in *Proc. ACM SIGMETRICS*, pp. 63-72, 2016.
[6] S. Ji, W. Li, M. Srivatsa and R. Beyah, "Structural data de-anonymization: Quantification, practice, and implications", in *Proc. ACM CCS*, pp. 1040-1053, 2014.
[7] S. Ji, W. Li, N. Z. Gong, P. Mittal and R. Beyah, "On your social network de-anonymizablity: Quantification and large scale evaluation with seed knowledge" in *NDSS* 2015.
[8] E. Onaran, G. Siddharth and E. Erkip, "Optimal de-anonymization in random graphs with community structure", arXiv preprint arXiv:1602.01409, 2016.
[9] G. Palla, I. Derenyi, L. J. Farkas and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society", in *Nature*, No. 7043, Vol. 435, pp. 814-818, 2005.
[10] P. Erdős and A. Rényi, "On random graphs", in *Publicationes Mathematicae*, pp. 290-297, 1959.
[11] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
[12] L. Yartseva and M. Grossglauser, "On the performance of percolation graph matching", in *Proc. ACM COSN*, pp. 119-130, 2013.
[13] E. Kazemi, S. H. Hassani and M. Grossglauser, "Growing a graph matching from a handful of seeds", in *Proc. the VLDB Endowment*, pp. 1010-1021, 2015.
[14] C. F. Chiasserini, M. Garetto and E. Leonardi, "Social network de-anonymization under scale-free user relations", in *IEEE/ACM Trans. on Networking*, Vol. 24, No. 6, pp. 3756-3769, 2016.
[15] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks", in *Proc. the VLDB Endowment*, pp. 377-388, 2014.
[16] C. F. Chiasserini, M. Garetto and E. Leonardi, "Impact of clustering on the performance of network de-anonymization", in *Proc. ACM COSN*, pp. 83-94, 2015.
[17] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth", in *Knowledge and Information Systems*, No. 42, Vol. 1, pp. 181-213, 2015.
[18] P. Latouche, E. Birmel and C. Ambroise, "Overlapping stochastic block models with application to the french political blogosphere", in *The Annals of Applied Statistics* pp.309–336, 2011.
[19] G. H. Hardy, J. E. Littlewood and G. Plya, "Inequalities. Reprint of the 1952 edition." in *Cambridge Mathematical Library*, 1988
[20] O. Kariv and S. L. Hakimi, "Algorithm approach to network location problems - 2. the p-medians", in *Siam Journal on Applied Mathematics*, No. 3, Vol. 37, pp. 539-560, 1979.
[21] M. Zaslavskiy, F. Bach and J. P. Vert, "A path following algorithm for the graph matching problem" , in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, No. 12, Vol. 31, pp. 2227-2242, 2009.
[22] L. Fu, X. Fu, Z. Hu, Z. Xu and X. Wang, De-anonymization of Social Networks with Communities: When Quantifications Meet Algorithms, arXiv preprint arXiv:1703.09028, 2017.
[23] X. Fu, Z. Hu, Z. Xu, L. Fu and X. Wang, De-anonymization of Networks with Communities: When Quantifications Meet Algorithms, to appear in IEEE Globecom, 2017.