上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

学士学位论文

BACHELOR'S THESIS

社交网络去匿名化理论分析与算法设计

学生姓名 吴昕宇

学生学号 5140219173

专　　业 信息工程

指导教师 王新兵

学院(系) 电子信息与电气工程学院

Submitted in total fulfillment of the requirements for the degree of
Bachelor in Information Engineering

Theoretical Analysis and Algorithm Design for Social Network De-anonymization

Xinyu Wu

Advisor
Prof. Xinbing Wang

Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, P.R.China

社交网络去匿名化理论分析与算法设计

摘要

社交网络的不断发展对网络中用户的隐私已经产生了严重的威胁：隐私攻击者能够通过将匿名用户映射到一个相关的跨领域非匿名网络中，从而识别匿名用户的身份。在未知真实映射的情况下，前人通过提出各式各样的代价函数，以期衡量映射的质量。然而，目前该领域的研究中普遍缺失对这些代价函数背后机理的探究，以及缺失对各种常用网络模型下去匿名问题可行性的根本性研究。

在本项工作中，我们首先研究在具有重叠社区结构的社交网络中的去匿名化问题。我们考虑重叠社区这一更贴近实际社交网络的模型，在此模型下解决前人尚未探究之处。重叠社区可以视作攻击者去匿名的辅助信息。我们首先发现，针对未知真实映射的问题，攻击者可以基于最小均方误差准则构建一种合理的代价函数，从而能够在所有可能的映射中选取一个使得误差期望达到最小值。然而，我们证明选择这样的一种映射是 NP 难的，因此我们考虑将原问题合理近似转化为一个更易求解的问题，称为“带权值边匹配问题”（WEMP）。转化成 WEMP 使得问题的效能和求解难度均显著优化：（1）WEMP 的最优值能够使得在较为普遍的情况下使得映射错误的数量相对于整个网络规模的比例趋近于 0，同时重叠社区的性质将有助于映射错误数量的减少；（2）WEMP 能够通过我们提出的基于凸凹优化的去匿名算法（CDBA）在多项式时间内求解。

更深一步，我们探究了在三个最常用的社交网络模型中（包括：Erdos-Renyi (ER) 随机图模型、随机块模型以及幂律模型）决定攻击者是否能够成功去匿名的参数取值界。对于每种模型，我们考虑“全采样”和“部分采样”两种情形。我们给出清晰直观的参数取值界，展示参数取值的变化将如何影响去匿名效果，同时得出在采样概率给定下，总体上去匿名成功性排序为：随机块模型>ER 随机图模型>幂律模型。此外，我们通过提出一个时间复杂度至多为 $O(n^3 \log n)$ 的高效算法（n 代表网络规模），证明我们理论推导的界是可达的，同时远超目前所知最优的 $O(n^6)$ 复杂度的算法。

关键词：社交网络去匿名化，重叠社区结构，理论界

Theoretical Analysis and Algorithm Design for Social Network De-Anonymization

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 and fundamental study of de-anonymization under different network models.

In this work, we firstly focus on de-anonymization under overlapping community structure. We jointly tackle above concerns under a more practical social network model parameterized by overlapping communities, which 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 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 our proposed convex-concave based de-anonymization algorithm (CBDA).

Furthermore, we insightfully probe into the theoretical bounds of parameters for successful de-anonymization under Erdos-Renyi (ER) model, Stochastic Block model

(SBM), and Power Law model, three most common models in social networks. For each model, we consider two cases: fully sampled and partly sampled. We give explicit parametric bounds to clearly show how each parameter will affect de-anonymization in two cases, and generally show the de-anonymizability among these three models as SBM>ER>Power Law. Moreover, we prove that these bounds are achievable by proposing a novel efficient algorithm with time complexity $O(n^3 \log n)$ in terms of network size n, greatly outperforming state of art $O(n^6)$.

Keywords: Social network de-anonymization, Overlapping community structure, Theoretical bound.

Contents

List of Figures

List of Tables

List of Algorithms

List of Symbols

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
Q	Total number of communities
q	One of the communities
w_{ij}	The weight of node pair (i, j)
$\boldsymbol{C_i}$	Community representation of node i
$p_{\boldsymbol{C_i}\boldsymbol{C_j}}$	Probability of edge existence between node i and j with community representation $\boldsymbol{C_i}$ and $\boldsymbol{C_j}$ respectively
$\boldsymbol{\theta}$	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 of de-anonymization problem (the corresponding mapping)
$\tilde{\mathbf{\Pi}}(\tilde{\pi})$	The minimizer of weighted-edge matching problem (the corresponding mapping)
$\mathbf{\Pi^n}$	The set of $n \times n$ permutation matrices.
$g(\mathbf{\Pi})$	The objective function of MMSE problem

Chapter 1 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. In this work, we do **two** in-depth and novel studies of de-anonymization problem: We first consider a specific but more practical situation: social network de-anonymization with *overlapping* communities, and build a systematic study from theory, to algorithm, and to validation on real networks; Furthermore, we stand on a higher level to dissect de-anonymization by deriving the theoretical bounds for common network models including Erdos-Renyi model, Stochastic Block model, and Power Law model, which acts as thresholds determining whether de-anonymization can be done. To verify that our theoretical bounds is valid, we propose a novel candidate-set based algorithm to show that these bounds are achievable factually, with time complexity outperforming state of art at least 2 orders in terms of network size n.

Work 1: De-anonymization with Overlapping Communities

As far as we know, the only work that shares the closest correlation with us belongs to Fu et. al. [23, 24], 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 dataset-

s. 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.

Work 2. Theoretical Bounds and Achievable Algorithm for De-anonymization in Different Network Models

Generally, in this piece of work, we analyze the de-anonymizability of three ubiquitous models in social network problems: Erdos-Renyi model, Stochasic Block model, and Power-Law model. For each model, we consider two cases: fully and partly sampled case. We also propose our Candidate Set based De-anonymization Algorithm (CASDA) to demonstrate that our derived theoretical bounds are achievable in real case. The main contributions of our work can be specified as follows:

1. We derive the upper and lower bound of de-anonymization in E-R graph $G(n,p)$ under both fully and partly sampled situation: We discover that for fully sampled case, $p=\Theta\left(\frac{1}{n}\right)$ functions as a threshold deciding whether the E-R graph can be de-anonymized or not. For partly sampled case, de-anonymizing the E-R graph relies heavily on the value of sampling probability s. The upper and lower bound of expected mapping error are $O(n(1-s))$ and $\Omega(n(1-s)^2)$ respectively, so in order to guarantee controllable mapping error, $s=1-o\left(\frac{1}{n}\right)$.

2. We derive the upper and lower bound of de-anonymization in Stochastic Block Model $SBM(n, p, q)$ under both fully and partly sampled situation: For fully sampled case, we discover that p and q mutually determine the phase transition, and we show the tightness of this threshold. For partly sampled case, the de-anonymization error is related to the probability ratio $\frac{q}{p}$ and the distribution of community size: (i) When $\frac{q}{p} > 1$, higher homogeneity of this distribution benefits de-anonymization; (ii) When $\frac{q}{p} < 1$, higher heterogeneity of this distribution benefits de-anonymization; (iii) Generally, $p > q$ outperforms $p < q$ in de-anonymization result which corresponds that community structure make adversaries easier to de-anonymize users. Meanwhile, compared with the case in E-R graph model, Stochastic Block model generally performs better in de-anonymization than E-R graph model with given sampling probability due to higher distinguishability brought by the community structure.

3. We derive the upper bound of de-anonymization under Power Law model $\Gamma(n, \gamma, k_{\min})$ under both fully and partly sampled situation. For both situations, we unveil that lower γ and higher $k_{\min}$ (but not arbitrarily close to n) promise higher de-anonymization accuracy. Meanwhile, compared with the case in E-R graph model, Power Law model generally performs worse in de-anonymization than E-R graph model with given sampling probability due to the restriction of degree which tends to arouse higher symmetry in the graph.

4. We propose an efficient algorithm: Candidate Set based De-anonymization Algorithm (CASDA), with time complexity $O(n^3 \log n)$ superior to the state of art $O(n^6)$, to demonstrate that our derived threshold is achievable. Concretely, if the parameters in each network model take their values in the intervals which ensures successful de-anonymization, then CASDA can output the de-anonymization result with the corresponding guarantee as proved theoretically.

The rest of this thesis is organized as follows. Chapter 2 summarizes prior work about de-anonymization. Chapter 3 presents the first work about de-anonymization with overlapping communities. Chapter 4 shows our further exploration about theoretical

bounds and corresponding achievable algorithm design of de-anonymization in different network models. Chapter 5 generalizes our results and gives our conclusion.

Chapter 2 Related Work

Social network de-anonymization problem has been in the dimelight in recent decades. Narayanan and Shimatikove [2] formulated this problem initially. They presented its framework and proposed a generic algorithm, which did not utilize any side information except the network structure and worked based on some pre-identified nodes, called seed nodes.

Predicated on this seminal paper, a large amount of work emerges focusing on different facets of de-anonymization. One major division is whether the anonymized network is seeded or seedless, i.e., whether pre-identified nodes exist. For seeded anonymized network, as the pioneering work [2], the common idea to solve the problem is to design algorithms based on *percolation*, which means that the re-identification process starts from the seed nodes and identify their neighbor nodes iteratively until all the nodes are de-anonymized [2, 12–15]. Yartseva et al. [12], Kazemi et al. [13] and Fabiana et al. [14] studied seeded problem under Erdos-Renyi graph model, while Korula and Lattenzi [15] shed light on preferential attachment model.

However, in real situations it is often the case that adversaries are difficult to obtain seeded nodes before de-anonymizing [23, 24] due to the limited access to user profiles. For seedless networks, the major methodology is to propose cost functions and obtain an estimation of the correct mapping between two networks by optimizing these cost functions. Pedarsani and Grossglauer [3] are the precursors in de-anonymizing seedless networks. They studied this problem under Erdos-Renyi graph and their cost function was the number of mismatched edges. With the same cost function, Kazemi et al. [4] considered the situation where the nodes in two networks are overlapping partially, and Cullina and Kiyavash [5] further investigated the information-theoretic threshold for exact identification in [3]. However, the cost functions in [3–5] were not justified by rationale. One cost function based on Maximum A Posterior (MAP) has been justified by [8, 23, 24]. Onaran et al. [8] theoretically proved the validity of MAP and Fu et al. [23, 24] provided two approximation algorithms to solve this problem.

Another facet for de-anonymization problem is the amount of side information adversaries have. A large amount of work [2–5, 12–15], either in seeded or seedless situations, studied this problem without any side information except the topological structure of two networks, i.e., the edge sets in two networks. However, the clustering effect exists in real social networks, which has not been considered in work above. To incorporate clustering effect, Chiasserini et al. [16] studied clustering under seeded problem and drew the conclusion that the impact of clustering is double-edged, which may dramatically reduce the required seed nodes but make the algorithm more fragile to errors. Onaran et al. [8] and Fu et al. [23, 24] both studied clustering by modeling it as communities in two networks, and Fu et al. [23, 24] showed that the side information of communities makes for higher accuracy of the algorithms intended for seedless problem. However, as far as we know, no existing work has ever focused on overlapping communities, which is omnipresent in real situations, especially the large-scale social networks nowadays. Meanwhile, no prior work has conducted the theoretical bound analysis for parameters in each network model, which fundamentally characterizes the de-anonymizability of different network models.

Chapter 3 De-anonymization with Overlapping Communities

3.1 Preliminaries

In this section we introduce some basic definitions and lemmas which will be used in our later analysis.

3.1.1 Definitions

Definition 3.1. *(**Trace**) Given an $n \times n$ square matrix $\mathbf{Y}$, the trace of $\mathbf{Y}$ is $\mathbf{trY} = \sum_{i=1}^{n} \mathbf{Y}_{ii}$, where $\mathbf{Y}_{ii}$ denotes the element at the i_{th} row and i_{th} column of $\mathbf{Y}$.*

Definition 3.2. *(**Expectation Over Matrix**) Given a random matrix variable $\mathbf{A}$ and a function of $\mathbf{A}$, denoted as $f(\mathbf{A})$, then the expectation of $f(\mathbf{A})$ over matrix $\mathbf{A}$ is denoted as $\boldsymbol{E}_{\mathbf{A}}(f(\mathbf{A}))$.*

Definition 3.3. *(**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 $\mathbf{X}_{ij}$ denotes the element at the i_{th} row and j_{th} column of $\mathbf{X}$.

Definition 3.4. *(**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 $\forall i, j \in \{1, 2, ..., n\}, (\mathbf{Y} \circ \mathbf{Z})_{ij} = \mathbf{Y}_{ij}\mathbf{Z}_{ij}$, where $\mathbf{Y} \circ \mathbf{Z}$ is still an $n \times n$ matrix.*

Definition 3.5. *(**Approximation Ratio**) Given a maximization problem $\mathcal{I}$ and its optimal value $OPT(\mathcal{I})$, if an algorithm $\mathcal{A}$ outputs a solution $\mathcal{S}$ such that $\mathcal{S} \geq \tau OPT(\mathcal{I})$, where $\alpha \in [0, 1]$. Then the approximation ratio of this algorithm $\mathcal{A}$ for problem $\mathcal{I}$ is τ.*

3.1.2 Lemmas

Lemma 3.1.1. *(**Sequence Inequality** [20]) For two nonnegative sequences $a_1 \leq a_2 \leq a_3 \cdots \leq a_n$ and $b_1 \leq b_2 \leq b_3 \cdots \leq b_n$, let $\eta = \sum_{k=1}^{n} a_{i_k} b_{j_k}$ where $\{i_1, i_2, ..., i_n\}$ and*

$\{j_1, j_2, ..., j_n\}$ are both permutations of $\{1, 2, ..., n\}$. Then we can obtain the Sequence Inequality that yields to

$$\sum_{k=1}^{n} a_k b_k \geq \eta \geq \sum_{k=1}^{n} a_k b_{n+1-k}.$$

Lemma 3.1.2. *Let $A(n)$, $B(n)$, $C(n)$ and $D(n)$ denote four functions with variable n, such that $A(n) = o(B(n))$ and $C(n) = o(D(n))$, then when $n \to \infty$,*

$$\frac{A(n) + B(n)}{C(n) + D(n)} = \frac{B(n)}{D(n)}.$$

Lemma 3.1.3. *(**Stirling's Formula**) Stirling's formula presents an approximation for the factorial, $n!$, when $n \to \infty$, as*

$$n! \sim \sqrt{2\pi n}(\frac{n}{e})^n.$$

Lemma 3.1.4. *Given an $n \times n$ matrix $\mathbf{R}$ and an $n \times n$ permutation matrix $\mathbf{\Pi}$, then $||\mathbf{\Pi R}||_F = ||\mathbf{R\Pi}||_F = ||\mathbf{R}||_F$, i.e., multiplying a permutation matrix keeps invariant of the Frobenius norm.*

3.2 Models and Definitions

In this section, we firstly introduce the social network models, then give the definition of the social network de-anonymization problem.

3.2.1 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. For instance, in reality G may characterize the invisible 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.

Underlying Social Network

Let $G = (V, E, \mathbf{U})$ be the underlying graph, where V is the node set, E is the edge set and $\mathbf{U}$[1] is the adjacency matrix of G. We regard G as an undirected network and assume that the total number of nodes is $|V| = n$. To reflect the property of overlapping communities, we suppose G is generated based on the overlapping stochastic block model [18], the idea of which 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 its Q elements are independent boolean variables $C_{iq} \in \{0, 1\}$, with C_{iq} being the q_{th} row (element) 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:

$$\boldsymbol{C_i} \sim \prod_{q=1}^{Q} (p_q)^{C_{iq}} (1 - p_q)^{1-C_{iq}}, \tag{3-1}$$

where p_q is the probability of any node in G falling into community q. Hence we have

$$Pr(\boldsymbol{C_i} = \{C_{i1}, C_{i2}, ..., C_{iQ}\}^T) = \prod_{q=1}^{Q} (p_q)^{C_{iq}} (1 - p_q)^{1-C_{iq}}. \tag{3-2}$$

Intuitively, Eqn. (3-2) shows the probability of node i belonging to communities $q_1, q_2, ...q_\ell$ which make the boolean variable $C_{iq_k} = 1, k = 1, 2, ..., \ell$ while not belonging to other communities. We call $\boldsymbol{C_i}$ as the *community representation* of node i, since $\boldsymbol{C_i}$ explicitly represents to which communities node i belongs and does not belong. For instance, if node i belongs to communities 1, 2 and 3, then the community representation of node i is $\boldsymbol{C_i} = \{1, 1, 1, 0, 0, ..., 0\}^T$.

Unlike the stochastic block model in [19] which can only represent disjoint communities, the overlapping stochastic block model can measure the property of communities overlapping, which allows one node to belong to multiple communities. For ease of understanding, let us consider an example where node i belongs to both communities 1 and 2. Then we have $Pr(\boldsymbol{C_i} = \{C_{i1}, C_{i2}, ..., C_{iQ}\}^T) = p_1 p_2 \prod_{p=3}^{Q} (1 - p_q)$. For an edge

[1] $\mathbf{U}(i, j) = 1$ if $(i, j) \in E$ and $\mathbf{U}(i, j) = 0$ if $(i, j) \notin E$

$(i,j) \in E$, it is natural that the probability of the existence of this edge is determined by $\boldsymbol{C_i}$ and $\boldsymbol{C_j}$. Therefore we can set $Pr\{(i,j) \in E\} = Pr\{\mathbf{U}(i,j) = 1\} = p_{\boldsymbol{C_i C_j}}$, where $p_{\boldsymbol{C_i C_j}}$ is a pre-defined parameter representing the probability of edge existence between two nodes belonging to any community representation. It has been demonstrated in [18] that the overlapping stochastic block model turns out to be more reasonable in reality since overlapping property exists in social networks widely, and the parameters in this model can be estimated efficiently.

Published Network and Auxiliary Network

Now we proceed to define the published and auxiliary networks. Specifically, we let $G_1(V_1, E_1, \mathbf{A})$ denote the published network, which can be interpreted as a graph that shares the same node labeling as the underlying graph, with its edges independently sampled from G with some probability s_1. In contrast, an auxiliary network, denoted by $G_2(V_2, E_2, \mathbf{B})$, does not necessarily have the same node labeling as the underlying network and the edges are independently sampled from G with some probability s_2. Again, here $\mathbf{A}$ and $\mathbf{B}$ respectively represent the adjacency matrix of published and auxiliary networks.

In correspondence to real situations, G_1 characterizes the publicly available 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. The adversary (attacker) can leverage the information of G_2, and tries to identify the users in G_1 based on the edge relationship between and community representation of both G_1 and G_2. In terms of edge relationship, the node of high degree in G_1 should be of higher possibility to correspond to a node which is also of high degree in G_2. Therefore while de-anonymizing any node in G_2, the adversary can harness this *degree similarity* in matched node pairs to predict its corresponding node in G_1. In terms of community representation, the nodes in G_1 and G_2 with the same community representation should be matched with higher probability. Then the adversary can make use of this *community representation similarity* while judging whether a node in G_1 is matched with the node in G_2 to be de-anonymized with high probability.

For the edge set E_k $(k \in \{1,2\})$ of either network, we have

$$Pr\{(i,j) \in E_k\} = \begin{cases} s_k & \text{if } (i,j) \in E, \\ 0 & \text{if } (i,j) \notin E. \end{cases}$$

For the node sets V_1 and V_2, we assume that the number of nodes in G, G_1 and G_2 are the same, i.e., $|V| = |V_1| = |V_2| = n$. By this assumption, there exists bijective mapping between G_1 and G_2, as will be defined in Section 3.2.2. Note that it is easy to extend to the situation where $|V_1| \neq |V_2|$. Although the mapping between G_1 and G_2 in such case is no longer bijective, we only need to modify the permutation matrix (defined in Section 3.2.2) between G_1 and G_2 from a square matrix into a non-square one, which will not influence our theoretical analysis.

Furthermore, we should clarify that in our model we render each node pair (i,j) a weight w_{ij}, which, as will be defined in Section 3.2.2, is dependent on the parameter set for the node pair (i,j), i.e., $\theta_{ij} = \{p_{\boldsymbol{C_i C_j}}, s_1, s_2\}$. Different pairs of nodes may have different weights. As we will state in Section 3.2.2, w_{ij} reflects the probability of edge existence between nodes i and j, and the weights facilitates the reduction of the average de-anonymization error, which makes our estimation of permutation matrix more accurate.

Remark: According to the description above, it can be seen that G, G_1 and G_2 are all random variables. For the convenience of representation, 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 together.

3.2.2 Social Network De-anonymization

Predicated on the side information provided by the published network G_1 and the auxiliary network G_2, the goal of social network de-anonymization problem is to find a bijective node mapping $\pi : V_1 \mapsto V_2$, which is the true matching of nodes in G_1 and G_2. We can equivalently express this bijective mapping by forming a permutation ma-

trix $\mathbf{\Pi} \in \{0,1\}^{n\times n}$, where $\mathbf{\Pi}(i,j) = 1$ if $\pi(i) = j$ and $\mathbf{\Pi}(i,j) = 0$ otherwise. We denote $\mathbf{\Pi_0}$ as the true permutation matrix between G_1 and G_2, with π_0 representing the corresponding true bijective mapping. Note that we do not have any prior knowledge of $\mathbf{\Pi_0}$, and we do not have access to the underlying graph G of G_1 and G_2. Now we can formally define the social network de-anonymization problem as follows.

Definition 3.6. *(**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 bijective mapping π_0 between V_1 and V_2 (the true permutation matrix $\mathbf{\Pi_0}$ equivalently).*

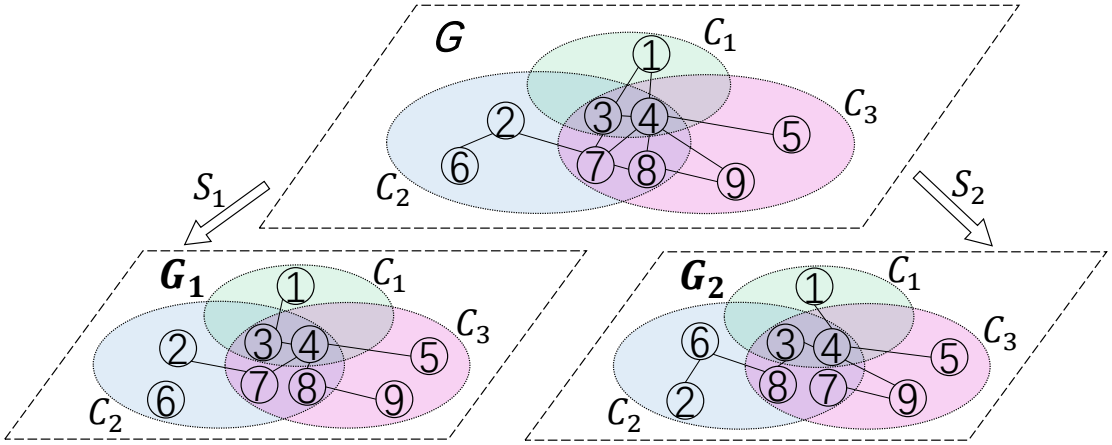


Figure 3.1 An example of the underlying graph (G), published graph (G_1) and auxiliary graph (G_2). The edges of $G_{1(2)}$ are sampled independently from G with probability $s_{1(2)}$. C_1, C_2, C_3 denote 3 different communities in our overlapping stochastic block model. Nodes 7 and 8 belong to 2 different communities. Nodes 3 and 4 belong to 3 different communities. The true mapping $\pi_0 = \{(1,1),(2,6),(3,3),(4,4),(5,5),(6,2),(7,8),(8,7),(9,9)\}$

Figure 1 illustrates an example of the proposed social de-anonymization problem that incorporates the feature of overlapping community. Here we note that our solution[2] to the social network de-anonymization problem, denoted as $\hat{\mathbf{\Pi}}$, is not necessarily equal to the $\mathbf{\Pi_0}$. To quantify the difference between our solution and true permutation matrix, we introduce a metric called "node mapping error (NME)", whose formal definition is provided as follows.

[2]Hereinafter our solution refers to the permutation matrix.

Definition 3.7. ***(Node Mapping Error)*** *Given the estimated permutation matrix $\hat{\mathbf{\Pi}}$ and the true permutation matrix $\mathbf{\Pi_0}$, the node mapping error (NME) between $\hat{\mathbf{\Pi}}$ and $\mathbf{\Pi_0}$ is defined as*

$$d(\hat{\mathbf{\Pi}}, \mathbf{\Pi_0}) = \frac{1}{2}||\hat{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2. \tag{3-3}$$

Obviously $d(\hat{\mathbf{\Pi}}, \mathbf{\Pi_0})$ equals to 0 if and only if two permutations are identical, and if there are k nodes mapped mistakenly, then it equals to k. Therefore this metric reveals how much the estimated permutation of nodes deviates from the true one. Based on the definition of NME, the goal of the social network de-anonymization problem is thus to minimize NME.

As we have mentioned earlier, we have no prior knowledge of $\mathbf{\Pi_0}$, the true permutation matrix. Moreover, with the given G_1 and G_2, $\mathbf{\Pi_0}$ in fact can be viewed as a random variable whose probability distribution is conditioned on these two networks. Note that regarding $\mathbf{\Pi_0}$ as a random variable does not contradict the fact that there is only one determined true mapping between G_1 and G_2 in real situations, because this true mapping can be perceived as a realization of the random variable $\mathbf{\Pi_0}$. Therefore, we consider selecting $\hat{\mathbf{\Pi}}$, an estimation of the permutation matrix which minimizes the expected or mean value of the node mapping error (NME). We call this estimation as "Minimum Mean Square Error (MMSE)", since in the following Definition 3.8 we can discover that it is the minimizer of the node mapping error in the form of mean square. The formal definition of MMSE is as follows.

Definition 3.8. ***(The MMSE Estimator)*** *Given the published network G_1, the auxiliary network G_2 and parameter set $\boldsymbol{\theta}$, the MMSE estimator is an estimation of permutation matrix which minimizes the number of mistakenly matched nodes in expectation, that is*

$$\hat{\mathbf{\Pi}} = \arg\min_{\mathbf{\Pi}\in\mathbf{\Pi}^n} \mathbf{E}_{\mathbf{\Pi_0}}\{d(\mathbf{\Pi}, \mathbf{\Pi_0})\} \quad = \arg\min_{\mathbf{\Pi}\in\mathbf{\Pi}^n} \sum_{\mathbf{\Pi_0}\in\mathbf{\Pi}^n} ||\mathbf{\Pi} - \mathbf{\Pi_0}||_F^2 Pr(\mathbf{\Pi_0}|G_1, G_2, \boldsymbol{\theta}), \tag{3-4}$$

where $\boldsymbol{E}_{\mathbf{\Pi_0}}$ means the expectation over all possible $\mathbf{\Pi_0}$. The posterior probability $Pr(\mathbf{\Pi_0}|G_1, G_2, \boldsymbol{\theta})$ means the probability of a possible true permutation matrix $\mathbf{\Pi_0}$ given G_1, G_2 and $\boldsymbol{\theta}$.

Table 3.1 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
Q	Total number of communities
q	One of the communities
w_{ij}	The weight of node pair (i, j)
$\boldsymbol{C_i}$	Community representation of node i
$p_{\boldsymbol{C_i}\boldsymbol{C_j}}$	Probability of edge existence between node i and j with community representation $\boldsymbol{C_i}$ and $\boldsymbol{C_j}$ respectively
$\boldsymbol{\theta}$	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 of de-anonymization problem (the corresponding mapping)
$\tilde{\mathbf{\Pi}}(\tilde{\pi})$	The minimizer of weighted-edge matching problem (the corresponding mapping)
$\mathbf{\Pi^n}$	The set of $n \times n$ permutation matrices.
$g(\mathbf{\Pi})$	The objective function of MMSE problem

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 used throughout the paper are summarized in Table 1.

3.3 Analytical Aspect of De-anonymization Problem

In this section, we start to provide analysis of the social network de-anonymization problem that we have defined earlier. In doing so, we firstly prove that this problem is NP-hard. To facilitate the problem analysis, we then give an approximation to the original MMSE estimator and verify it under the expectation of different possible network structures. Furthermore, we validate this approximation by proving that the approximation ratio is not small for a single possible network structure.

3.3.1 Transformation of MMSE Estimator

As can be seen from the definition of MMSE (Eqn. (3-4) in Section 3.2.2), the posterior probability $Pr(\mathbf{\Pi_0}|G_1, G_2, \boldsymbol{\theta})$ still needs to be expressed more explicitly. Inspired by the derivation in [8], we have the following theorem about the transformation of MMSE estimator.

Theorem 3.9. *Given the published graph G_1, the auxiliary graph G_2 and the parameter set $\boldsymbol{\theta}$, the MMSE estimator can be equivalently transformed into*

$$\begin{aligned}\hat{\mathbf{\Pi}} &= \arg\max_{\mathbf{\Pi}\in\mathbf{\Pi^n}} \sum_{\mathbf{\Pi_0}\in\mathbf{\Pi^n}} ||\mathbf{\Pi}-\mathbf{\Pi_0}||_F^2 ||\mathbf{W}\circ(\mathbf{\Pi_0 A}-\mathbf{B\Pi_0})||_F^2 \\ &= \arg\max_{\mathbf{\Pi}\in\mathbf{\Pi^n}} g(\mathbf{\Pi}),\end{aligned} \tag{3-5}$$

where $g(\mathbf{\Pi}) = \sum_{\mathbf{\Pi_0}\in\mathbf{\Pi^n}} ||\mathbf{\Pi}-\mathbf{\Pi_0}||_F^2 ||\mathbf{W}\circ(\mathbf{\Pi_0 A}-\mathbf{B\Pi_0})||_F^2$ is the objective function of the MMSE problem, $\mathbf{W}$ is the weight matrix in which $W(i,j) = \sqrt{w_{ij}} = W(j,i)$, $w_{ij} = \log\left(\frac{1-p_{\boldsymbol{C_i C_j}}(s_1+s_2-s_1 s_2)}{p_{\boldsymbol{C_i C_j}}(1-s_1)(1-s_2)}\right)$ is weight between nodes i and j, and "$\circ$" denotes the Hadamard product.

Proof. 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

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

Let us focus on the conditional probability $Pr(G, \mathbf{\Pi_0}|G_1, G_2, \boldsymbol{\theta})$ in Eqn. (2). Ac-

cording to Bayesian's formula, along with the fact that G_1 and G_2 are sampled independently from each other, we obtain

$$Pr(G, \mathbf{\Pi_0}|G_1, G_2, \boldsymbol{\theta}) = \frac{Pr(G, G_1, G_2, \mathbf{\Pi_0})}{Pr(G_1, G_2)} \sim Pr(G)Pr(G_1|G)Pr(G_2|G, \mathbf{\Pi_0}), \tag{3-6}$$

where $a \sim b$ means that a and b are different only in parameters unrelated to $\mathbf{\Pi_0}$, which will not change the value of $\arg\max$ or $\arg\min$.[3] Note that the parameter set $\boldsymbol{\theta}$ remains invariant, so we need not add $\boldsymbol{C_i}$ and $\boldsymbol{\theta}$ into further consideration.

Set E^{ij} as the indicator variable about whether an edge exists between nodes i and j in the edge set E. If an edge exists then $E^{ij} = 1$, otherwise $E^{ij} = 0$. The same rule also holds for indicators E_1^{ij} and E_2^{ij}. Therefore Eqn. (3-6) can be further written as

$$\begin{aligned}
\sum_{G\in\mathcal{G}_{\mathbf{\Pi}}} Pr(G)Pr(G_1|G)Pr(G_2|G, \Pi_0) &= \sum_{G\in\mathcal{G}_{\mathbf{\Pi}}} \prod_{i<j} s_1^{E_1^{ij}}(1-s_1)^{E^{ij}-E_1^{ij}} s_2^{E_2^{\pi_0(i)\pi_0(j)}} \\
&\quad \cdot (1-s_2)^{E^{ij}-E_2^{\pi_0(i)\pi_0(j)}} p_{\boldsymbol{C_iC_j}}^{E^{ij}}(1-p_{\boldsymbol{C_iC_j}})^{1-E^{ij}} \\
&= \prod_{i<j}\left(\frac{s_1}{1-s_1}\right)^{E_1^{ij}}\left(\frac{s_2}{1-s_2}\right)^{E_2^{\pi_0(i)\pi_0(j)}} \\
&\quad \cdot \sum_{G\in\mathcal{G}_{\mathbf{\Pi}}}\left((1-s_1)(1-s_2)\frac{p_{\boldsymbol{C_iC_j}}}{1-p_{\boldsymbol{C_iC_j}}}\right)^{E^{ij}} \\
&\sim \sum_{G\in\mathcal{G}_{\mathbf{\Pi}}}\left((1-s_1)(1-s_2)\frac{p_{\boldsymbol{C_iC_j}}}{1-p_{\boldsymbol{C_iC_j}}}\right)^{E^{ij}}.
\end{aligned} \tag{3-7}$$

Note that the last equivalence in Eqn. (3-7) holds since the term $\left(\frac{s_1}{1-s_1}\right)^{E_1^{ij}}$ does not depend on π_0 and the product $\prod_{i<j}\left(\frac{s_2}{1-s_2}\right)^{E_2^{\pi_0(i)\pi_0(j)}}$ is independent of π_0 due to the bijective property of π_0.

Then we define $G_{\pi_0}^*$ as the graph which has the smallest number of edges in $\mathcal{G}_{\mathbf{\Pi}}$. Equivalently $G_{\pi_0}^* = (V, E_1 \cup \pi_0(E_1))$, where $\pi_0(E_1) = \{(\pi_0(i), \pi_0(j))|(i,j) \in E_1\}$. Now we set $E_{\pi_0}^*$ as the edge set of $G_{\pi_0}^*$, and $E_{\pi_0}^{*ij}$ as the indicator variable between nodes i and j, i.e., $E_{\pi_0}^{*ij} = 1$ if $(i,j) \in E_{\pi_0}^*$ and $E_{\pi_0}^{*ij} = 0$ otherwise. Then we sum up all the

[3] There is a notation abuse for $\sim$ between the one in Eqn. (3-1) and here.

graphs in $\mathcal{G}_{\mathbf{\Pi}}$

$$\sum_{G\in\mathcal{G}_{\mathbf{\Pi}}}\left((1-s_1)(1-s_2)\frac{p_{\boldsymbol{C_iC_j}}}{1-p_{\boldsymbol{C_iC_j}}}\right)^{E^{ij}} = \prod_{i<j}^{n}\left((1-s_1)(1-s_2)\frac{p_{\boldsymbol{C_iC_j}}}{1-p_{\boldsymbol{C_iC_j}}}\right)^{E^{*ij}_{\pi_0}} \cdot \sum_{k=0}^{E_{ij}-E^{*ij}_{\pi_0}} C^{k}_{E_{ij}-E^{*ij}_{\pi_0}}\left((1-s_1)(1-s_2)\frac{p_{\boldsymbol{C_iC_j}}}{1-p_{\boldsymbol{C_iC_j}}}\right)^{k}. \tag{3-8}$$

Note that in Eqn. (3-8) last multiplicative factor ,

$$\sum_{k=0}^{E_{ij}-E^{*ij}_{\pi_0}} C^{k}_{E_{ij}-E^{*ij}_{\pi_0}}\left((1-s_1)(1-s_2)\frac{p_{\boldsymbol{C_iC_j}}}{1-p_{\boldsymbol{C_iC_j}}}\right)^{k},$$

yields as a Bernoulli sum, therefore Eqn. (3-8) can be further written as

$$\begin{aligned}\sum_{G\in\mathcal{G}_{\mathbf{\Pi}}}\left((1-s_1)(1-s_2)\frac{p_{\boldsymbol{C_iC_j}}}{1-p_{\boldsymbol{C_iC_j}}}\right)^{E^{ij}} &= \prod_{i<j}^{n}\left((1-s_1)(1-s_2)\frac{p_{\boldsymbol{C_iC_j}}}{1-p_{\boldsymbol{C_iC_j}}}\right)^{E^{*ij}_{\pi_0}} \cdot \left(1+(1-s_1)(1-s_2)\frac{p_{\boldsymbol{C_iC_j}}}{1-p_{\boldsymbol{C_iC_j}}}\right)^{1-E^{*ij}_{\pi_0}} \\ &\sim \prod_{i<j}^{n}\left(\frac{p_{\boldsymbol{C_iC_j}}(1-s_1)(1-s_2)}{1-p_{\boldsymbol{C_iC_j}}(s_1+s_2-s_1s_2)}\right)^{E^{*ij}_{\pi_0}} \\ &\sim \sum_{i<j}^{n} E^{*ij}_{\pi_0}\log\left(\frac{p_{\boldsymbol{C_iC_j}}(1-s_1)(1-s_2)}{1-p_{\boldsymbol{C_iC_j}}(s_1+s_2-s_1s_2)}\right).\end{aligned} \tag{3-9}$$

Here the last line in Eqn. (3-9) holds since the log operator keeps the minimum $\mathbf{\Pi_0}$ invariant. Note that $G^*_{\pi_0} = (V, E_1 \cup \pi_0(E_1))$. Then we can find that $E^{*ij}_{\mathbf{\Pi_0}} = 0$ if and only if both E^{ij}_1 and E^{ij}_2 are equal to 0, and $E^{*ij}_{\mathbf{\Pi_0}} = 1$ occurs in the following three conditions:

- $(i,j) \in E_1$ but $(i,j) \notin E_2$. Note that this condition also ensures that $(\pi_0(i), \pi_0(j)) \in E_2$.
- $(i,j) \in E_2$ but $(i,j) \notin E_1$. Note that this condition also ensures that $(\pi_0(i), \pi_0(j)) \notin E_2$.

- $(i,j) \in E_1$ and $(i,j) \in E_2$. Note that this condition also ensures that $(\pi_0(i), \pi_0(j)) \in E_2$.

Synthesizing all the above conditions, we can express $E_{\pi_0}^{*ij}$ as

$$E_{\pi_0}^{*ij} = \frac{1}{2}(E_1^{ij} + E_2^{ij} + |\mathbb{1}\{(i,j) \in E_1\} - \mathbb{1}\{(\pi_0(i), \pi_0(j)) \in E_2\}|), \tag{3-10}$$

where $\mathbb{1}\{P\} = 1$ if the random event P happens and $\mathbb{1}\{P\} = 0$ otherwise. Substituting Eqn. (3-10) into the last line in Eqn. (3-9), we get

$$\begin{aligned} &\arg\min_{\mathbf{\Pi}\in\mathbf{\Pi^n}} \sum_{i<j}^{n} E_{\pi_0}^{*ij} \log\left(\frac{p_{\boldsymbol{C_i C_j}}(1-s_1)(1-s_2)}{1-p_{\boldsymbol{C_i C_j}}(s_1+s_2-s_1s_2)}\right) \\ &= \arg\max_{\mathbf{\Pi}\in\mathbf{\Pi^n}} \sum_{i<j}^{n} w_{ij}|\mathbb{1}\{(i,j) \in E_1\} - \mathbb{1}\{(\pi_0(i), \pi_0(j)) \in E_2\}| \\ &= \arg\max_{\mathbf{\Pi}\in\mathbf{\Pi^n}} ||\mathbf{W} \circ (\mathbf{\Pi_0 A} - \mathbf{B\Pi_0})||_F^2, \end{aligned} \tag{3-11}$$

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

Substituting Eqn. (3-11) into Eqn. (3-6), now we can formulate the MMSE estimator as

$$\hat{\mathbf{\Pi}} = \arg\max_{\mathbf{\Pi}\in\mathbf{\Pi^n}} \sum_{\mathbf{\Pi_0}\in\mathbf{\Pi^n}} ||\mathbf{\Pi} - \mathbf{\Pi_0}||_F^2 ||\mathbf{W} \circ (\mathbf{\Pi_0 A} - \mathbf{B\Pi_0})||_F^2. \tag{3-12}$$

□

Remark: Additionally, to simplify the form of $||\mathbf{W} \circ (\mathbf{\Pi_0 A} - \mathbf{B\Pi_0})||_F^2$, we use $\mathbf{\Pi_0}\hat{\mathbf{A}}$ to represent $\mathbf{W} \circ \mathbf{\Pi_0 A}$, and $\hat{\mathbf{B}}\mathbf{\Pi_0}$ to represent $\mathbf{W} \circ \mathbf{B\Pi_0}$[4]. Therefore we can rewrite the MMSE estimator in Eqn. (3-12) as

$$\hat{\mathbf{\Pi}} = \arg\max_{\mathbf{\Pi}\in\mathbf{\Pi^n}} \sum_{\mathbf{\Pi_0}\in\mathbf{\Pi^n}} ||\mathbf{\Pi} - \mathbf{\Pi_0}||_F^2 ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2, \tag{3-13}$$

[4]We should clarify that we only provide a simpler form to represent $\mathbf{W} \circ \mathbf{\Pi_0 A}$ and $\mathbf{W} \circ \mathbf{B\Pi_0}$, and it does NOT imply that $\mathbf{W} \circ \mathbf{A} = \hat{\mathbf{A}}$ and $\mathbf{W} \circ \mathbf{B} = \hat{\mathbf{B}}$. But some operations under this new notation still hold, for example, multiplying a permutation matrix does not change the value of the Frobenius norm, i.e., $||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 = ||\mathbf{W} \circ (\mathbf{\Pi_0 A} - \mathbf{B\Pi_0})||_F^2 = ||\mathbf{W} \circ \mathbf{\Pi_0^T}(\mathbf{\Pi_0 A} - \mathbf{B\Pi_0}))||_F^2 = ||\mathbf{W} \circ (\mathbf{A} - \mathbf{\Pi_0^T B\Pi_0})||_F^2$ and $||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 = ||\mathbf{\Pi_0}\hat{\mathbf{A}}\mathbf{\Pi_0^T} - \hat{\mathbf{B}}||_F^2$.

and $g(\mathbf{\Pi}) = \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\mathbf{\Pi} - \mathbf{\Pi_0}||_F^2 ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2$. In the following analysis, we use the form in Eqn. (3-13). In Section 3.4.1, we will discuss the condition under which $\mathbf{W} \circ \mathbf{A} = \hat{\mathbf{A}}$ and $\mathbf{W} \circ \mathbf{B} = \hat{\mathbf{B}}$.

3.3.2 NP-hardness of Solving the MMSE Estimator

Since we have derived a more explicit form of MMSE estimator, we are interested in whether there exists a polynomial-time algorithm that can solve the MMSE problem. However, as we will prove in the sequel, this problem is NP-hard, meaning that no polynomial time (pseudo-polynomial time) approximation algorithm exists for solving the MMSE estimator.

Proposition 3.3.1. *Solving the MMSE estimator is an NP-hard problem. There is no polynomial time or pseudo-polynomial time approximation algorithm for this problem with any multiplicative approximation guarantee unless P=NP.*

Proof. We derive the proof in two steps: 1. modeling this problem as a clique with weighted nodes and edges, and 2. reducing the 1-median problem to our MMSE problem[5]. Here the 1-median problem [21] 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 (also nonnegative) between nodes v and v^*.

Reduction from 1-median problem: Our construction of the clique works as follows: Suppose there are n nodes in G_1 and G_2. Then for any permutation matrix

[5]Note that 1-median itself is not NP-hard, but we demonstrate that the lower bound of 1-median is of $O(n)$ and when applied in our problem it becomes larger than polynomial.

$\mathbf{\Pi} \in \mathbf{\Pi^n}$, we have

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

in which all the multiplicative factors are nonnegative. Since the number of elements in $\mathbf{\Pi^n}$ is $n!$, then we construct a clique with $n!$ nodes, with every node representing an $n \times n$ permutation matrix. We set the distance between two nodes i and j as $D(i,j) = 4n - ||\mathbf{\Pi}(i) - \mathbf{\Pi}(j)||_F^2$. Note that this distance satisfies the triangular equality $D(i,k) + D(k,j) \geq D(i,j)$, which assures that the edge directly connecting nodes i and j has the minimum distance among all possible paths between them. So the shortest path length between nodes i and j is just the distance $D(i,j)$. We define the weight of node i as $\omega(i) = ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2$ (Note that each $\mathbf{\Pi_0}$ is a node in the graph with $n!$ nodes). For ease of understanding, Fig. 3.2 illustrates the constructed clique with 5 nodes.

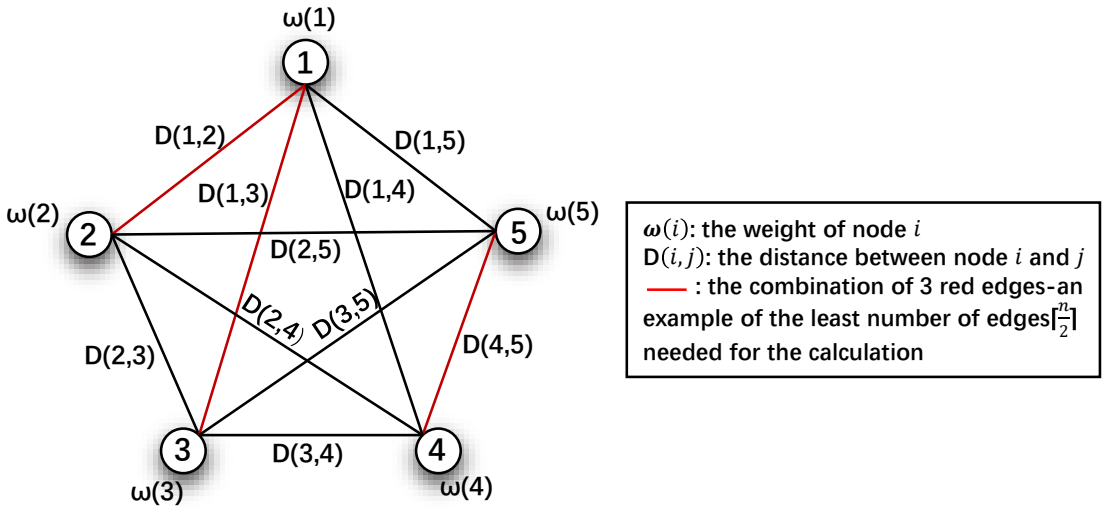


Figure 3.2 An Illustration of the Constructed Clique with 5 Nodes

The Lower Bound for 1-Median Problem: Based on the above construction, we equivalently transform our problem into the form of a 1-median problem:

$$i_0 = \arg\min_{i^*\in V} \sum_{i\in V} \omega(i) \cdot D(i, i^*).$$

For a 1-median problem with n nodes, it is easy to discover that we need to calculate at least $\lceil n/2 \rceil$ times, since we need at least $\lceil n/2 \rceil$ edges to form an edge set such that the endpoints of all edges in this edge set cover all the vertices in the graph. Or else one node will not be calculated for any edge connecting it, thus no information about this node is revealed, and then we can not judge whether this node is the one we intend to find. The red lines in Fig. 3.2 illustrates an example that when there are 5 nodes, the least number of edges needed to be calculated is $\lceil 5/2 \rceil = 3$. For our MMSE estimator problem we have $n!$ nodes, thus the calculation times is at least $(n/2)!$, which means that we need to calculate $(n/2)!$ permutation matrices. Compared with the size of the problem, n^2, the complexity turns out to be $\Omega(((\sqrt{n})/2)!) = \Omega(\sqrt{n}!)$, which exceeds polynomial.

□

The NP-hardness of MMSE estimator shows the impossibility to pursue an exact algorithm or any approximation algorithm with multiplicative guarantee. Thus we need to simplify this problem by conducting reasonable approximation to make it possible to solve this problem, with certain tolerance of mapping error. In the following we propose one way to approximate this problem, the analysis of which will indicate that the error arose by this approximation can be bounded.

3.3.3 Approximation of the MMSE estimator

As we have just stated above, the NP-hardness of MMSE problem urges us to find proper approximation for the original problem. Recall that MMSE involves all the possible true mappings, the number of which is $n!$, thus leading to fairly prohibitive computational cost. To tackle the difficulty, we firstly transform the original MMSE problem into a weighted-edge matching problem (WEMP), which, as we will define and present more details later, simplifies the form of objective function of the original MMSE problem and makes it tractable. Then we demonstrate that this transformation is valid, meaning that the solution of WEMP will not deviate much from the solution of the original MMSE problem by proving its high approximation ratio. Definition 3.10

provides the formal statement of WEMP.

Definition 3.10. ***(Weighted-Edge Matching Problem)*** *Given the adjacent matrices of G_1 and G_2, denoted as $\mathbf{A}$ and $\mathbf{B}$ respectively, and the weight matrix $\mathbf{W}$, the weight-edge matching problem is to find*

$$\tilde{\mathbf{\Pi}} = \arg\min_{\mathbf{\Pi}\in\mathbf{\Pi^n}} ||\mathbf{\Pi}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi}||_F^2.$$

Hereinafter we discuss the following two aspects of WEMP:

- How do we transform from the original MMSE problem into WEMP?
- How is the validity of this transformation?

The Idea of Transformation

We intend to transform the original problem of solving the MMSE estimator into WEMP. The idea of this transformation can be interpreted in the following sense: for any fixed $\mathbf{\Pi}$, define a set $S_k(\mathbf{\Pi}), 0 \le k \le n$, any element of which is an $n \times n$ permutation matrix $\mathbf{\Pi_0}$ such that $d(\mathbf{\Pi}, \mathbf{\Pi_0}) = 2k$. It is obvious that $S_0(\mathbf{\Pi}) = \{\mathbf{\Pi}\}$ and $S_1(\mathbf{\Pi}) = \emptyset$. Then we can reform the original problem as

$$\hat{\mathbf{\Pi}} = \arg\max_{\mathbf{\Pi}\in\mathbf{\Pi^n}} \sum_{k=0}^{n} k \left(\sum_{\mathbf{\Pi_0}\in S_k(\mathbf{\Pi})} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 \right). \tag{3-14}$$

Zooming in on Eqn. (3-14), we propose our idea of transforming it into WEMP. To present our idea clearly, we divide our analysis into three parts; First we analyze a single term, $||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2$, where $\mathbf{\Pi_0} \in S_2(\mathbf{\Pi})$; Then we analyze $\mathbf{\Pi_0} \in S_k(\mathbf{\Pi})$ based on the analysis of $\mathbf{\Pi_0} \in S_2(\mathbf{\Pi})$; Finally we analyze the R.H.S of Eqn. (3-14) based on Lemma 3.1.1. In the sequel we unfold the three parts in a more detailed manner:

1. Analysis of $||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2$ where $\mathbf{\Pi_0} \in S_2(\mathbf{\Pi})$

Now we focus on the value of $||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2$, where $\mathbf{\Pi_0} \in S_2(\mathbf{\Pi})$. Note that any permutation in $S_2(\mathbf{\Pi})$ only causes matching error on one pair of nodes. Thus if we consider $\mathbf{\Pi} = \tilde{\mathbf{\Pi}}$ and set one specific $\mathbf{\Pi_0} \in S_2(\tilde{\mathbf{\Pi}})$, which differs from $\tilde{\mathbf{\Pi}}$ only in the

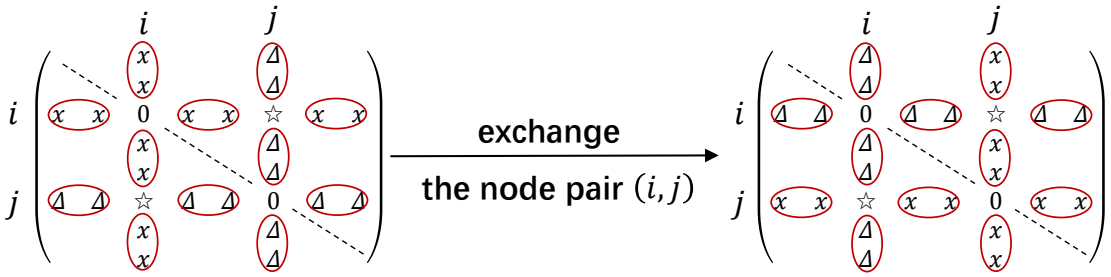


Figure 3.3 An example of the effect of $\mathbf{\Pi_0}$ which differs from $\tilde{\mathbf{\Pi}}_0$ only in the i_{th} and j_{th} row. The triangles denote the j_{th} row and column the "x"es denote the i_{th} row and column of $\mathbf{W} \circ (\mathbf{\Pi_0 A \Pi_0^T} - \mathbf{B})$. And the triangles denote the i_{th} row and column the "x"es denote the j_{th} row and column of $\mathbf{W} \circ (\tilde{\mathbf{\Pi}} \mathbf{A} \tilde{\mathbf{\Pi}}^{\mathbf{T}} - \mathbf{B})$. Note that the difference between $\mathbf{W} \circ (\mathbf{\Pi_0 A \Pi_0^T} - \mathbf{B})$ and $\mathbf{W} \circ (\tilde{\mathbf{\Pi}} \mathbf{A} \tilde{\mathbf{\Pi}}^{\mathbf{T}} - \mathbf{B})$ exists in the i_{th} and j_{th} row and column except the intersections (those 0s and stars).

i_{th} and j_{th} row, we can derive that

$$
\begin{aligned}
&||\mathbf{\Pi_0 \hat{A}} - \mathbf{\hat{B} \Pi_0}||_F^2 - ||\tilde{\mathbf{\Pi}} \mathbf{\hat{A}} - \mathbf{\hat{B}} \tilde{\mathbf{\Pi}}||_F^2 \\
&= ||\mathbf{W} \circ (\mathbf{\Pi_0 A \Pi_0^T} - \mathbf{B})||_F^2 - ||\mathbf{W} \circ (\tilde{\mathbf{\Pi}} \mathbf{A} \tilde{\mathbf{\Pi}}^{\mathbf{T}} - \mathbf{B})||_F^2 \\
&= 2\Big(\sum_{k \neq i,j}^{n} [(\mathbf{W} \circ (\mathbf{\Pi_0 A \Pi_0^T} - \mathbf{B}))_{ik}^2 - (\mathbf{W} \circ (\tilde{\mathbf{\Pi}} \mathbf{A} \tilde{\mathbf{\Pi}}^{\mathbf{T}} - \mathbf{B}))_{ik}^2] \\
&\quad + \sum_{k \neq i,j}^{n} [(\mathbf{W} \circ (\mathbf{\Pi_0 A \Pi_0^T} - \mathbf{B}))_{jk}^2 - (\mathbf{W} \circ (\tilde{\mathbf{\Pi}} \mathbf{A} \tilde{\mathbf{\Pi}}^{\mathbf{T}} - \mathbf{B}))_{jk}^2] \Big) \\
&= 2\Big(\sum_{k \neq i,j}^{n} w_{ik} [(\mathbf{\Pi_0 A \Pi_0^T} - \mathbf{B})_{ik}^2 - (\tilde{\mathbf{\Pi}} \mathbf{A} \tilde{\mathbf{\Pi}}^{\mathbf{T}} - \mathbf{B})_{ik}^2] \\
&\quad + \sum_{k \neq i,j}^{n} w_{jk} [(\mathbf{\Pi_0 A \Pi_0^T} - \mathbf{B})_{jk}^2 - (\tilde{\mathbf{\Pi}} \mathbf{A} \tilde{\mathbf{\Pi}}^{\mathbf{T}} - \mathbf{B})_{jk}^2] \Big) \\
&= 2\Big(\sum_{k \neq i,j}^{n} w_{ik} [\mathbf{\Pi_0 A \Pi_0^T} - \tilde{\mathbf{\Pi}} \mathbf{A} \tilde{\mathbf{\Pi}}^{\mathbf{T}}]_{ik} \psi(\mathbf{B}_{ik}) \\
&\quad + \sum_{k \neq i,j}^{n} w_{jk} [\mathbf{\Pi_0 A \Pi_0^T} - \tilde{\mathbf{\Pi}} \mathbf{A} \tilde{\mathbf{\Pi}}^{\mathbf{T}}]_{jk} \psi(\mathbf{B}_{jk}) \Big),
\end{aligned}
\tag{3-15}
$$

where $\psi(x) = -1$ if $x = 1$ and $\psi(x) = 1$ if $x = 0$. Fig. 3.3 illustrates how Eqn. (3-15) can be derived intuitively. Note that if $\mathbf{\Pi_0}$ and $\tilde{\mathbf{\Pi}}$ are different only in the i_{th} and j_{th} rows, then the difference between $||\mathbf{W} \circ (\mathbf{\Pi_0 A \Pi_0^T} - \mathbf{B})||_F^2$ and $||\mathbf{W} \circ (\tilde{\mathbf{\Pi}} \mathbf{A} \tilde{\mathbf{\Pi}}^{\mathbf{T}} - \mathbf{B})||_F^2$ exists in the red circles in Fig. 3.3, which corresponds to the third line in Eqn. (3-15). Note that the intersection part, i.e., the stars in Fig. 3.3, does not contribute to the $||\mathbf{W} \circ (\mathbf{\Pi_0 A \Pi_0^T} - \mathbf{B})||_F^2$ and $||\mathbf{W} \circ (\tilde{\mathbf{\Pi}} \mathbf{A} \tilde{\mathbf{\Pi}}^{\mathbf{T}} - \mathbf{B})||_F^2$.

Note that since $\mathbf{\Pi_0}$ and $\tilde{\mathbf{\Pi}}$ are different in the i_{th} and j_{th} rows, then $(\tilde{\mathbf{\Pi}}\mathbf{A}\tilde{\mathbf{\Pi}}^{\mathbf{T}})_{ik} = (\mathbf{\Pi_0}\mathbf{A}\mathbf{\Pi_0^T})_{jk}$. Therefore

$$
\begin{aligned}
&||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 - ||\tilde{\mathbf{\Pi}}\hat{\mathbf{A}} - \hat{\mathbf{B}}\tilde{\mathbf{\Pi}}||_F^2 \\
&= 2\Big(\sum_{k\neq i,j}^{n} w_{ik}\psi(\mathbf{B}_{ik})([\mathbf{\Pi_0}\mathbf{A}\mathbf{\Pi_0^T}]_{ik} - [\mathbf{\Pi_0}\mathbf{A}\mathbf{\Pi_0^T}]_{jk}) \\
&\quad + \sum_{k\neq i,j}^{n} w_{jk}\psi(\mathbf{B}_{jk})([\mathbf{\Pi_0}\mathbf{A}\mathbf{\Pi_0^T}]_{jk} - [\mathbf{\Pi_0}\mathbf{A}\mathbf{\Pi_0^T}]_{ik})\Big) \\
&= 2\Big(\sum_{k\neq i,j}^{n} (w_{ik}\psi(\mathbf{B}_{ik}) - w_{jk}\psi(\mathbf{B}_{jk})) \\
&\quad \cdot [(\mathbf{\Pi_0}\mathbf{A}\mathbf{\Pi_0^T})_{ki} - (\mathbf{\Pi_0}\mathbf{A}\mathbf{\Pi_0^T})_{kj}]\Big),
\end{aligned}
\tag{3-16}
$$

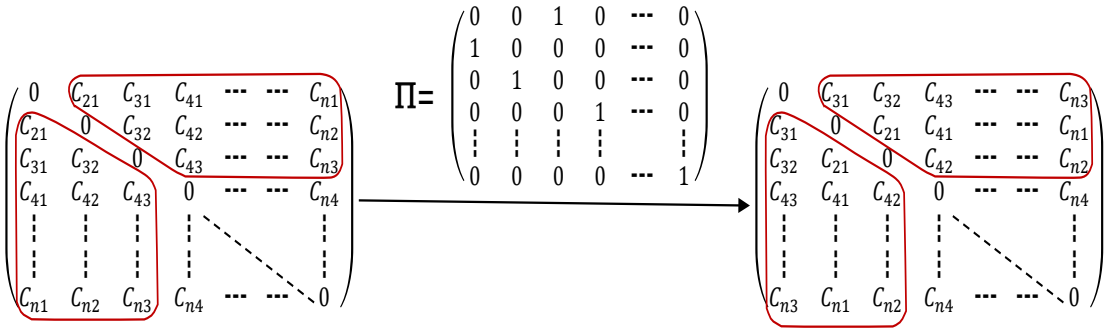


Figure 3.4 An example of the effect of $\mathbf{\Pi} \in S_3(\tilde{\mathbf{\Pi}})$, where we set $\tilde{\mathbf{\Pi}} = \mathbf{I}$. I is the identity matrix. Note that under the $\mathbf{\Pi}$ above the arrow, which differs from I only in the first three rows (columns). Thus the possible difference between two matrices only exists in the red circles, with $6n - 6$ elements in the matrix involved.

Since G_1 and G_2 are independently sampled from G, then $\mathbf{A}$ and $\mathbf{B}$ are independent. Thus we can first take the expectation of $\mathbf{B}$ on both sides of Eqn. (3-16). Note that the probability for the edge existence between nodes i and j in $\mathbf{B}$ is $p_{\boldsymbol{C_iC_j}}s_2$, therefore $\mathbf{E}[\psi(B_{ij})] = (-1)p_{\boldsymbol{C_iC_j}}s_2 + (1 - p_{\boldsymbol{C_iC_j}}s_2) = 1 - 2p_{\boldsymbol{C_iC_j}}s_2$. Hence, taking the expectation of $\mathbf{B}$ on both sides of Eqn. (3-16) and we obtain

$$
\begin{aligned}
&\mathbf{E_B}(||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 - ||\tilde{\mathbf{\Pi}}\hat{\mathbf{A}} - \hat{\mathbf{B}}\tilde{\mathbf{\Pi}}||_F^2) \\
&= 2\sum_{k\neq i,j}^{n} [w_{ik}[(1 - 2p_{\boldsymbol{C_iC_k}}s_2) - w_{jk}(1 - 2p_{\boldsymbol{C_jC_k}}s_2)) \\
&\quad \cdot ((\mathbf{\Pi_0}\mathbf{A}\mathbf{\Pi_0^T})_{ki} - (\mathbf{\Pi_0}\mathbf{A}\mathbf{\Pi_0^T})_{kj})].
\end{aligned}
$$

Similarly, taking the expectation of $\mathbf{A}$ on both sides, we have

$$\begin{aligned}
&\mathbf{E}_{\mathbf{A},\mathbf{B}}(||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 - ||\tilde{\mathbf{\Pi}}\hat{\mathbf{A}} - \hat{\mathbf{B}}\tilde{\mathbf{\Pi}}||_F^2) \\
&= 2\sum_{k\neq i,j}^{n}(w_{ik}(1-2p_{\boldsymbol{C_iC_k}}s_2) - w_{jk}(1-2p_{\boldsymbol{C_jC_k}}s_2)) \\
&\quad \cdot (p_{\boldsymbol{C_{\pi_0(i)}C_{\pi_0(k)}}} - p_{\boldsymbol{C_{\pi_0(j)}C_{\pi_0(k)}}})s_1 \\
&= 2\sum_{k\neq i,j}^{n}\Delta_{i,j,k,\pi_0},
\end{aligned}$$

where

$$\begin{aligned}
\Delta_{i,j,k,\pi_0} &= (w_{ik}(1-2p_{\boldsymbol{C_iC_k}}s_2) - w_{jk}(1-2p_{\boldsymbol{C_jC_k}}s_2)) \\
&\quad \cdot (p_{\boldsymbol{C_{\pi_0(i)}C_{\pi_0(k)}}} - p_{\boldsymbol{C_{\pi_0(j)}C_{\pi_0(k)}}})s_1.
\end{aligned}$$

Δ_{i,j,k,π_0} reflects a part of the difference $||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 - ||\tilde{\mathbf{\Pi}}\hat{\mathbf{A}} - \hat{\mathbf{B}}\tilde{\mathbf{\Pi}}||_F^2$ caused by the difference of a single element in matrices $\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}$ and $\tilde{\mathbf{\Pi}}\hat{\mathbf{A}} - \hat{\mathbf{B}}\tilde{\mathbf{\Pi}}$.[6] Since we consider the average case of all possible $\mathbf{\Pi_0}$, we also consider the average value of Δ_{i,j,π_0}, which we set as $\hat{\Delta} = \mathbf{E}_{i,j,\pi_0}(\Delta_{i,j,\pi_0})$. Note that $\mathbf{E}_{\mathbf{A},\mathbf{B}}(||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 - ||\tilde{\mathbf{\Pi}}\hat{\mathbf{A}} - \hat{\mathbf{B}}\tilde{\mathbf{\Pi}}||_F^2) > 0$ since $\tilde{\mathbf{\Pi}}$ is the minimizer of $||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2$. Therefore $\hat{\Delta} = \mathbf{E}_{i,j,\pi_0}(\Delta_{i,j,\pi_0}) > 0$.

2. Analysis of $\sum_{\mathbf{\Pi_0}\in S_k(\mathbf{\Pi})}||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2$

Now we move to the second part involved in our idea. We first focus on $S_k(\mathbf{\Pi_0})$, and count the number of elements in $S_k(\mathbf{\Pi_0})$, denoted as $|S_k|$. Note that if there are k mismatched nodes in a graph with n nodes, there are C_n^k possible sets of mismatched nodes. We define $|T_k|$ as the number of elements in each possible set, and can get $|S_k| = C_n^k|T_k|$. For $|T_k|$, we can find that it satisfies

$$\begin{aligned}
|T_k| &= (k-1)(|T_{k-2}| + (k-2)(|T_{k-3}| + (k-3)(|T_{k-4}| + ...))) \\
&= \sum_{t=1}^{k-1}(\prod_{i=1}^{t}(k-i))|T_{k-t-1}|.
\end{aligned} \tag{3-17}$$

[6]For example, the difference of the corresponding element (with the same notation, e.g., (i,k) in the left matrix and (j,k) in the right matrix, both of which are triangles.) in two matrices in Fig. 3.3 inside one of the red circles

Consider $|T_k|$ and $|T_{k-1}|$ in Eqn. (3-17), we can discover that

$$|T_k| = (k-1)(|T_{k-2}| + |T_{k-1}|) \geq (k-1)|T_{k-1}|, k \geq 2.$$

Therefore we obtain the relationship between $|S_k|$ and $|S_{k-1}|$ as

$$|S_k| = C_n^k |T_k| \geq (k-1)\frac{C_n^k}{C_n^{k-1}}|S_{k-1}| = (1-\frac{1}{k})(n-k+1)|S_{k-1}|, \tag{3-18}$$

where $k \geq 2$. Eqn. (3-18) shows that when k is much smaller than n, then $\frac{|S_k|}{|S_{k-1}|} = (1-\frac{1}{k})(n-k+1)$ is large; when k gets close to n, then $\frac{|S_k|}{|S_{k-1}|}$ approaches 1, which means that $|S_k|$ and $|S_{k-1}|$ are almost the same.

Now we consider $\mathbf{\Pi_0} \in S_k(\mathbf{\Pi})$. Note that for any $\mathbf{\Pi_0} \in S_k(\mathbf{\Pi})$, there are k rows and columns that may cause the difference between $||\mathbf{W} \circ (\mathbf{\Pi_0 A \Pi_0^T} - \mathbf{B})||_F^2$ and $||\mathbf{W} \circ (\mathbf{\tilde{\Pi} A \tilde{\Pi}^T} - \mathbf{B})||_F^2$. Fig. 3.4 illustrates an example of $\mathbf{\Pi_0} \in S_3(\mathbf{\Pi})$. Therefore we can discover for any $\mathbf{\Pi_0} \in S_k(\mathbf{\Pi})$, the number of node pairs (i,j) which may influence the difference between $||\mathbf{W} \circ (\mathbf{\Pi_0 A \Pi_0^T} - \mathbf{B})||_F^2$ and $||\mathbf{W} \circ (\mathbf{\tilde{\Pi} A \tilde{\Pi}^T} - \mathbf{B})||_F^2$ is approximately $\sum_{i=1}(n-i) = \frac{(2n-k-1)k}{2}$.[7] Thus, denoting N_k as this number of node pair, we can obtain

$$\begin{aligned} N_k &= \frac{(2n-k-1)k}{2}|S_k| \\ &\geq \frac{(2n-k-1)k}{2}(1-\frac{1}{k})(n-k+1)|S_{k-1}| \\ &= (1-\frac{1}{k})(n-k+1)\frac{(2n-k-1)k}{(2n-k)(k-1)}N_{k-1} \\ &= (n-k+1)\frac{2n-k-1}{2n-k}N_{k-1}. \end{aligned}$$

[7] For example, in Fig. 3.4 when $k=3$ the number is $6n-6$. Although there may be some elements which do not cause error, such as the two stars in Fig. 3.3, the number of this kinds of node pairs can be neglected when n is large enough.

Therefore in average, we have

$$\begin{aligned}\sum_{\mathbf{\Pi_0}\in S_k}||\mathbf{\Pi_0}\hat{\mathbf{A}}-\hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 &= N_k\hat{\Delta}\\ &\geq (n-k+1)\frac{2n-k-1}{2n-k}N_{k-1}\hat{\Delta}\\ &\geq (n-k+1)\frac{2n-k-1}{2n-k}\sum_{\mathbf{\Pi_0}\in S_{k-1}}||\mathbf{\Pi_0}\hat{\mathbf{A}}-\hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2\\ &\approx (n-k+1)\sum_{\mathbf{\Pi_0}\in S_{k-1}}||\mathbf{\Pi_0}\hat{\mathbf{A}}-\hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2,\end{aligned} \tag{3-19}$$

where the last approximation holds since $k \leq n$ and when $n \to \infty$, $\frac{2n-k-1}{2n-k} \to 1$.

Therefore, we can claim that in average, if $k_1 > k_2$, then

$$\sum_{\mathbf{\Pi_0}\in S_{k_1}}||\mathbf{\Pi_0}\hat{\mathbf{A}}-\hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 > \sum_{\mathbf{\Pi_0}\in S_{k_2}}||\mathbf{\Pi_0}\hat{\mathbf{A}}-\hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2. \tag{3-20}$$

3. Maximum Value Under Sequence Inequality

Based on the analysis above, if we set $\mathbf{\Pi} = \tilde{\mathbf{\Pi}}$, then we find that when $k = 0$, the minimum value in the set $\{0, 2, 3, \cdots, n\}$. Thus $\sum_{\mathbf{\Pi_0}\in S_0(\tilde{\mathbf{\Pi}})}||\mathbf{\Pi_0}\hat{\mathbf{A}}-\hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 = ||\tilde{\mathbf{\Pi}}\hat{\mathbf{A}}-\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}||_F^2$ is also the minimum value in the set

$$\left\{\begin{array}{cc}\sum_{\mathbf{\Pi_0}\in S_0(\tilde{\mathbf{\Pi}})}||(\mathbf{\Pi_0}\hat{\mathbf{A}}-\hat{\mathbf{B}}\mathbf{\Pi_0})||_F^2, & \sum_{\mathbf{\Pi_0}\in S_2(\tilde{\mathbf{\Pi}})}||(\mathbf{\Pi_0}\hat{\mathbf{A}}-\hat{\mathbf{B}}\mathbf{\Pi_0})||_F^2,\\ \sum_{\mathbf{\Pi_0}\in S_3(\tilde{\mathbf{\Pi}})}||(\mathbf{\Pi_0}\hat{\mathbf{A}}-\hat{\mathbf{B}}\mathbf{\Pi_0})||_F^2, ..., & \sum_{\mathbf{\Pi_0}\in S_n(\tilde{\mathbf{\Pi}})}||(\mathbf{\Pi_0}\hat{\mathbf{A}}-\hat{\mathbf{B}}\mathbf{\Pi_0})||_F^2\end{array}\right\}.$$

Thus according to Lemma 3.1.1, we know that in average case, by setting $\mathbf{\Pi}$ in the original MMSE objective function

$$\sum_{\mathbf{\Pi_0}\in\Pi^n}||\mathbf{\Pi}-\mathbf{\Pi_0}||_F^2||\mathbf{W}\circ(\mathbf{\Pi_0}\mathbf{A}-\mathbf{B}\mathbf{\Pi_0})||_F^2$$

equal to $\tilde{\mathbf{\Pi}}$, the minimizer of WEMP, then this original MMSE objective function reaches its largest value under Sequence Inequality.

Moreover, note that if we do not set $\tilde{\mathbf{\Pi}} = \hat{\mathbf{\Pi}}$, for example set $\tilde{\mathbf{\Pi}} = \mathbf{\Pi} \in S_k(\mathbf{\Pi})$, we can verify that $\mathbf{\Pi}$ does not make the objective function in Eqn. (3-14) larger than $\mathbf{\Pi_0}$

since

$$0||\mathbf{\Pi}\hat{\mathbf{A}}\mathbf{\Pi}^{\mathbf{T}}-\hat{\mathbf{B}}||_{\mathbf{F}}^{\mathbf{2}}+2k||\mathbf{\Pi}\hat{\mathbf{A}}\mathbf{\Pi}^{\mathbf{T}}-\hat{\mathbf{B}}||_{\mathbf{F}}^{\mathbf{2}}\geq 2k||\hat{\mathbf{\Pi}}\hat{\mathbf{A}}\hat{\mathbf{\Pi}}^{\mathbf{T}}-\hat{\mathbf{B}}||_{\mathbf{F}}^{\mathbf{2}}+0||\mathbf{\Pi}\hat{\mathbf{A}}\mathbf{\Pi}^{\mathbf{T}}-\hat{\mathbf{B}}||_{\mathbf{F}}^{\mathbf{2}},$$

which means that the Sequency Inequality preserves that when $||\mathbf{\Pi_0}\hat{\mathbf{A}}\mathbf{\Pi_0^T}-\hat{\mathbf{B}}||_F^2$ achieves its minimum, then $||\Pi-\Pi_0||_F^2$ also achieves its minimum. Therefore by setting $\tilde{\mathbf{\Pi}}=\hat{\mathbf{\Pi}}$ we can achieve the largest value of the original MMSE problem under this sequence inequality.

However, as we noted earlier, we can only transform the original MMSE problem into WEMP in an average case of network structures. This implies that the transformation is not necessarily the best approximation of a single network structure. In the following we further analyze the validity of this transformation in a possible network structure by showing the approximation ratio of our transformation is large (at least larger than 0.5).

The Validity of Transformation

As we have stated above, $\tilde{\mathbf{\Pi}}$ does not necessarily achieve the maximum of the original MMSE problem for a specific network structure. That is to say there may exist error in $g(\tilde{\mathbf{\Pi}})$ and $g(\hat{\mathbf{\Pi}})$, where $g(\hat{\mathbf{\Pi}})$ is the maximum value of the original MMSE objective function and $g(\tilde{\mathbf{\Pi}})$ is the value of MMSE objective function when $\mathbf{\Pi}$ equals to the minimizer of WEMP. If we demonstrate that this error can be bounded within a small range, then we can say that this approximation is *valid*. Theorem 3.11 shows that under the mild condition indicated by Inequality (3-19), we can get approximation ratio $g(\tilde{\mathbf{\Pi}})/g(\hat{\mathbf{\Pi}})$ larger than 0.5, which, to some extent, makes our estimation reasonable.

Theorem 3.11. *Given the published graph G_1, the auxiliary graph G_2, the parameter set $\boldsymbol{\theta}$ and the weight matrix $\mathbf{W}$, in average case we have the approximation ratio $g(\tilde{\mathbf{\Pi}})/g(\hat{\mathbf{\Pi}})$ larger than 0.5.*

Proof. First we have

$$g(\hat{\mathbf{\Pi}}) - g(\tilde{\mathbf{\Pi}}) = \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} (||\hat{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 - ||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2)||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2. \tag{3-21}$$

Then we divide the set $\mathbf{\Pi^n}$ into two subsets:

$$\mathbf{\Pi_1^n} = \{\mathbf{\Pi} \in \mathbf{\Pi^n} | ||\hat{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 > ||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2\};$$
$$\mathbf{\Pi_2^n} = \{\mathbf{\Pi} \in \mathbf{\Pi^n} | ||\hat{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 < ||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2\}.$$

Following that we divide the Eqn. (3-21) into two sets, $\mathbf{\Pi_1^n}$ and $\mathbf{\Pi_2^n}$:

$$\begin{aligned} g(\hat{\mathbf{\Pi}}) - g(\tilde{\mathbf{\Pi}}) = & \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi_1^n}} (||\hat{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 - ||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2)||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 \\ & - \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi_2^n}} (||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 - ||\hat{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2)||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 \\ \leq & ||\tilde{\mathbf{\Pi}} - \hat{\mathbf{\Pi}}||_F^2 \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi_1^n}} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2. \end{aligned} \tag{3-22}$$

where the last inequality holds due to the triangular inequality $||\hat{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 - ||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 \leq ||\tilde{\mathbf{\Pi}} - \hat{\mathbf{\Pi}}||_F^2$ and the term $\sum_{\mathbf{\Pi_0} \in \mathbf{\Pi_2^n}} (||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 - ||\hat{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2)||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2$ is positive. Then we have

$$\begin{aligned} \frac{g(\hat{\mathbf{\Pi}}) - g(\tilde{\mathbf{\Pi}})}{g(\tilde{\mathbf{\Pi}})} &= \frac{(||\tilde{\mathbf{\Pi}} - \hat{\mathbf{\Pi}}||_F^2) \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi_1^n}} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2}{\sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2} \\ &\leq \frac{2\beta n \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\mathbf{\Pi_0}\tilde{\mathbf{A}} - \tilde{\mathbf{B}}\mathbf{\Pi_0}||_F^2}{\sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 ||\mathbf{\Pi_0}\tilde{\mathbf{A}} - \tilde{\mathbf{B}}\mathbf{\Pi_0}||_F^2}. \end{aligned} \tag{3-23}$$

where $||\tilde{\mathbf{\Pi}} - \hat{\mathbf{\Pi}}||_F^2 = 2\beta n$ and $\beta \in [0, 1]$ is the ratio between the number of mistakenly matched nodes and that of all the nodes. The last inequality in $(3-23)$ holds because $\mathbf{\Pi_1^n} \subset \mathbf{\Pi^n}$.

Now we divide the sum $\sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2$ into two parts:

$$D_1 = \sum_{k \leq \rho n} \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2;$$

$$D_2 = \sum_{\rho n < k \le n} \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2.$$

where ρ is any real number in $[0, 1]$ and we assume that ρn is an integer[8].

For D_1, in average case we can obtain

$$\begin{aligned} D_1 &\le \sum_{i=1}^{\rho n} \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 \le \sum_{i=1}^{\rho n} \prod_{j=1}^{i} 2(n-j+1) \\ &\le \sum_{i=1}^{\rho n} (2n)^i = 2n\frac{(2n)^{\rho n} - 1}{2n-1} \approx (2n)^{\rho n}. \end{aligned}$$

For D_2, according to Inequality (3-19), in average case we can get

$$\begin{aligned} D_2 &\ge \sum_{k=\rho n+1}^{n} \prod_{j=1}^{k} (n-j+1) = \sum_{k=\rho n+1}^{n} \frac{n!}{(n-k)!} \\ &\ge \sum_{k=\rho n+1}^{n} \frac{n!}{((1-\rho)n)!} = (1-\rho)n\frac{n!}{((1-\rho)n)!}. \end{aligned}$$

Note that if we set $\rho = \Omega(1) = c_0$, where $c_0 \to 1$, then $\rho \to 1$ and

$$D_2 \ge c_0\frac{n!}{c_0!} = cn! \sim c\sqrt{2\pi n}(\frac{n}{e})^n,$$

where c is a constant and the last step holds due to the Stirling's formula. Therefore we can upper bound $\frac{D_2}{D_1}$ as

$$\frac{D_2}{D_1} \ge c\frac{\sqrt{2\pi n}(\frac{n}{e})^n}{(2n)^{\rho n}} = c\sqrt{2\pi n}\left(\frac{n^{1-\rho}}{2^\rho e}\right)^n.$$

Then if ρ is a constant which approaches 1 but does not equal to 1, then we find that when $n \to \infty$, D_2 is of higher order of n than D_1. Therefore we can easily verify that in the denominator of the last term in Inequality (3-23), $\sum_{\rho n < k \le n} \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2$ is of higher order of n than $\sum_{k \le \rho n} \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2$, since for $k_1 > \rho n$ and $k_2 < \rho n$, $\mathbf{\Pi'_1} \in S_{k_1}(\tilde{\mathbf{\Pi}})$ and $\mathbf{\Pi'_2} \in S_{k_2}(\tilde{\mathbf{\Pi}})$, we have $||\mathbf{\Pi'_1} - \tilde{\mathbf{\Pi}}||_F^2 \ge ||\mathbf{\Pi'_2} - \tilde{\mathbf{\Pi}}||_F^2$. Therefore according to Lemma 3.1.2, we can leave the term with highest order of n in the denominator and numerator in the last term in

[8] If it is not an integer, we can easily modify it by rounding.

Inequality (3-23) when $n \to \infty$ and thus we can obtain

$$\frac{2\beta n \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2}{\sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2} \approx \frac{2\beta n \sum_{\rho n < k \leq n} \sum_{\mathbf{\Pi_0} \in S_k(\tilde{\mathbf{\Pi}})} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2}{\sum_{\rho n < k \leq n} \sum_{\mathbf{\Pi_0} \in S_k(\tilde{\mathbf{\Pi}})} ||\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}}||_F^2 ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2}$$
$$\leq \frac{2\beta n}{2\rho n} \frac{\sum_{\rho n < k \leq n} \sum_{\mathbf{\Pi_0} \in S_k(\tilde{\mathbf{\Pi}})} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2}{\sum_{\rho n < k \leq n} \sum_{\mathbf{\Pi_0} \in S_k(\tilde{\mathbf{\Pi}})} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2} = \frac{\beta}{\rho}.$$

Thus we have the approximation ratio

$$\frac{g(\tilde{\mathbf{\Pi}})}{g(\hat{\mathbf{\Pi}})} \geq \frac{1}{1 + \frac{\beta}{\rho}} \approx \frac{1}{1+\beta} \geq \frac{1}{2}.$$

□

Note that in the proof of Theorem 3.11, we use several times of inequality scaling method to derive the lower bound of approximation ratio, which is 0.5. These inequality scaling may cause this lower bound to be smaller than the real approximation ratio. That is to say, the approximation ratio 0.5 may be even worse than the approximation ratio in the worst case in real situations. For example, in Inequality (3-23) we directly use

$$\sum_{\mathbf{\Pi_0} \in \mathbf{\Pi_1^n}} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 \leq \sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2,$$

which may cause a big gap. Therefore, for a more general situation we have the following corollary.

Corollary 3.3.1. *Given the published graph G_1, the auxiliary graph G_2, the parameter set $\boldsymbol{\theta}$ and the weight matrix $\mathbf{W}$, and we let*

$$\chi = \Big(\sum_{\mathbf{\Pi_0} \in \mathbf{\Pi_1^n}} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 \Big) \Big/ \Big(\sum_{\mathbf{\Pi_0} \in \mathbf{\Pi^n}} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 \Big),$$

then in average case, the approximation $g(\tilde{\mathbf{\Pi}})/g(\hat{\mathbf{\Pi}})$ ratio is larger than $\frac{1}{1+\beta\chi}$.

This corollary can be easily proved by slightly changing the form of Eqn. (3-23). To take an example to illustrate the gap of approximation ratio caused by χ more intuitively,

we assume that $\sum_{\mathbf{\Pi}\in\mathbf{\Pi_1^n}} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2 = \sum_{\mathbf{\Pi}\in\mathbf{\Pi_2^n}} ||\mathbf{\Pi_0}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi_0}||_F^2$ [9]. Then $\chi = \frac{1}{2}$ and $(\frac{1}{1+\beta\chi}) > \frac{2}{3}$, which causes the gap of the lower bound of approximation ratio to be $\frac{2}{3} - \frac{1}{2} = \frac{1}{6}$.

Note that we still claim that the approximation ratio is *larger* than $(\frac{1}{1+\beta\chi})$. This is because we eliminate the sum $\sum_{\mathbf{\Pi_0}\in\mathbf{\Pi_2^n}}(||\hat{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2 - ||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2)||\mathbf{\Pi}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi}||_F^2$ in Eqn. (3-22), which also generates a gap between the lower bound $\frac{1}{1+\beta\chi}$ and the real approximation ratio. We leave it a future direction to find a proper estimation of this gap. However, the current gap still ensures the real approximation ratio strictly larger than $\frac{1}{1+\beta\chi}$, which further strengthens our claim at the beginning of Section 3.3.3 that the transformation of the original MMSE problem is valid.

3.4 Algorithmic Aspect of De-anonymization Problem

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).

3.4.1 The Influence of Transformation to WEMP on 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 caus ed 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 ben-

[9] This is only a very special situation, which we use it to make an intuitive example to explain how χ causes the gap of approximation ratio. It is not necessarily the same as real situations

efits brought by overlapping communities in NME reduction. Theorem 3.12 formally presents our result mentioned above. Before that, we give Lemma 3.4.1, a prerequisite in proving Theorem 3.12.

Lemma 3.4.1. *Suppose the permutation matrix* $\mathbf{\Pi}$ *keeps invariant of the community representation of all the nodes, i.e.,* $\forall \mathbf{\Pi}$ *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*

$$||\mathbf{\Pi}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi}||_F = ||\mathbf{W} \circ (\mathbf{\Pi}\mathbf{A}\mathbf{\Pi}^{\mathbf{T}} - \mathbf{B})||_F = ||\mathbf{\Pi}\hat{\mathbf{A}}\mathbf{\Pi}^{\mathbf{T}} - \hat{\mathbf{B}}||_F. \tag{3-24}$$

Proof. We know $||\mathbf{\Pi}\hat{\mathbf{A}} - \hat{\mathbf{B}}\mathbf{\Pi}||_F = ||\mathbf{W} \circ (\mathbf{\Pi}\mathbf{A} - \mathbf{B}\mathbf{\Pi})||_F$, thus we only need to prove that $\mathbf{W} \circ \mathbf{\Pi}\mathbf{A} = \mathbf{\Pi}\mathbf{W} \circ \mathbf{A}$. Note that w_{ij} only depends on $p_{\boldsymbol{C_iC_j}}$, s_1 and s_2, therefore for some nodes i, j, s, t, if $\boldsymbol{C_i} = \boldsymbol{C_s}$ and $\boldsymbol{C_j} = \boldsymbol{C_t}$, then $\mathbf{W}(i,j) = \mathbf{W}(s,t)$. This fact tells that the weight is invariant within communities. Therefore, since $\mathbf{\Pi}$ keeps invariant of the community representation of all the nodes, it is easy to verify that $\mathbf{W} \circ \mathbf{\Pi}\mathbf{A} = \mathbf{\Pi}\mathbf{W} \circ \mathbf{A}$. Thus we have $\hat{\mathbf{A}} = \mathbf{W} \circ \mathbf{A}$ and similarly, $\hat{\mathbf{B}} = \mathbf{W} \circ \mathbf{B}$. Then Eqn. (3-24) holds naturally. □

Remark: According to Lemma 3.4.1, we can similarly show that $||\mathbf{W} \circ (\mathbf{A} - \mathbf{\Pi}\mathbf{B}\mathbf{\Pi}^{\mathbf{T}})||_F = ||\hat{\mathbf{A}} - \mathbf{\Pi}\hat{\mathbf{B}}\mathbf{\Pi}^{\mathbf{T}}||_F$, and there are no differences in form between $||\mathbf{\Pi_1}\hat{\mathbf{A}}\mathbf{\Pi}_{\mathbf{1}}^{\mathbf{T}} - \hat{\mathbf{B}}||_F$ and $||\hat{\mathbf{A}} - \mathbf{\Pi_2}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{2}}^{\mathbf{T}}||_F$ since the mappings are bijections and we can simply set $\mathbf{\Pi_2} = \mathbf{\Pi}_{\mathbf{1}}^{\mathbf{T}}$. Therefore, in the following we do not distinguish the forms $||\mathbf{\Pi}\hat{\mathbf{A}}\mathbf{\Pi}^{\mathbf{T}} - \hat{\mathbf{B}}||_F$ and $||\hat{\mathbf{A}} - \mathbf{\Pi}\hat{\mathbf{B}}\mathbf{\Pi}^{\mathbf{T}}||_F$.

Theorem 3.12. *Given the published network* G_1, *the auxiliary network* G_2, *the parameter set* $\boldsymbol{\theta}$, *the weight matrix* $\mathbf{W}$. *Set* $\mathbf{A}$ *as the adjacent matrix of* G_1, *and* $\mathbf{B}$ *as the adjacent matrix of* G_2. *Set* $\tilde{p}_{\boldsymbol{C_iC_j}} = w_{ij}p_{\boldsymbol{C_iC_j}}$ *and*

$$K = \min_{s,t,j}\{(\tilde{p}_{\boldsymbol{C_sC_j}} + \tilde{p}_{\boldsymbol{C_tC_j}})\min\{s_1, s_2\}\},$$

$$L = \max_{s,t,j}\{[(\tilde{p}_{\boldsymbol{C_sC_j}} + \tilde{p}_{\boldsymbol{C_tC_j}})\max\{s_1, s_2\}]^2\}.$$

If the following four conditions:

- $\frac{L}{K} = o(1)$;
- *the minimizer of WEMP,* $\tilde{\mathbf{\Pi}}$*, satisfies that* $||\hat{\mathbf{A}} - \mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0^T}||_F^2/||\hat{\mathbf{A}} - \tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}||_F^2 = \Omega(1)$;
- $||\hat{\mathbf{A}} - \mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0^T}||_F^2 = o(Kn^2)$;
- $\mathbf{\Pi_0}$ *and* $\tilde{\mathbf{\Pi}}$ *keep invariant of the community representation of all the nodes,*

hold, then the relative NME, $\frac{||\tilde{\mathbf{\Pi}}-\mathbf{\Pi_0}||_F^2}{||\mathbf{\Pi_0}||_F^2}$*, can be upper bounded by the minimum value of WEMP, i.e.,* $||\hat{\mathbf{A}} - \tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}||_F^2$*, and as* $n \to \infty$*,* $\frac{||\tilde{\mathbf{\Pi}}-\mathbf{\Pi_0}||_F^2}{||\mathbf{\Pi_0}||_F^2} \to 0$.

Proof. We divide our proof into four main parts. Firstly, we start from $||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F$ and upper bound it using $||(\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}||_F$ (or equivalently $||(\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{A}}||_F$). Secondly, we find the relationship between $||(\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}||_F$ and $\mathbf{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}})\hat{\mathbf{A}})$. Thirdly we upper bound the $\mathbf{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}})\hat{\mathbf{A}})$ and finally we upper bound $\frac{||(\mathbf{\Pi_0}-\tilde{\mathbf{\Pi}})||_F^2}{||\mathbf{\Pi_0}||_F^2}$, the relative NME, based on the first three steps.

1. Relationship Between $||\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}}||_F$ and $||(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}||_F$

We start with the first part and focus on $||(\mathbf{\Pi_0}-\tilde{\mathbf{\Pi}})\hat{\mathbf{B}}||_F$. For the i_{th} row of $(\mathbf{\Pi_0}-\tilde{\mathbf{\Pi}})$, there are two possibilities: (i) If $\mathbf{\Pi_0}$ and $\tilde{\mathbf{\Pi}}$ map node i in G_2 to the same node in G_1, then the i_{th} row of $(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}$ is a zero row vector; (ii) If $\mathbf{\Pi_0}$ and $\tilde{\mathbf{\Pi}}$ map node i to node s and t respectively ($s \neq t$), then 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, $([(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}]_{ij})^2 = (\sqrt{w_{sj}}\mathbf{B}_{sj} - \sqrt{w_{tj}}\mathbf{B}_{tj})^2$. Taking the expectation on both sides, we can derive that

$$\begin{aligned}&\mathbf{E}[(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}]_{ij}^2 = \mathbf{E}(\hat{\mathbf{B}}_{sj} - \hat{\mathbf{B}}_{tj})^2\\ &= (\tilde{p}_{\boldsymbol{C_sC_j}} + \tilde{p}_{\boldsymbol{C_tC_j}} - 2\sqrt{w_{sj}w_{tj}}p_{\boldsymbol{C_sC_j}}p_{\boldsymbol{C_tC_j}}s_2)s_2 \sim (\tilde{p}_{\boldsymbol{C_sC_j}} + \tilde{p}_{\boldsymbol{C_tC_j}})s_2.\end{aligned}$$

So by summing up all the columns, we have

$$\begin{aligned}&\mathbf{E}\sum_{j=1}^{n}[(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}]_{ij}^2\\ &= \sum_{j=1}^{n}(\tilde{p}_{\boldsymbol{C_sC_j}} + \tilde{p}_{\boldsymbol{C_tC_j}})s_2.\end{aligned}$$

Then summing up all the rows, we can obtain

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

where $\mathbf{1}\{\pi_0(i)\neq\tilde{\pi}(i)\}=1$ if π_0 and $\tilde{\pi}$ map node i in G_2 to the same node in G_1 and $\mathbf{1}\{\pi_0(i)\neq\tilde{\pi}(i)\}=0$ otherwise. Thus it eliminates rows with all zero elements.

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 $K=\min_{s,t,j}(p_{\boldsymbol{C_sC_j}}+p_{\boldsymbol{C_tC_j}})s_2$, we have

$$||\mathbf{\Pi_0}-\tilde{\mathbf{\Pi}}||_F^2 \leq \frac{2}{nK}||(\mathbf{\Pi_0}-\tilde{\mathbf{\Pi}})\hat{\mathbf{B}}||_F^2. \tag{3-25}$$

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

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

In the second part, note that

$$\begin{aligned}||(\mathbf{\Pi_0}-\tilde{\mathbf{\Pi}})\hat{\mathbf{B}}||_F &= ||\hat{\mathbf{B}}(\mathbf{\Pi_0}-\tilde{\mathbf{\Pi}})^{\mathbf{T}}||_F = ||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}(\mathbf{\Pi_0}-\tilde{\mathbf{\Pi}})^{\mathbf{T}}||_F \\ &\leq ||(\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi}_0^T-\hat{\mathbf{A}})-(\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^T-\hat{\mathbf{A}})||_F \\ &\leq ||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}-\hat{\mathbf{A}}||_F+||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}}-\hat{\mathbf{A}}||_F,\end{aligned}$$

where the second equation holds since the permutation matrix $\tilde{\mathbf{\Pi}}$ keeps invariant of Frobenius norm, and the second inequality holds due to the triangular inequality of Frobenius norm. Then we obtain

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

For the term $||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi_0^T} - \hat{\mathbf{A}}||_F^2$,

$$\begin{aligned}
&||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi_0^T} - \hat{\mathbf{A}}||_F^2 = \mathbf{tr}((\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi_0^T} - \hat{\mathbf{A}})^{\mathbf{T}}(\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi_0^T} - \hat{\mathbf{A}})) \\
&= \mathbf{tr}(\hat{\mathbf{A}}^{\mathbf{T}}\hat{\mathbf{A}}) + \mathbf{tr}(\hat{\mathbf{B}}^{\mathbf{T}}\hat{\mathbf{B}}) - 2\mathbf{tr}(\mathbf{\Pi_0}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}\hat{\mathbf{A}}) \\
&= ||\hat{\mathbf{A}}||_F^2 + ||\hat{\mathbf{B}}||_F^2 - 2\mathbf{tr}(\mathbf{\Pi_0}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}\hat{\mathbf{A}}) \\
&= \frac{1}{2}(||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}} - \hat{\mathbf{A}}||_F^2 + ||\mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0^T} - \hat{\mathbf{A}}||_F^2) \\
&\quad + \mathbf{tr}(\mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0^T}\hat{\mathbf{A}}) + \mathbf{tr}(\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}\hat{\mathbf{A}}) - 2\mathbf{tr}(\mathbf{\Pi_0}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}\hat{\mathbf{A}}) \\
&\leq ||\mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0}^T - \hat{\mathbf{A}}||_F^2 + \mathbf{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}})\hat{\mathbf{A}}),
\end{aligned} \tag{3-26}$$

where the last equation can be verified by the first three equations, and the last inequality holds since $||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}} - \hat{\mathbf{A}}||_F^2 \leq ||\mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0^T} - \hat{\mathbf{A}}||_F^2$.

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 $\mathbf{Z} = (\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}}\hat{\mathbf{A}}$. Now we focus on $\mathbf{tr}(\mathbf{Z})$. Note that the i_{th} row of $\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}$ is composed of either zeros if $\tilde{\mathbf{\Pi}}$ and $\mathbf{\Pi_0}$ map node i in G_2 to the same node in G_1, or zeros expect one 1 and one -1 if $\tilde{\mathbf{\Pi}}$ and $\mathbf{\Pi_0}$ map node i in G_2 to different nodes in G_1. 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 $s \neq t$. 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 $\mathbf{Z} = \mathbf{YX}$. Then 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 number of the $1(-1)$ in the i_{th} column of $\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}$, when $\tilde{\mathbf{\Pi}}$ and $\mathbf{\Pi_0}$ map node i in G_2 into different nodes in G_1. If they map node j in G_2 to the same node in G_1, then we set $\mathbf{X}_{ji} = 0$. Therefore for a single value on the diagonal of

$\mathbf{Z}$, i.e., $\mathbf{Z}_{ii}$, we can bound its absolute value as

$$\begin{aligned}|\mathbf{Z}_{ii}| &= |\langle \mathbf{Y_{i\cdot}}, \mathbf{X_{\cdot i}} \rangle| \leq ||\mathbf{Y_{i\cdot}}||_F ||\mathbf{X_{\cdot i}}||_F \\ &\leq n \max_k |\hat{\mathbf{B}}_{sk} - \hat{\mathbf{B}}_{tk}| \max_\ell |\hat{\mathbf{A}}_{p_\ell \ell} - \hat{\mathbf{A}}_{q_\ell \ell}|.\end{aligned} \tag{3-27}$$

Taking the expectation of $\mathbf{A}$ and $\mathbf{B}$ on both sides of Inequality (3-27), we can obtain that

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

based on the Jensen's Inequality. Hence $|\hat{\mathbf{B}}_{sk} - \hat{\mathbf{B}}_{tk}| = |\sqrt{w_{sk}}\mathbf{B}_{sk} - \sqrt{w_{tk}}\mathbf{B}_{tk}| \leq |\mathbf{B}_{sk} - \mathbf{B}_{tk}|$, and $|\hat{\mathbf{A}}_{p_\ell \ell} - \hat{\mathbf{A}}_{q_\ell \ell}|$ is similar. Hence

$$|\mathbf{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}})\hat{\mathbf{A}})| \leq n \max_i |\langle \mathbf{Y_{i\cdot}}, \mathbf{X_{\cdot i}} \rangle| \leq n^2 L. \tag{3-28}$$

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

Upon completion of the former three parts, now we can move to the final part. Specifically, from Inequalities (3-25), (4-41) and (3-28), we can obtain

$$\begin{aligned}||(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})||_F^2 &\leq \frac{2}{nK}||(\mathbf{\Pi_0} - \tilde{\mathbf{\Pi}})\hat{\mathbf{B}}||_F^2 \\ &\leq \frac{8}{nK}||\mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0^T} - \hat{\mathbf{A}}||_F^2 + 2\mathbf{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\hat{\mathbf{B}}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}})\hat{\mathbf{A}}) \\ &\leq \frac{8}{nK}||\mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0^T} - \hat{\mathbf{A}}||_F^2 + \frac{4nL}{K}.\end{aligned}$$

Since $\tilde{\mathbf{\Pi}}$ is the minimizer of $||\hat{\mathbf{A}} - \mathbf{\Pi}\hat{\mathbf{B}}\mathbf{\Pi}^{\mathbf{T}}||_F^2$ and $||\hat{\mathbf{A}} - \mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0^T}||_F^2 / ||\hat{\mathbf{A}} - \tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}||_F^2 = \Omega(1)$ holds, there exists a constant $\tilde{c} \geq 1$ such that $||\hat{\mathbf{A}} - \mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0}^{\mathbf{T}}||_F \leq \tilde{c}||\hat{\mathbf{A}} - \tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}||_F$. Therefore since $||\mathbf{\Pi_0}||_F^2 = 2n$ and the first and third condition, we can bound the relative NME when $n \to \infty$ as:

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

This completes our proof. □

Theorem 3.12 demonstrates that under certain conditions, the relative NME goes to 0 when the size of network tends to be infinity. Although this result does not show that the NME, expressed as $||\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}||_F^2$, vanishes under the conditions, it shows that compared with the number of nodes in the network, the NME can be neglected when the size of network is very large. This phenomenon makes sense in de-anonymization since it demonstrates that by minimizing the weighted-edge matching problem (WEMP), we can neglect the NME in large social networks and map most of the nodes correctly.

The Positive Impact of Overlapping Communities on Theorem 3.12: Now we demonstrate that the overlapping communities exert a positive impact on diminishing the relative NME through making the conditions in Theorem 3.12 more prone to be satisfied. Specifically, when the overlapping strength in the networks becomes stronger, then the condition 3 is easier to be met. We claim that condition 3 is a decisive prerequisite for the vanish of relative NME, since conditions 2 and 4 are easy to meet by the common assumption that true mapping keeps invariant of the community representations and the additive constraint about communities, which we will discuss in Section 3.4.2. Therefore the overlapping strength holds a balance in vanishing the relative NME.

For convenience, we assume that $s = s_1 = s_2$ in the following setting. Note that when the correct mapping π_0 keeps invariant of community representations, then on average condition 3 can be written as

$$2 \sum_{1 \le i < j \le n} \log \left(\frac{1 - p_{\boldsymbol{C}_i \boldsymbol{C}_j}(2s - s^2)}{p_{\boldsymbol{C}_i \boldsymbol{C}_j}(1-s)^2} \right) p_{\boldsymbol{C}_i \boldsymbol{C}_j} s = o(Kn^2). \tag{3-29}$$

To characterize the global situation in the networks, we define an average probability $\hat{p}$ such that

$$\begin{aligned} &\sum_{1 \le i < j \le n} \log \left(\frac{1 - p_{\boldsymbol{C}_i \boldsymbol{C}_j}(2s - s^2)}{p_{\boldsymbol{C}_i \boldsymbol{C}_j}(1-s)^2} \right) p_{\boldsymbol{C}_i \boldsymbol{C}_j} s \\ &= \frac{n(n-1)}{2} \log \left(\frac{1 - \hat{p}(2s - s^2)}{\hat{p}(1-s)^2} \right) \hat{p} s, \end{aligned} \tag{3-30}$$

where $\hat{p}$ is positively correlated to the overlapping strength of the whole networks. Tak-

ing 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

$$\frac{d(\log\left(\frac{1-\hat{p}(2s-s^2)}{\hat{p}(1-s)^2}\right)\hat{p}s)}{d\hat{p}} = \log\left(\frac{1-\hat{p}(2s-s^2)}{\hat{p}(1-s)^2}\right)\hat{p}s) - \frac{1}{1-\hat{p}(2s-s^2)}, \tag{3-31}$$

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

$$\log\left(\frac{1-\hat{p}(2s-s^2)}{\hat{p}(1-s)^2}\right)\hat{p}s = \log\left(1+\frac{1-\hat{p}}{\hat{p}(1-s)^2}\right)\hat{p}s \sim \frac{1-\hat{p}}{(1-s)^2}s = o(1), \tag{3-32}$$

where $s = \Omega 1$. Therefore if $\hat{p}$ is asymptotically close to 1 as the overlapping strength enhances, then the order of $||\hat{\mathbf{A}} - \mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0^T}||_F^2$ turns smaller, which is more prone to satisfy $||\hat{\mathbf{A}} - \mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi_0^T}||_F^2 = o(Kn^2)$.

Taking a vivid example of the overlapping stochastic block model (OSBM) in which

$$p_{\boldsymbol{C_i}\boldsymbol{C_j}} = \frac{1}{1+ae^{-x}}, \tag{3-33}$$

where a is an adjustable parameter and x is the number of overlapping communities. We find that $\min_{i,j} p_{\boldsymbol{C_i}\boldsymbol{C_j}} = \frac{1}{1+a}$ 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

$$\begin{aligned}\log\frac{1-\hat{p}(2s-s^2)}{p(1-s)^2}p &= \log(1+\frac{1-\hat{p}}{\hat{p}(1-s)^2})p \\ &\approx \frac{1-\hat{p}}{(1-s)^2} = o(\min_{i,j} p_{\boldsymbol{C_i}\boldsymbol{C_j}}) = o(1),\end{aligned} \tag{3-34}$$

thus condition 3 holds. Meanwhile note that $s = o(1)$ makes condition 1 hold as well. Therefore all the four conditions in Theorem 3.12 hold, thus the relative NME vanishes to 0.

3.4.2 Algorithm Design and Convergence Analysis

In Sections 3.3.3 and 3.4.1 we have verified the validity of the transformation from MMSE estimator to the weighted-edge matching problem (WEMP). In this section, we will propose an algorithm to solve WEMP and analyze its convergence.

Formulation of WEMP in Constrained Optimization Form

Before designing the algorithm, we first restate WEMP in the form of the following constrained optimization problem:

$$\begin{aligned} &\text{minimize} \|(\hat{\mathbf{A}} - \mathbf{\Pi}\hat{\mathbf{B}}\mathbf{\Pi}^{\mathbf{T}})\|_{\mathrm{F}}^2 \\ \textbf{s.t. } &\forall i \in V_1,\ \textstyle\sum_i \mathbf{\Pi}_{ij} = 1 && \text{(3-35)} \\ &\forall j \in V_2,\ \textstyle\sum_j \mathbf{\Pi}_{ij} = 1 && \text{(3-36)} \\ &\forall i, j,\ \mathbf{\Pi}_{ij} \in \{0, 1\}, && \text{(3-37)} \end{aligned}$$

Additionally, note that in previous sections we have assume that the true mapping between G_1 and G_2 should keep invariant of the community representation of every node before and after mapping. That is to say, the same user in G_1 and G_2 belongs to the same subset of communities, which is in line with real situations where there is no difference in communities in G_1 and G_2. To elaborate, let us recall Fig. 3.1, where the communities in G_1 and G_2 are with no differences since the number of communities are the same and the corresponding communities in two networks contain the same subset of users. Here we point out that we keep this assumption in our algorithm design. Therefore, in order to obtain the correct mapping π_0, another constraint about community representation should be added, which is

$$\forall i \in V_1, \boldsymbol{C_i} = \boldsymbol{C_{\pi(i)}}. \tag{3-38}$$

Eqn. 3-38 means that our estimated mapping π should keep the community rep-

resentation of all the nodes in V_1 unchanged before and after mapping. Note that it is hard to implement this constraint directly in the optimization problem since it is not in the form of permutation matrix. However, we can easily convert it into a suitable one by defining a new matrix to characterize the community representation of all the nodes, which we call as "Community Representation Matrix", denoted as $\mathbf{M}$. Its formal definition is as follows.

Definition 3.13. *(**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.*

Take Fig. 3.1 as an instance again. The community representation matrix of G, denoted as $\mathbf{M_G}$, satisfies

$$\mathbf{M_G^T} = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

Note that the community representation matrices for G, G_1 and G_2 are identical. So we set all of them to be $\mathbf{M}$. Hence the constraint (3-38) can be rewritten as $||\mathbf{\Pi M} - \mathbf{M}||_F^2 = 0$. According to optimization theory, we can form this constraint into the objective function by regarding it as the penalty term and obtain a new objective function

$$F_0(\mathbf{\Pi}) = ||\mathbf{\hat{A}} - \mathbf{\Pi \hat{B} \Pi^T}||_{\mathrm{F}}^2 + \mu||\mathbf{\Pi M} - \mathbf{M}||_F^2,$$

where μ is an adjustable penalty parameter, which is large enough such that when the objective function reaches its minimum value, $||\mathbf{\Pi M} - \mathbf{M}||_F^2$ is exactly or very close to 0. Note that this transformation of objective function does not affect the previous analytical results of WEMP since we have the assumption that the true mapping keeps invariant of the community representation of every single node before and after mapping. Then with the aim of finding the true permutation matrix $\mathbf{\Pi_0}$, we must have $||\mathbf{\Pi_0 M} - \mathbf{M}||_F^2 = 0$, thus the objective function is the same as that of WEMP.

Problem Relaxation and Idea of Algorithm Design

Hereinafter, we focus on how we design our algorithm targeting the WEMP.

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 [22] 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.

Implementation of CCOM and Algorithm Design

Although [22] has proposed the general framework of CCOM, the way it presents to obtain $F_1(\mathbf{\Pi})$ and $F_2(\mathbf{\Pi})$ is rather complex, as it involves Kronecker product and the Laplacian matrix of graphs. Here we provide a simple way, as defined in Lemma 3.4.2, to get the convex relaxation and concave relaxation, for simplifying the objective function compared with that in [22].

Lemma 3.4.2. *A proper way to get the convex relaxation and concave relaxation is*

$$F_1(\mathbf{\Pi}) = F_0(\mathbf{\Pi}) + \frac{\lambda_{min}}{2}(n - ||\mathbf{\Pi}||_F^2);$$

$$F_2(\mathbf{\Pi}) = F_0(\mathbf{\Pi}) + \frac{\lambda_{max}}{2}(n - ||\mathbf{\Pi}||_F^2).$$

Therefore 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 $\xi = (1-\alpha)\lambda_{min} + \alpha\lambda_{max}$, $\xi \in [\lambda_{min}, \lambda_{max}]$.

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[10]. 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.

Meanwhile, one of the sufficient and necessary conditions for a function whose variable is matrix is concave is that the Hessian matrix of this function is negative semi-definite. Similar to the analysis of $F_1(\mathbf{\Pi})$, we can verify that $F_2(\mathbf{\Pi})$ is a concave function. Thus we complete the proof. □

Lemma 3.4.2 presents a simple way to implement CCOM algorithmically, since $F_0(\mathbf{\Pi})$ is just our objective function in Section 3.4.2 and $||\mathbf{\Pi}||_F^2$ can be computed easily.

[10]The identity matrix I means all the elements on the diagonal of I are all 1s while others are all 0s. Note that here I is an $n^2 \times n^2$ matrix since the first order derivative of a function whose variable is a matrix is a $n \times n$ matrix, thus the second derivative of F_0 (F_1) is $n^2 \times n^2$ matrix.

We can modify $F(\mathbf{\Pi})$ step by step from a convex function to a concave function by modifying the value of ξ or α. In the following analysis, we set $F_\xi(\mathbf{\Pi})$ equivalent to $F(\mathbf{\Pi})$ since ξ is an adjustable parameter in $F(\mathbf{\Pi})$.

A vivid example of the CCOM under the formulation of $F_\xi(\mathbf{\Pi})$ by Lemma 3.4.2 is illustrated in Fig. 3.5. As can be seen in the figure, when ξ starts at λ_{min}, $F_\xi(\mathbf{\Pi})$ is a convex function, thus we can obtain the minimizer of this objective function. After we find the minimizer, we modify α to be 0.2, thus $\xi = 0.8\lambda_{min} + 0.2\lambda_{max}$, which makes the objective function become less convex. To obtain the minimizer of this new objective function, we have the prior knowledge of the previous minimizer, and since we only slightly modify the objective function, the optimal solution of new objective function should not deviate much from the previous one intuitively. Therefore we can start from the previous minimizer to find the new minimizer. Gradually, as α becomes increasingly larger, the objective function tends to be concave while the minimizer of it tends to get close to the boundary, on which the optimal solution of the original WEMP exists. The trail for the minimizer can be referred to the red line with arrows in Fig. 3.5.

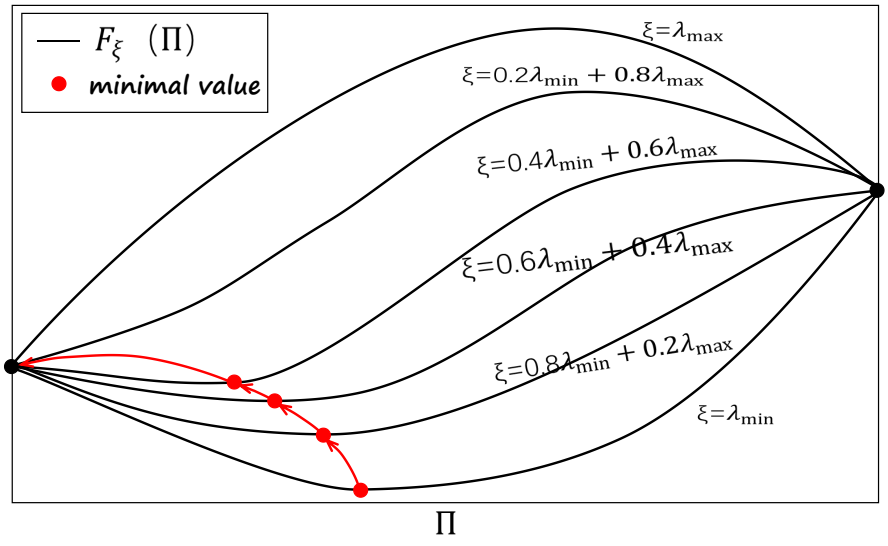


Figure 3.5 An Illustration of the Implementation of CCOM by Lemma 3.4.2.

Based on the above analysis, we propose Algorithm 3.1 as our main algorithm for the weighted-edge matching problem (WEMP) under CCOM. We call Algorithm 3.1 *Convex-concave Based De-anonymization Algorithm (CBDA)*. 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 3.1: 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 $\mathbf{X}^\perp \in \Omega$. //Finding the optimal descent direction
6: $\gamma_k \leftarrow \arg\min_\gamma F(\mathbf{\Pi_k} + \gamma(\mathbf{X}^\perp - \mathbf{\Pi_k}))$, where $\gamma_k \in [0, 1]$. //Finding the optimal step size
7: $\mathbf{\Pi_{k+1}} \leftarrow \mathbf{\Pi_k} + \gamma_k(\mathbf{X}^\perp - \mathbf{\Pi_k})$, $k \leftarrow k + 1$. //Estimation Update
8: **end while**
9: $\xi \leftarrow \xi + \Delta\xi$.
10: **end while**

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 $O\left(\frac{n^6 T\xi}{\Delta\xi}\right)$. As far as we know, a dearth of algorithmic analysis of seedless de-anonymization exists except for [23,24], with their proposed algorithm sharing identical complexity of $O(n^6)$ with ours.

Convergence: There are two loops in CBDA and we provide convergence analysis on them respectively. Before that, we first clarify that:

- We set $\mathbf{\Pi_k}$ as the estimation after k rounds in the inner loop, thus $\mathbf{\Pi_{k+1}}$ is the estimation after $k+1$ rounds in the inner loop and $\mathbf{\Pi_{k+1}} = \mathbf{\Pi_k} + \gamma_k(\mathbf{X}_\perp - \mathbf{\Pi_k})$.
- We set $F_\xi(\mathbf{\Pi}) = F_0(\mathbf{\Pi}) + \xi(n - ||\mathbf{\Pi}||_F^2)$ and $\mathbf{\Pi}^\xi$ as the minimizer of $F_\xi(\mathbf{\Pi})$. Thus $F_{\xi+\Delta\xi}(\mathbf{\Pi}) = F_0(\mathbf{\Pi}) + (\xi+\Delta\xi)(n - ||\mathbf{\Pi}||_F^2)$ and $\mathbf{\Pi}^{\xi+\mathbf{\Delta\xi}}$ is the minimizer of $F_{\xi+\Delta\xi}(\mathbf{\Pi})$.

Then we propose Lemma 3.4.3 to discuss the convergence of CBDA.

Lemma 3.4.3. *CBDA converges and the final output is a permutation matrix in the original feasible region* Ω_0.

Proof. As stated above, showing the convergence of CBDA is equivalent to showing the convergence of both inner and outer loops.

1. Inner Loop: We focus on $F_\xi(\mathbf{\Pi_{k+1}})$ and $F_\xi(\mathbf{\Pi_{k+1}})$. Since $\mathbf{\Pi_{k+1}} = \mathbf{\Pi_k} + \gamma_k(\mathbf{X}_\perp - \mathbf{\Pi_k})$, according to Taylor's Theorem,

$$\begin{aligned} F_\xi(\mathbf{\Pi_{k+1}}) &= F_\xi(\mathbf{\Pi_k} + \gamma_k(\mathbf{X}_\perp - \mathbf{\Pi_k})) \\ &= F_\xi(\mathbf{\Pi_k}) + \gamma_k \mathbf{tr}(\nabla F_\xi^T(\mathbf{\Pi_k})(\mathbf{X}^\perp - \mathbf{\Pi_k})) + \gamma_k \mathbf{R_k} \\ &\leq F_\xi(\mathbf{\Pi_k}) + \gamma_k \mathbf{tr}(\nabla F_\xi^T(\mathbf{\Pi_k})(\mathbf{\Pi}^\xi - \mathbf{\Pi_k})) + \gamma_k \mathbf{R_k}, \end{aligned} \tag{3-39}$$

where $\gamma_k \mathbf{R_k}$ is the remainder of this Taylor series, and this form makes sense since the remainder must contain a multiplicative factor of γ_k. The last inequality holds since $\mathbf{X}^\perp$ is the minimizer of $\mathbf{tr}(\nabla F_\xi^T(\mathbf{\Pi_k})(\mathbf{\Pi}^\xi - \mathbf{\Pi_k}))$.

In terms of $F_\xi(\mathbf{\Pi}^\xi)$, we have

$$\begin{aligned} F_\xi(\mathbf{\Pi}^\xi) &= F_\xi(\mathbf{\Pi_k} + \mathbf{\Pi}^\xi - \mathbf{\Pi_k}) \\ &= F_\xi(\mathbf{\Pi_k}) + \mathbf{tr}(\nabla F_\xi^T(\mathbf{\Pi_k})(\mathbf{\Pi}^\xi - \mathbf{\Pi_k})) + \mathbf{R'_k}, \end{aligned} \tag{3-40}$$

where $\mathbf{R'_k}$ is the remainder of this Taylor series.

Combining Eqn. (3-39) and (3-40), we can obtain

$$F_\xi(\mathbf{\Pi_{k+1}}) \leq F_\xi(\mathbf{\Pi_k}) + \gamma_k(F_\xi(\mathbf{\Pi}^\xi) - F_\xi(\mathbf{\Pi_k})) + \gamma_k(\mathbf{R_k} - \mathbf{R'_k}). \tag{3-41}$$

Denote $\Delta\mathbf{R_k} = \mathbf{R_k} - \mathbf{R'_k}$ and by simple transformation of Inequality (3-41), we obtain

$$F_\xi(\mathbf{\Pi_{k+1}}) - F_\xi(\mathbf{\Pi}^\xi) \leq (1-\gamma_k)(F_\xi(\mathbf{\Pi_k}) - F_\xi(\mathbf{\Pi}^\xi)) + \gamma_k \Delta\mathbf{R_k}. \tag{3-42}$$

Note that Inequality (3-42) builds up the relationship between $F_\xi(\mathbf{\Pi_{k+1}})$ and $F_\xi(\mathbf{\Pi_k})$, and we obtain

$$\begin{aligned} &F_\xi(\mathbf{\Pi_{k+1}}) - F_\xi(\mathbf{\Pi}^\xi) \\ &\leq \prod_{i=1}^{k}(1-\gamma_i)(F_\xi(\mathbf{\Pi_1}) - F_\xi(\mathbf{\Pi}^\xi)) + \sum_{i=1}^{k}\gamma_i \prod_{j=1}^{k-i}(1-\gamma_j)\Delta\mathbf{R}_i. \end{aligned} \tag{3-43}$$

For $F_\xi(\mathbf{\Pi_1}) - F_\xi(\mathbf{\Pi}^\xi)$, note that $\mathbf{\Pi_1} = \mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}$, then

$$\begin{aligned} F_\xi(\mathbf{\Pi}^\xi) &= F_0(\mathbf{\Pi}^\xi) + \xi(n - ||\mathbf{\Pi}^\xi||_F^2) \\ &= F_0(\mathbf{\Pi}^\xi) + (\xi - \Delta\xi)(n - ||\mathbf{\Pi}^\xi||_F^2) - \Delta\xi(n - ||\mathbf{\Pi}^\xi||_F^2) \\ &\geq F_0(\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}) + (\xi - \Delta\xi)(n - ||\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}||_F^2) \\ &\quad - \Delta\xi(n - ||\mathbf{\Pi}^\xi||_F^2) \\ &= F_0(\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}) + \xi(n - ||\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}||_F^2) \\ &\quad + \Delta\xi(||\mathbf{\Pi}^\xi||_F^2 - ||\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}||_F^2) \\ &= F_\xi(\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}) + \Delta\xi(||\mathbf{\Pi}^\xi||_F^2 - ||\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}||_F^2). \end{aligned} \tag{3-44}$$

Hence

$$F_\xi(\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}) - F_\xi(\mathbf{\Pi}^\xi) \leq \Delta\xi(||\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}||_F^2 - ||\mathbf{\Pi}^\xi||_F^2). \tag{3-45}$$

Therefore by combining Inequalities (3-45) and (3-43), we can obtain if $\Delta\xi$ is small enough, or if $k \to \infty$, then the term $\prod_{i=1}^{k}(1-\gamma_i)(F_\xi(\mathbf{\Pi_1}) - F_\xi(\mathbf{\Pi}^\xi))$ in last expression of Inequality (3-43) goes to 0.

For the second term $\sum_{i=1}^{k}\gamma_i\prod_{j=1}^{k-i}(1-\gamma_j)\Delta\mathbf{R}_i$, we note that when $k \to \infty$, then $\forall\epsilon > 0, \exists K > 0, \delta_1 > 0$, when $i > K$, $\gamma_i\prod_{j=1}^{k-i}(1-\gamma_j) < \gamma_i < \frac{\epsilon}{2^{\delta_1 i}}$, and meanwhile when $i \leq K$, $\gamma_k(1-\gamma_j) < \prod_{j=1}^{k-i}(1-\gamma_j) < \frac{\epsilon}{2^{\delta_2 i}}$. Setting $\delta^* = \min\{\delta_1, \delta_2\}$, then we can upper bound the sum $\sum_{i=1}^{k}\gamma_i\prod_{j=1}^{k-i}(1-\gamma_j)\Delta\mathbf{R}_i \leq \sum_{i=1}^{\infty}\frac{\epsilon}{2^{\delta i}} = 0$. Therefore we prove

that the inner loop converges.

2. Outer Loop: Note that from Eqn. (3-45), we know $(||\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}||_F^2 - ||\mathbf{\Pi}^{\xi}||_F^2)$ is nonnegative since $\Delta\xi > 0$ and $\mathbf{\Pi}^{\xi}$ is the minimizer of $F_{\xi}(\mathbf{\Pi})$. Thus $||\mathbf{\Pi}^{\xi}||_F^2 \leq ||\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}||_F^2$. Note that for all the $\mathbf{\Pi} \in \Omega$, the maximum value of $||\mathbf{\Pi}||_F^2$ is n, and the maximizer is in Ω_0. Therefore $||\mathbf{\Pi}||_F^2 - n \leq 0$. From Inequality (3-44), we find that

$$\begin{aligned} F_{\xi}(\mathbf{\Pi}^{\xi}) &\geq F_0(\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}) + (\xi - \Delta\xi)(n - ||\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}||_F^2) - \Delta\xi(n - ||\mathbf{\Pi}^{\xi}||_F^2) \\ &= F_{\xi-\Delta\xi}(\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}) - \Delta\xi\mathbf{tr}(||\mathbf{\Pi}^{\xi}||_F^2 - n). \end{aligned}$$

Therefore

$$\begin{aligned} |F_{\xi}(\mathbf{\Pi}^{\xi}) - F_{\xi-\Delta\xi}(\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi})| &\leq \Delta\xi|||(\mathbf{\Pi}^{\xi})||_F^2 - n| \leq \Delta\xi|||\mathbf{\Pi}^{\xi-\mathbf{\Delta}\xi}||_F^2 - n| \\ &\leq \Delta\xi|||\mathbf{\Pi}^{\xi_0}||_F^2 - n| \leq \Delta\xi(n-1), \end{aligned}$$

where the third inequality holds since $\mathbf{\Pi}^{\xi_0}$ is the minimizer of $F_{\lambda_{min}}(\mathbf{\Pi})$, i.e., the convex relaxation of $F_0(\mathbf{\Pi})$, and the fourth inequality holds since $\min_{\mathbf{\Pi}\in\Omega} ||\mathbf{\Pi}||_F^2 = 1$ and $\mathbf{\Pi} = \mathbf{1}_{n\times n}./n$ is the minimizer. Therefore, the analysis tells us if $\Delta\xi = o\big(\frac{1}{n}\big)$, then we can ensure that the outer loop converges.

Combining the convergence analysis of both inner and outer loops above, we complete the proof of the convergence of CBDA. □

Lemma 3.4.3 shows that CBDA can exactly find $\tilde{\mathbf{\Pi}}$, the minimizer of the objective function $F_0(\mathbf{\Pi})$, meanwhile ensuring that CBDA can perfectly solve WEMP, which vanishes the relative NME under mild conditions (Recall Theorem 3.12). Therefore CBDA is an algorithmic approach for seedless de-anonymization with high feasibility and good performance, especially for networks with larger size.

3.5 Experimental Aspect of Social Network De-anonymization Problem

In this section, we utilize three datasets: synthetic networks, sampled real social networks and true cross-domain networks, to conduct the experimental validation in terms of our analytical results and the performance of our proposed algorithm CBDA. Before

we start, we need to clarify that our theoretical results are based on asymptotical analysis when the size of the network goes to infinity, thus it is hard to validate them under finite computability. However, we can also observe some expected phenomenons under networks with finite size. In our experiments, the number of nodes in cross-domain co-author networks is 3176, larger than previous work in [23, 24] which is 2093. The performance validation of algorithms for seedless de-anonymization on large-scale real social networks, adopted by studies on seeded de-anonymization, as far as we know, is still an open problem.

3.5.1 Experiment Setup

Before presenting our experimental results, we first introduce the basic experimental settings.

Main Parameters

We list our adjustable parameters involved in our experiments in Table 3.2. Three parameters are in need of further explanations:

(i) a. This is a parameter in the overlapping stochastic block model (OSBM) which determines the $p_{\boldsymbol{C}_i\boldsymbol{C}_j}$, the probability of edge existence between nodes i and j in underlying graph. Specifically, $p_{\boldsymbol{C}_i\boldsymbol{C}_j}$ can be expressed as[11]

$$p_{\boldsymbol{C}_i\boldsymbol{C}_j} = \frac{1}{1+ae^{-x}}, \tag{3-46}$$

where x is the number of communities that both nodes i and j belong to. Note that $p_{\boldsymbol{C}_i\boldsymbol{C}_j}$ increases as x rises, which corresponds to the real case that nodes with more overlapping communities are more possibly related. Meanwhile, if a becomes larger (smaller), then $p_{\boldsymbol{C}_i\boldsymbol{C}_j}$ is smaller (larger) so that the graph becomes sparser (denser).

(ii) η. This is the community ratio. It means the ratio between the number of communities and nodes. This ratio reflects the fact that when the size of network becomes larger, the number of communities also increases. In performance validation of CBDA

[11]Note that this expression is equivalent to that in [18], though their forms are different.

Table 3.2 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
a	OSBM Parameter	{3, 5, 7, 9}
η	Community Ratio	{0.05, 0.1}
OL/NOL	Overlapping or Non-Overlapping	{OL, NOL}

we set $\eta = 0.05$ or 0.1, while when studying the influence of η on de-anonymization accuracy, it will be endowed with more values.

(iii) OL/NOL. OL means that communities are overlapping while NOL means not. This makes for illustrating the impact of the overlapping property of communities on the mapping accuracy.

Experimental Datasets

We discuss three adopted datasets in an order from model-based to real social networks.

1. Synthetic Networks: When we generate synthetic networks, there are two main steps: (i) randomly setting the community representation of every node and (ii) judging whether an edge exists between any two nodes. For step (i), since the nodes and communities in our model are both independently distributed, step (i) can be viewed as a Bernoulli trial for every node: Setting the probability that node i belongs to any one community as p_{c_i}, then the probability that node i belongs to k different communities is $p_{c_{ik}} = C_m^k p_{c_i}^k (1 - p_{c_i})^{m-k}$. In our experiment we set the same p_{c_i} for all nodes as we view them equally. For step (ii), we can set the probability of edge existence between any two nodes based on Eqn. (3-46)[12], with x determined by the community representation matrix (Recall Definition 3.13). In experiments on this dataset, we adjust the parameters based on Table 3.2 to validate the performance of our algorithm under different network settings.

2. Sampled Real Social Networks: In sampled real social networks, the underlying

[12]Unlike existing work [3] which determines the edge existence in the graph based on different distributions like Poisson, power law or exponential expected degree distributions, we strictly follow the OSBM and alter the edge distribution by modifying the parameter a.

Table 3.3 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

social network G is extracted from the dataset LiveJournal [17] without changes. The published and auxiliary networks (G_1 and G_2) are artificially sampled from G with the same probability s. To compare with the results in synthetic networks, we keep the settings of N, s and η as Table 3.2. However, since G is not generated from the OSBM, the OSBM parameter a does not exist anymore. In experiments on this dataset, we adjust N, s and η to characterize different possible situations based on real underlying networks.

3. 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 3.2) 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. In experiments on this dataset, the results accurately reflect the practical situations.

Algorithms for Comparison and Performance Metric

Note that the main point of our experiments is to show the influence of overlapping communities on the accuracy, and our algorithm can effectively harness this overlapping property. Therefore, 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

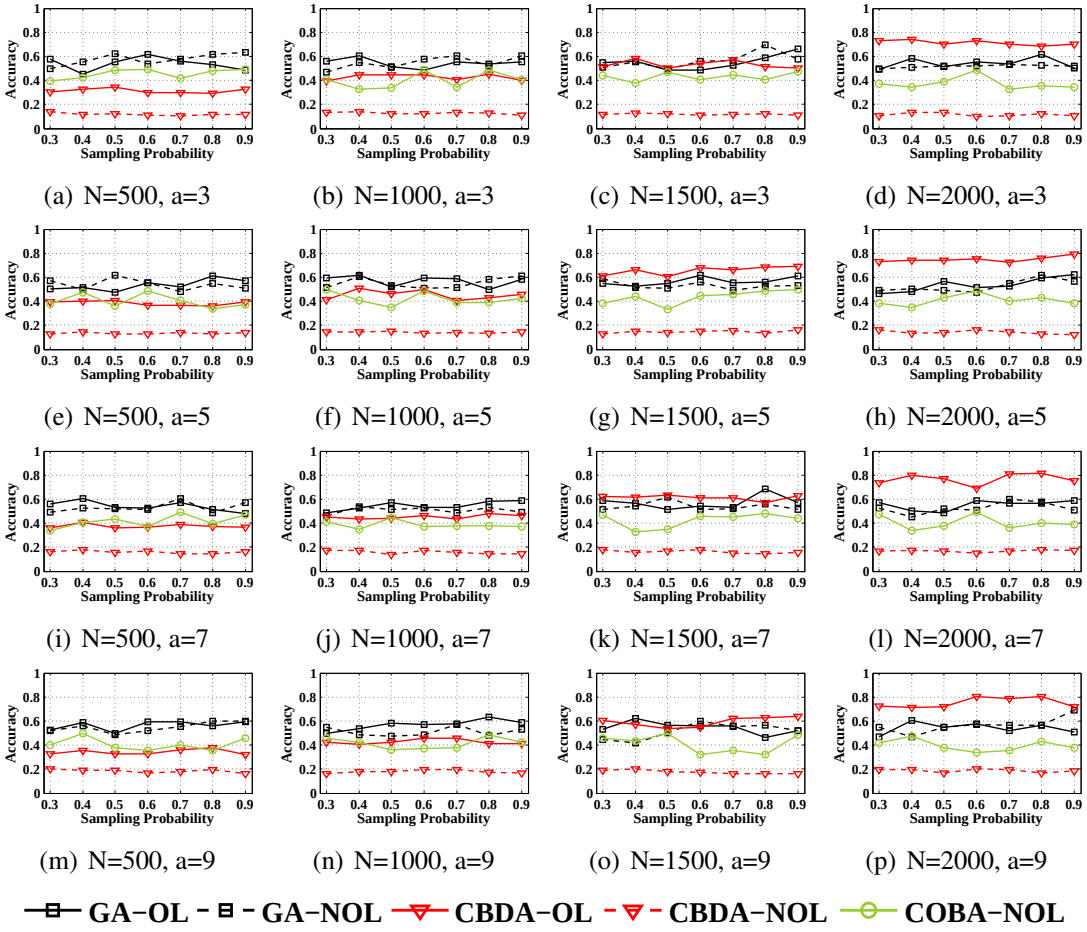


Figure 3.6 Experiments on Synthetic Networks with $\eta = 0.05$.

(GA), an epitome of heuristic algorithms, however due to its instability[13], we run 10 times and average these results as the accuracy of GA in every experiment; (ii) the Convex Optimization-Based Algorithm (COBA) in [23, 24], assigning a node to a unique community, which primarily suits non-overlapping cases. The performance metric is *accuracy*, the proportion of correctly mapped nodes.

Supplementary Experiments

To make our experimental validation more comprehensive and convincing, we supplement three experiments: (i) We study the effect of different community ratios (η) on the accuracy based on sampled real social networks. We modify η from 0.025 to 0.2

[13]The instability of GA will be shown in experimental results.

with interval 0.025; (ii) We study whether the weight matrix $\mathbf{W}$ in our cost function makes for the higher accuracy, compared with the cost function without appending $\mathbf{W}$ in existing work [3] . Appending $\mathbf{W}$ means adding the community information in the cost function. (iii) We study the instability of genetic algorithm (GA) and reveals the reason why GA lacks practical usage even if it achieves acceptable average accuracy in our main experiments.

3.5.2 Experiment Results

Synthetic Networks

Fig. 3.6 and 3.7 illustrate our experimental results on synthetic networks, where community ratio $\eta = 0.05$ in Fig. 3.6 and $\eta = 0.1$ in Fig. 3.7. Firstly looking at Fig. 3.6, with lower community ratio, we observe that: (i) The *average* accuracy of genetic algorithm (GA) under different settings keeps at levels around $40\% - 60\%$, which illustrates that based on OSBM, different sizes, densities and whether the communities overlap or not do not make a difference on the performance of GA averagely. This is because GA examines the edges one by one to make the cost function as small as possible, like a greedy algorithm which searches for the local optimum, therefore GA is not seriously affected by the global setting of the networks. (ii) The accuracy of COBA also keeps at a stable level in different situations. However, COBA can only cope with non-overlapping situations, and generally its performance is inferior to GA when communities are not overlapped, which is in line with the results in [23, 24]. (iii) The accuracy of CBDA, our algorithm, keeps stable under one specific situation but varies a lot in different networks when the communities overlap each other. This variation is mainly caused by the value of N. When the network size N becomes larger, the accuracy of CBDA rises up as well. Specifically, when N goes from 500 to 2000, the accuracy rises from approximately 40% to 80%. This striking phenomenon demonstrates that our CBDA is suitable for larger size of networks under networks with relatively sparse communities, which corresponds to our Theorem 3.12 that as the size of networks be-

comes larger, the relative NME becomes smaller[14]. On the other hand, however, when dealing with non-overlapping situations, our CBDA works stably but not as efficiently as GA or COBA, with the accuracy only around 20%.

Now we focus on Fig. 3.7 and compare it with Fig. 3.6. Fig. 3.7 shows the results under higher community ratio, i.e., denser communities. We can discover that the performance of GA follows that in Fig. 3.6, which makes sense since, as mentioned above, the performance of GA is not at the mercy of global information like community density. When communities are non-overlapping, the COBA and our CBDA keep similar trends as they do in lower η, showing that the community density under non-overlapping situations does not affect the performance of all these algorithms. However, what is noticeable is that our CBDA always performs better than other algorithms when the communities are overlapping each other. Moreover, compared with Fig. 3.6 in which η is low, the community parameter a is dominant in the accuracy of CBDA when the η is high, and when $a = 5$ the accuracy can keep stable at around 90%. This vivid comparison tells us that our CBDA is very suitable for high accuracy de-anonymization when the community density is large. Moreover, when the community density is large, the performance of CBDA is mainly decided by the edge density (a), positively correlated to community density; when the community density is small, then the performance of CBDA is mainly decided by the size of the networks (N). This shows that the community ratio (density) determines the dominant factor (a or N) in de-anonymization accuracy in networks with overlapping communities.

Sampled Real Social Networks

In sampled real social networks, we utilize the real underlying network, thus no modifications on a exist. The results are in Fig. 3.8. We can observe: (i) GA performs better in larger networks and under denser communities, either overlapping or non-overlapping; (ii) The performance of COBA is also enhanced when the size of networks become larger and the community becomes denser; (iii) The performance of CBDA under non-overlapping situations does not outperform other algorithms, but a rising

[14]Here when N is larger, the NME is smaller, thus the relative NME becomes smaller as well.

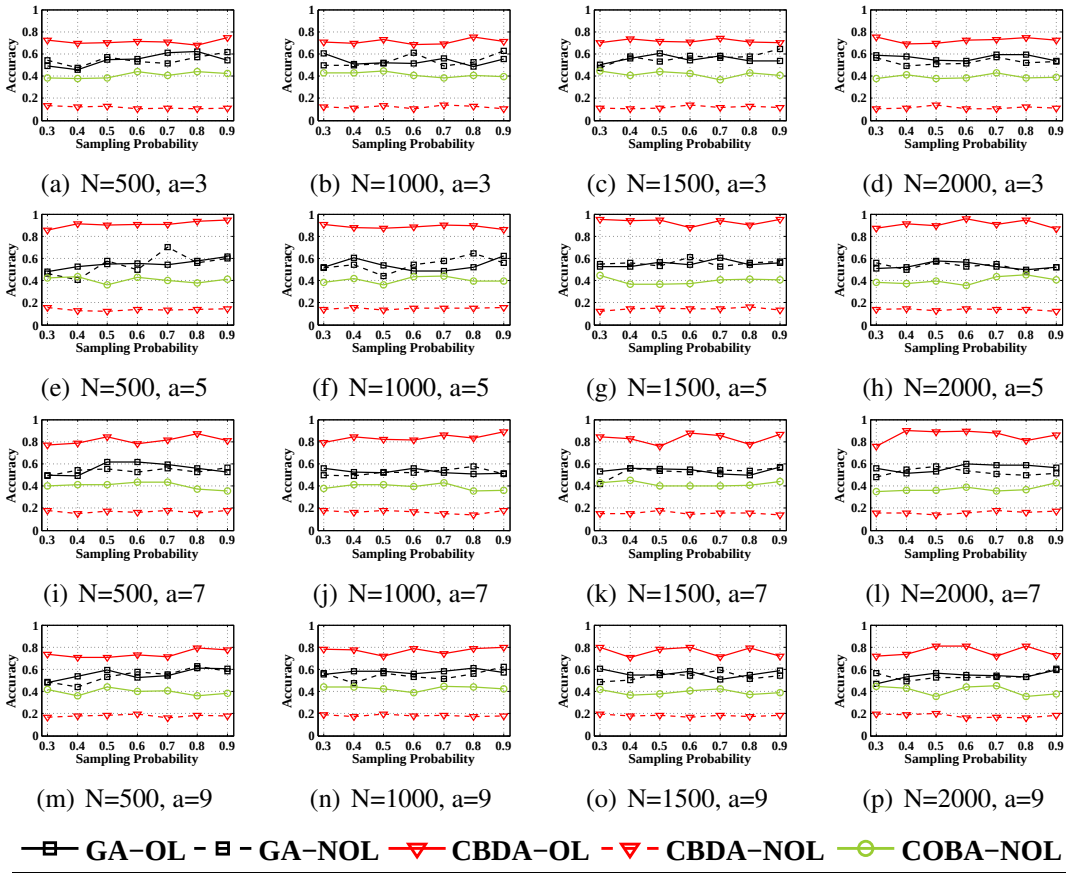


Figure 3.7 Experiments on Synthetic Network with $\eta = 0.1$.

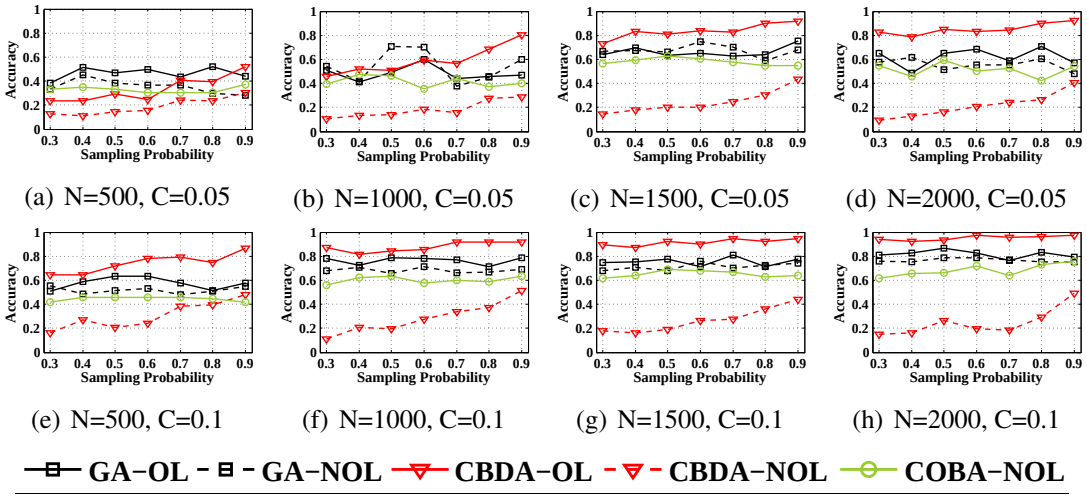


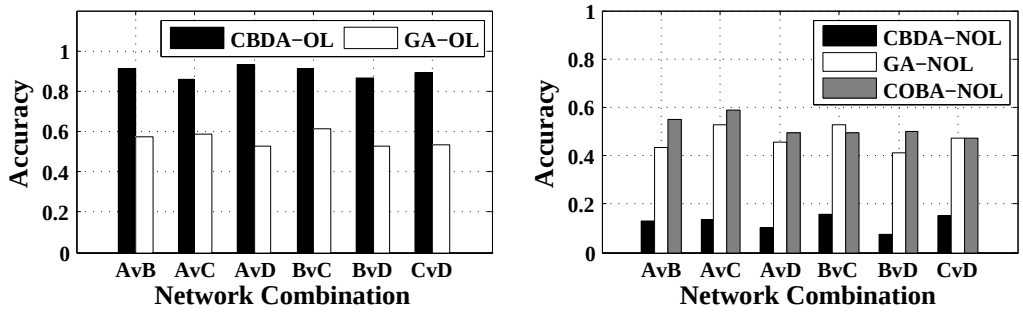
Figure 3.8 Experiments on Sampled Real Social Networks.

tendency exists as the sampling probability s becomes larger; (iv) The performance of CBDA under overlapping situations still performs well under denser communities and larger network size, with the highest point 95% and the highest average level around 90% when $N = 2000$ and $\eta = 0.1$, the largest size and densest communities in Table 3.3.

Synthesizing the above four observations, we can learn that the OSBM does not reflect the real social networks very precisely, since the performance of all three algorithms under non-overlapping or overlapping communities differs in two datasets. Moreover, with the same experimental setting, we discover that the performance of our CBDA is better in sampled real social networks than in OSBM-based synthetic networks, which further undergirds the high performance of our algorithm in practical use. Additionally, the results in Fig. 3.8 also meet Theorem 3.12 that as the network size becomes larger, the relative NME is much smaller and close to 0, indicating that Theorem 3.12 also works in real social networks.

Cross-Domain Co-author Networks

In cross-domain co-author networks, we pick up four networks with the same set of 3176 users. Fig. 3.9 illustrates our results. We find that in non-overlapping situation, the results correspond to those in previous datasets that our CBDA does not perform



(a) Overlapping (b) Non-Overlapping

Figure 3.9 Experiments on Cross-Domain Co-author Networks.

well, while GA and COBA work well. On the other hand, in overlapping situation, we find our CBDA reaches accuracy around 90%, outstripping GA whose accuracy is averagely 60%. This phenomenon places the significance of our CBDA in a higher level in de-anonymization with overlapping communities since it characterizes the real case totally. Moreover, due to the fact that overlapping situations are much more broadly in real large social networks than non-overlapping situations, our CBDA has wider usage than GA and COBA.

The Effect of Community Density

After presenting the results of three basic datasets, we further study the effect of community density on accuracy with more details by using our CBDA. Note that the community ratio η directly controls the community density, thus we apply the sampled real social networks under which we can adjust the community ratio η. We modify η from 0.025 to 0.2, with interval 0.025. The results are shown in Fig. 3.10. We can observe that in most cases our CBDA performs better when the network size is larger, which again echoes the conclusion in Theorem 3.12. Moreover, with the larger community ratio, the accuracy of CBDA rises up, showing that CBDA is suitable for social networks with highly overlapping communities. If we observe more carefully, the huge difference of accuracy occurs between $\eta = 0.025$ and $\eta = 0.075$, and when $\eta \geq 0.01$, the accuracy of CBDA under all the network sizes involved keeps at high levels, around 80% or higher. The results further illustrate that the higher community ratio η, the better

de-anonymizing result will be.

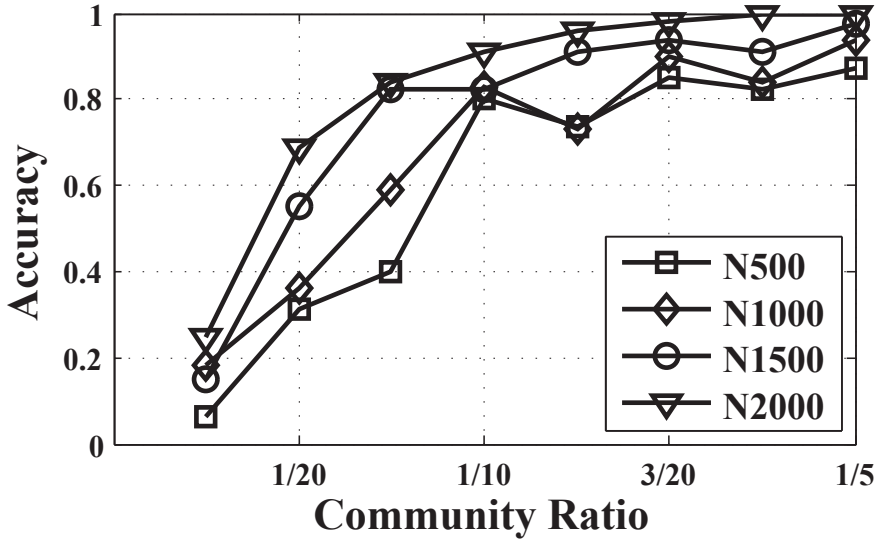
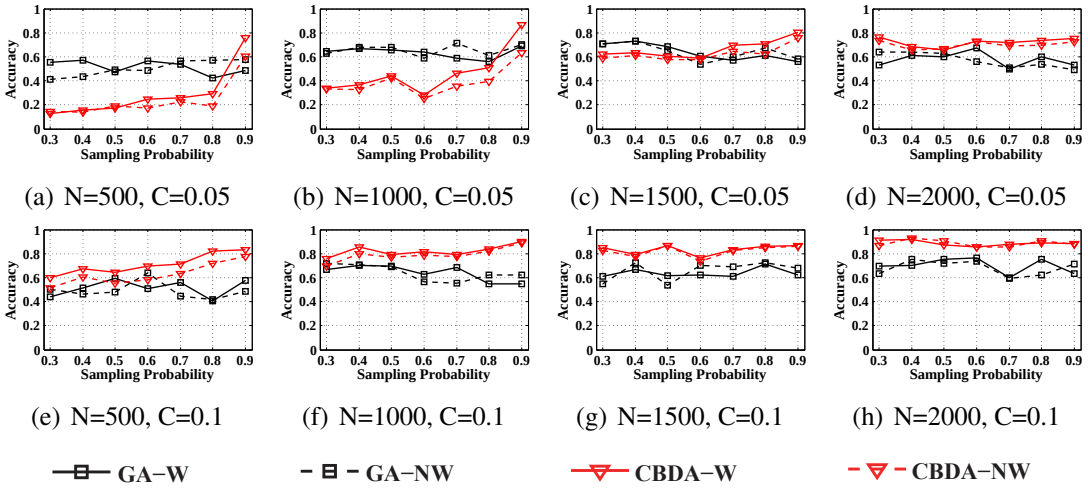


Figure 3.10 The Influence of Community Ratio on Accuracy.

The Effect of Weight Matrix W

In addition to previous experiments, we intend to supplement a study on the effect of weight matrix $\mathbf{W}$. The purpose of this study is to show that whether minimizing the cost function with $\mathbf{W}$ is of higher accuracy than minimizing the cost function without $\mathbf{W}$, proposed in [3]. Embedding $\mathbf{W}$ in the cost function means that We do this experiment



(a) N=500, C=0.05 (b) N=1000, C=0.05 (c) N=1500, C=0.05 (d) N=2000, C=0.05

(e) N=500, C=0.1 (f) N=1000, C=0.1 (g) N=1500, C=0.1 (h) N=2000, C=0.1

Figure 3.11 Experiments on Weighted and Non-weighted Cost Function.

under real sampled social networks. Fig. 3.11 illustrates the results. We can observe: (i) The performance of GA does not depend on whether the cost function is appended with $\mathbf{W}$. The curves under weighted and non-weighted cost functions interleave each other. This phenomenon, we suggest, is attributed to the instability of GA. (ii) The performance of our CBDA under weighted cost function is higher than that under non-weighted cost function in almost all the situations. One exception exists when $N = 2000$ and $\eta = 0.1$. In this situation two curves are almost overlapping each other, which tells us that in larger networks, embedding the community information in the cost function is less significant compared with the increasing network size. In smaller network size ($N \leq 1500$), however, the embedding of community information performs visible increment in accuracy.

The Instability of Genetic Algorithm

Now we discuss the weakness of GA in detail. Due to the fact that GA is a heuristic algorithm searching for a local minimum, we will obtain different results when trailing GA multiple times. Fig. 3.12 illustrates the results running GA for 10 times under real social networks with different sizes. Note that the performance of GA fluctuates violently, for example it swings from 30% to 84% when $N = 1000$ and from 42% to 80% when $N = 2000$. Therefore, although in average case GA keeps stable at around 40% to 60%, users who adopt GA cannot determine whether the solution GA outputs this time is of good or bad quality. This instability in output quality inhibits the usage of GA in practical situations.

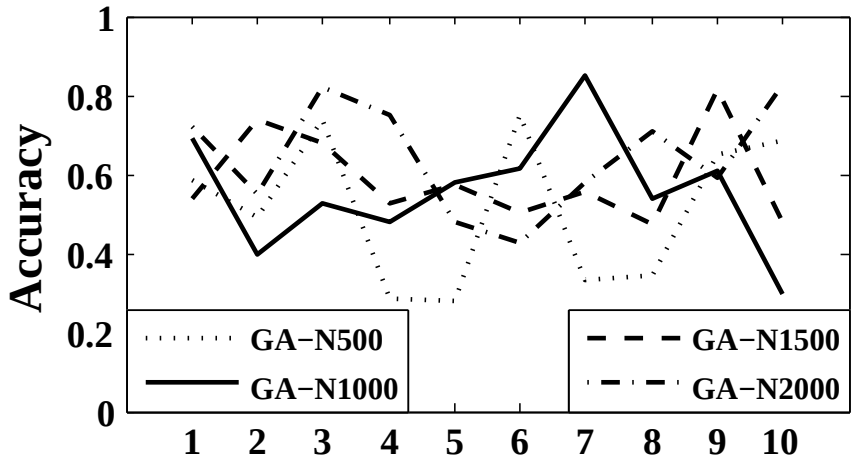


Figure 3.12 The Instability of Genetic Algorithm.

Chapter 4 Theoretical Bounds and Achievable Algorithm for De-anonymization in Different Network Models

4.1 De-anonymizability of Erdos-Renyi Graph Model

In this section, we will study the de-anonymizability in E-R graph. An E-R graph is denoted as $G(n,p)$, where n is the number of nodes and p is the edge existence probability between any pair of nodes. The edge existence is independent among different node pairs. Under E-R graph model, we will first target at the fully sampled situation, i.e. $s=1$, where both the published network G_1 and auxiliary network G_2 are identical, and then we will extend the results to the partially sampled situation, i.e. $s<1$, which corresponds the common de-anonymization problem in real cases.

4.1.1 Fully Sampled Situation: $s=1$

When G_1 and G_2 are fully sampled, they are isomorphic obviously, therefore the topology of G itself determines the de-anonymizability. Intuitively, when the nodes in G are of less similarity, it would be easier to de-anonymize due to the higher distinguishability of different nodes. For example, if the nodes in G are all of different degrees (from 0 to $n-1$), then it is trivial to de-anonymize G_1 by matching the node with same degree in G_1 to that in G_2. However, node degree is only one aspect of similarity. The kernel of the *similarity*, as we will show, is the *symmetry* among nodes in a network. Here we say two nodes are "symmetric" if and only if when we exchange these two nodes, the network does not change. If two nodes are symmetric, they can be viewed exactly the same, thus the node similarity between them is full and they are not de-anonymizable. Therefore, to characterize the conditions under which de-anonymization can be successful for fully sampled situation, the key problem is to characterize the probability of the existence of symmetric node pairs in G, denoted as P_{sym}. Hereinafter, we prove Theorem 4.1 and 4.2 to show the conditions under which P_{sym} asymptotically approaches 0 or not, indicating whether the network is de-anonymizable or not respectively. Inter-

estingly, we discover that the edge existence probability $p = \Omega(\frac{1}{n})$ plays as a threshold distinguishing whether de-anonymization will be successful in E-R model. Note that here we assume $p \leq \frac{1}{2}$ as for $p \geq \frac{1}{2}$, the situation is exactly the same by turning p to $1 - p$.

Before we give Theorem 4.1 and 4.2, we first provide an important lemma used in the proof of these two theorems.

Lemma 4.1.1. *Let* $T(a,k) = \binom{n-2}{a}\binom{n-2-a}{k-a}\binom{n-2-k}{k-a}$, *where* $a \in [0, \lfloor \frac{n}{4} \rfloor]$ *and* $k \in [\sqrt{an} - 1, \sqrt{(a+1)n} - 1] = S$. *Then we have:*

(i) If $a = o(n)$, *then*

$$K(a) = \max_{k \in S} T(a,k) = \frac{n}{a} \exp\left(\sqrt{an}(1 + o(1))\right) \tag{4-1}$$

(ii) If $a = \Theta(n)$, *let* $F(c) = \left(\frac{1}{c}\right)^{2c^2} \left(\frac{1}{c-c^2}\right)^{2(c-c^2)} \left(\frac{1}{1-c}\right)^{2(1-c)^2}$, *then*

$$K(a) = \max_{k \in S} T(a,k) = F^n(c) n^{\frac{1}{2} + \frac{1}{c} + o(1)} \tag{4-2}$$

Proof. Let $K(a) = T(a, k^*)$.

(i): $a = o(n)$

Note that

$$\begin{aligned} K(a) &= \frac{n!}{a!((k^* - a)!)^2 (n - 2k^* + a)!} \\ &\leq \frac{n!}{a!((\sqrt{an} - a - 1)!)^2 (n - 2\sqrt{(a+1)n} + a + 2)!} \\ &\leq \frac{n!}{a!((\sqrt{an} - a - 1)!)^2 (n - 2(\sqrt{a} + \frac{1}{2\sqrt{a}})\sqrt{n} + a + 2)!} \\ &\leq \frac{n^{2\sqrt{an} + \frac{\sqrt{n}}{\sqrt{a}} - a + 2}}{a!((\sqrt{an} - a)!)^2} \end{aligned} \tag{4-3}$$

Then we have

$$\begin{aligned}\ln K(a) \leq & (2\sqrt{an}+\frac{\sqrt{n}}{\sqrt{a}}-a+2)\ln n-\ln(a!)\\ & -2(\sqrt{an}-a)\ln(\sqrt{an}-a)-\ln(\sqrt{an}-a)+o(\ln n)\\ \leq & 2(\sqrt{an}-a)\ln\left(\frac{\sqrt{n}}{\sqrt{a}}+o\left(\frac{\sqrt{n}}{\sqrt{a}}\right)\right)+\left(a+2+\frac{\sqrt{n}}{\sqrt{a}}\right)\ln n\\ & -\ln a!+2(\sqrt{an}-a)-\frac{1}{2}\ln n+o(\ln n)\\ \leq & \left(\sqrt{an}+\frac{\sqrt{n}}{\sqrt{a}}+2\right)\ln n-(\sqrt{an}-a)\ln a+2\sqrt{an}\\ & -\ln a!+o(\sqrt{an})\\ \leq & (\sqrt{an}(1+o(1)))\ln\frac{n}{a}\end{aligned} \tag{4-4}$$

(ii): $a=\Theta(n)$

Let $a=c^2n$ where $c=(0,\frac{1}{2}]$. Then $K(a)\leq\frac{n^{2+\frac{1}{c}}n!}{(c^2n)!(((c-c^2)n)!)^2((1-c)^2n)!}$. According to Stirling's Formula, we can obtain

$$\begin{aligned}\ln K(a)\leq & -(c^2\ln c^2+2(c-c^2)\ln(c-c^2)+(1-c)^2\ln(1-c)^2)n\\ & +\left(\frac{1}{2}+\frac{1}{c}\right)\ln n+o(\ln n))\end{aligned} \tag{4-5}$$

Then we can obtain our conclusion.

□

Proposition 4.1.1. *For constant p, c and $\tilde{c}$ such that $0<p<1$ and $0<\tilde{c}<c<1$, then*

$$\frac{p^{2c}(1-p)^{2-c+\tilde{c}}(1-p^2)^{1-\tilde{c}}}{\tilde{c}^{\tilde{c}}(c-\tilde{c})^{2(c-\tilde{c})}(1-2c+\tilde{c})^{1-2c+\tilde{c}}}<1. \tag{4-6}$$

Theorem 4.1. *For an E-R graph $G(n,p)$, when $p=\omega(\frac{1}{n^{1-\epsilon}})$ where $\epsilon=\omega(\frac{1}{\log(n)})$, then G_1 can be successfully de-anonymized with probability 1 for the fully sampled situation.*

Proof. First we give some definitions used in the proof.

- ξ_{sym}: the set of symmetric node pairs;
- $d(i)$: the degree of node i;

- $d(i,j)$: the number of nodes both i and j are connected to.

Then for any node pair (i,j), we have

$$\begin{aligned}
&P((i,j)\in\xi_{sym}) = P(d(i)=d(j))P((i,j)\in\xi_{sym}|d(i)=d(j))\\
&=\sum_{k=0}^{n-1}P(d(i)=d(j)=k)P((i,j)\in\xi_{sym}|d(i)=d(j)=k)\\
&=\sum_{k=0}^{n-1}\sum_{t=0}^{k}P(d(i)=d(j)=k,d(i,j)=t)\times\\
&\quad P((i,j)\in\xi_{sym}|d(i)=d(j)=k,d(i,j)=t)\\
&\le\sum_{k=0}^{n-1}\sum_{t=0}^{k}P(A(i,j)=1)P(d(i)=d(j)=k,d(i,j)=t|A(i,j)=1)\\
&\quad+P(A(i,j)=0)P(d(i)=d(j)=k,d(i,j)=t|A(i,j)=0)\\
&=\sum_{k=0}^{n-1}\sum_{t=0}^{k}p\times P(d(i)=d(j)=k,d(i,j)=t|A(i,j)=1)\\
&\quad+(1-p)\times P(d(i)=d(j)=k,d(i,j)=t|A(i,j)=0).
\end{aligned}\tag{4-7}$$

In Eqn. (4-7), we utilize a necessary condition of the symmetric node pair that both nodes should have equal degree. Meanwhile, we take $d(i,j)$, the number of nodes both i and j are connected to, into consideration. It is of greater possibility for the symmetry of i and j than the different nodes connected by i and j, since whatever the degree of the common connected node is, if i and j are both connected to it, then they approach being symmetric directly, while for different connected nodes of i and j, we still need to check the symmetry of these connected nodes. Therefore, we consider the impact of common connected nodes in Eqn. (4-7).

Then we focus on $P(d(i)=d(j)=k,d(i,j)=t|A(i,j)=1)$. Note that under the condition $A(i,j)=1$, there will be $k-1$ nodes connecting i and j respectively.

Therefore, we have

$$\begin{aligned} X &= P(d(i) = d(j) = k, d(i,j) = t | A(i,j) = 1) \\ &= \binom{n-2}{t}(p^2)^t(1-p^2)^{(n-2-t)}\binom{n-2-t}{k-1-t} \\ &\quad p^{k-1-t}(1-p)^{n-1-k}\binom{n-k-1}{k-1-t}p^{k-1-t}(1-p)^{n-2k+t}. \end{aligned} \tag{4-8}$$

Similarly we have

$$\begin{aligned} Y &= P(d(i) = d(j) = k, d(i,j) = t | A(i,j) = 0) \\ &= \binom{n-2}{t}(p^2)^t(1-p^2)^{n-2-t}\binom{n-2-t}{k-t} \\ &\quad p^{k-t}(1-p)^{n-1-k}\binom{n-2-k}{k-t}p^{k-t}(1-p)^{n-2-2k+t}. \end{aligned} \tag{4-9}$$

Thus,

$$\begin{aligned} P((i,j) \in \xi_{sym}) &\leq \sum_{k=0}^{n-1}\sum_{t=0}^{k}(pX + (1-p)Y) \\ &= \sum_{k=0}^{n-2}\sum_{t=0}^{k}\binom{n-2}{t}\binom{n-2-t}{k-t}\binom{n-2-k}{k-t} \\ &\quad p^{2k}(1-p^2)^{n-2-t}(1-p)^{2n-3k+t-4} \\ &= \sum_{k=0}^{n-2}\sum_{t=0}^{k}Z(n,p,k,t)p^{2k}(1-p^2)^{n-2-t}(1-p)^{2n-3k+t-4} \end{aligned} \tag{4-10}$$

Let

$$C = \sum_{k=0}^{\frac{n-2}{2}}\sum_{t=0}^{k}Z(n,p,k,t)p^{2k}(1-p^2)^{n-2-t}(1-p)^{2n-3k+t-4}, \tag{4-11}$$

and

$$D = \sum_{k=\frac{n}{2}}^{n-2}\sum_{t=2k-n+2}^{k}Z(n,p,k,t)p^{2k}(1-p^2)^{n-2-t}(1-p)^{2n-3k+t-4}. \tag{4-12}$$

Then we discuss C and D respectively.

1. Upper Bound of C:

For $Z(n,p,k,t)$, we intend to find t^* which maximize it over all possible t. We

study the conditions:

$$\frac{Z(n,p,k,t+1)}{Z(n,p,k,t)} \leq 1, \quad \frac{Z(n,p,k,t-1)}{Z(n,p,k,t)} \leq 1.$$

We can obtain an unique t^*,

$$t^* = \left\lfloor \frac{(k+1)^2}{n} \right\rfloor \tag{4-13}$$

Let $Z(n,p,k,t^*) = H(t^*)$. Therefore,

$$\begin{aligned}
C &\leq \sum_{k=0}^{\frac{n-2}{2}} \sum_{t=0}^{k} H(t^*) p^{2k} (1-p)^{3n-3k-6} (1+p)^{n-2-t} \\
&\leq \sum_{k=0}^{\frac{n-2}{2}} H(t^*) p^{2k-1} (1-p)^{3n-3k-6} (1+p)^{n-1} \\
&\leq \sum_{k=0}^{\frac{n-2}{2}} H(t^*) p^{2k-1} (1-p)^{2n-2k-5} (1+p)^{n-1} (1-p)^{\frac{n}{2}} \\
&\leq \sum_{a=0}^{\lfloor \frac{n}{4} \rfloor} \binom{n-2}{a} \sum_{\sqrt{an}-1 \leq k \leq \sqrt{(a+1)n}-1} \binom{n-2-a}{k-a} \\
&\binom{n-2-k}{k-a} p^{2k-1} (1-p)^{\frac{5}{2}n-2k-5} (1+p)^{n-1} \\
&= \sum_{a=0}^{\lfloor \frac{n}{4} \rfloor} \binom{n-2}{a} p^{-1} (1-p)^{-1} (1+p)^{n-1} (1-p)^{\frac{n}{2}} \\
&\sum_{\sqrt{an}-1 \leq k \leq \sqrt{(a+1)n}-1} \binom{n-2-a}{k-a} p^k (1-p)^{n-2-k} \\
&\binom{n-2-k}{k-a} p^k (1-p)^{n-2-k},
\end{aligned} \tag{4-14}$$

where the last inequality means to dividing the sum into $\lfloor \frac{n}{4} \rfloor + 1$ parts since H^* has $\lfloor \frac{n}{4} \rfloor + 1$ different values when k is from 0 to n.

If we set k^* maximizing $\binom{n-2-a}{k-a}\binom{n-2-k}{k-a}$, and let

$$K(a) = \binom{n-2}{a}\binom{n-2-a}{k^*-a}\binom{n-2-k^*}{k^*-a},$$

then

$$\begin{aligned} C &\leq \sum_{a=0}^{\lfloor \frac{n}{4} \rfloor} K(a)(1-p^2)^{n-1}(1-p)^{\frac{3}{2}n-2\sqrt{an}}p^{2\sqrt{an}-4}\frac{1}{1-2p} \\ &= \frac{1}{g(p)}\sum_{a=0}^{\lfloor \frac{n}{4} \rfloor} K(a)(1-p^2)^{n}(1-p)^{\frac{3}{2}n-\sqrt{an}}p^{2\sqrt{an}} = \frac{1}{g(p)}\sum_{a=0}^{\lfloor \frac{n}{4} \rfloor} f(a), \end{aligned} \tag{4-15}$$

where $g(p) = (1-p^2)(1-2p)p^4$ and $f(a) = K(a)(1-p^2)^n(1-p)^{\frac{3}{2}n-\sqrt{an}}p^{2\sqrt{an}}$.

When $a = \Theta(n)$, if we let $a = c^2 n$ where $c \in (0, \frac{1}{2}]$, then according to Lemma 4.1.1,

$$f(a) \leq \left(F(c)(1-p^2)(1-p)^{\frac{3}{2}-2c}p^{2c}\right)^n. \tag{4-16}$$

Therefore when $p = o(1)$ we can ensure that $f(a) \to 0$ under $a = \Theta(n)$.

As for $p = \Theta(1)$, we need to explore a tighter bound since the RHS of Eqn. (4-16) can no longer guarantee to be arbitrarily close to 0 when $n \to \infty$. Recall Eqn. (4-11), $Z(n,p,k,t) = \binom{n-2}{t}\binom{n-2-t}{k-t}\binom{n-2-k}{k-t}$, and we set $\tilde{G}(t) = Z(n,p,s,t)\frac{1}{(1+p)^t}$. Now we intend to find the 'turning' point of $\tilde{G}(t)$, specifically t^* such that,

$$\frac{\tilde{G}(t+1)}{\tilde{G}(t)} = \frac{1}{1+p}\frac{(k-t)^2}{(t+1)(n-2k-1+t)} \leq 1; \tag{4-17}$$

$$\frac{\tilde{G}(t-1)}{\tilde{G}(t)} = (1+p)\frac{t(n-2k-2+t)}{(k-t+1)^2} \leq 1. \tag{4-18}$$

Note that we do not say t^* is the maximizer of $\tilde{G}(t)$, but we will show that it is later when $n \to \infty$. We rewrite Eqn. (4-17) and (4-18) respectively as

$$pt^2 + ((1+p)n - 2pk)t + (1+p)(n-2k) - k^2 \geq 0 \tag{4-19}$$

$$pt^2 + ((1+p)(n-2k-2) + 2(k+1))t - (k+1)^2 \leq 0 \tag{4-20}$$

Discuss k in different cases:

1. $k = o(\sqrt{n})$: we can rewrite Eqn. (4-19) and (4-20) as $pt^2 + ((1+p)n)t + (1+p)n \geq 0$ and $pt^2 + ((1+p)n)t \leq 0$. Note that $t = 0$ if and only if the latter inequality holds. It means that $\forall t \in \{0, 1, 2, \cdots, k\}$, $\tilde{G}(t)$ is monotonically non-increasing, so in this case, $t^* = 0$;

2. $k = \Omega(\sqrt{n}) \wedge k = o(n)$: we can rewrite Eqn. (4-19) and (4-20) as $pt^2 + ((1+p)n)t - k^2 \geq 0$ and $pt^2 + ((1+p)n)t - (k+1)^2 \leq 0$. Note that as t is an integer, when t adds 1, then there will be an increment of $(1+p)n = \Theta(n)$ on the LHS of the two inequalities, which outweighs the difference between two LHS: $(k+1)^2 - k^2 = 2k+1 = o(n)$. Therefore there will be only one $t^* \leq k$ satisfying both inequalities simultaneously. We can obtain that $t^* = \Theta(\frac{(k+1)^2}{n})$.

3. $k = \Theta(n)$: we set $k = cn$ where $c < \frac{1}{2}$. Then we can rewrite Eqn. (4-19) and (4-20) as $pt^2 + (1+p-2cp)nt + (1+p)(1-2c)n - c^2n^2) \geq 0$ and $pt^2 + (1+p-2cp)nt - c^2n^2 - 2cn - 1 \leq 0$. We can observe that $t^* = \Theta(n)$. Similarly, when t adds 1, then there will be an increment of $(1+p-2cp)n + 2tp + p$, which outweighs the difference between two LHS: $(1+p-2cp)n + 1$ as $t^* = \Theta(n)$.

Therefore above all, we have $t^* = \Theta(\frac{(k+1)^2}{n})$ is unique when $t^* \leq k$. Recall Eqn. (4-11), we can obtain

$$\begin{aligned} C &= \sum_{k=0}^{\frac{n-2}{2}} \sum_{t=0}^{k} Z(n,p,k,t) p^{2k} (1-p^2)^{n-2-t} (1-p)^{2n-3k+t-4} \\ &= \sum_{k=0}^{\frac{n-2}{2}} \sum_{t=o(n)} Z(n,p,k,t) p^{2k} (1-p^2)^{n-2-t} (1-p)^{2n-3k+t-4} \\ &+ \sum_{k=0}^{\frac{n-2}{2}} \sum_{t=\Theta(n)} Z(n,p,k,t) p^{2k} (1-p^2)^{n-2-t} (1-p)^{2n-3k+t-4} \end{aligned} \tag{4-21}$$

Hereinafter we discuss in cases of $k = o(n)$ and $k = \Theta(n)$.

1. $k = o(n)$: Then

$$\begin{aligned}
&Z(n,p,k,t)p^{2k}(1-p)^{3n-3k-6}(1+p)^{n-2-t} \\
&\leq Z(n,p,k,t^*)p^{2k}(1-p)^{3n-3k-6}(1+p)^{n-2-t^*} \\
&\sim \frac{(n-2)!}{(k!)^2(n-2k-2)!}p^{2k}(1-p)^{3n-3k-6}(1+p)^{n-2} \\
&\sim \frac{\sqrt{2\pi(n-2)}(\frac{n-2}{e})^{n-2}}{2\pi k(\frac{k}{e})^{2k}\sqrt{2\pi(n-2k-2)}}(\frac{n-2k-2}{e})^{n-2k-2} \\
&= \sqrt{\frac{n-2}{4\pi^2k^2(n-2k-2)}}(\frac{n-2}{k})^{2k}(\frac{n-2}{n-2k-2})^{n-2k-2} \\
&\sim \frac{1}{2\pi k}e^{2k\log\frac{n}{k}+(n-2k)\log\frac{n}{n-2k}}
\end{aligned} \tag{4-22}$$

For the function $h(k) = 2k\log\frac{n}{k} + (n-2k)\log\frac{n}{n-2k}$, taking the derivative in terms of k, we can obtain

$$\frac{d}{dk}f(k) = -2\log k + 2\log(n-2k). \tag{4-23}$$

Set k^* such that $\frac{d}{dk}f(k) = 0$, we have $k^* = \frac{n}{3} = \Theta(n)$. However $k = o(n)$, so $\frac{d}{dk}f(k) > 0$. Therefore we set $k = cn$ where $c = o(1)$, and we can obtain

$$\begin{aligned}
h(k) &\sim 2cn\log\frac{1}{c} + (1-2c)n\log\frac{1}{1-2c} \\
&\sim 2cn\log\frac{1}{c} = o(n)
\end{aligned} \tag{4-24}$$

Therefore $e^{h(k)} = o(e^n)$ and we can easily show that when $p = \Theta(1)$ then $Z(n,p,k,t)p^{2k}(1-p)^{3n-3k-6}(1+p)^{n-2-t} = o(\frac{1}{n})$.

2. $k=\Theta(n)$: Then $t^*=\Theta(n)$. Set $k=cn$, $t^*=\tilde{c}n$, where $0<\tilde{c}<c<\frac{1}{2}$. Then

$$\begin{aligned}
&Z(n,p,k,t^*)p^{2k}(1-p)^{3n-3k-6}(1+p)^{n-2-t^*}\\
&=\frac{(n-2)!p^{2cn}(1-p)^{(3-c)n-6}(1+p)^{(1-\tilde{c})n-2}}{(\tilde{c}n)!((c-\tilde{c})n!)^2((1-2c+\tilde{c})n)!}\\
&\sim\frac{n!p^{2cn}(1-p)^{(2-c+\tilde{c})n}(1-p^2)^{(1-\tilde{c})n}}{(\tilde{c}n)!((c-\tilde{c})n!)^2((1-2c+\tilde{c})n)!}\\
&=e^{\log n!-\log(\tilde{c}n)!-2\log((c-\tilde{c})n)!-\log((1-2c+\tilde{c})n)!}\\
&p^{2cn}(1-p)^{(2-c+\tilde{c})n}(1-p^2)^{(1-\tilde{c})n}\\
&\overset{\triangle}{\sim}\left(e^{\tilde{c}\log\tilde{c}-2(c-\tilde{c})\log(c-\tilde{c})-(1-2c+\tilde{c})\log(1-2c+\tilde{c})}\right)^n\\
&p^{2cn}(1-p)^{(2-c+\tilde{c})n}(1-p^2)^{(1-\tilde{c})n},\\
&=\left(\frac{p^{2c}(1-p)^{2-c+\tilde{c}}(1-p^2)^{1-\tilde{c}}}{\tilde{c}^{\tilde{c}}(c-\tilde{c})^{2(c-\tilde{c})}(1-2c+\tilde{c})^{1-2c+\tilde{c}}}\right)^n
\end{aligned}\tag{4-25}$$

where $\overset{\triangle}{\sim}$ is due to $\log(n!)\sim n\log n-n+\frac{1}{2}\log n+o(\log n)$. Then since $\frac{p^{2c}(1-p)^{2-c+\tilde{c}}(1-p^2)^{1-\tilde{c}}}{\tilde{c}^{\tilde{c}}(c-\tilde{c})^{2(c-\tilde{c})}(1-2c+\tilde{c})^{1-2c+\tilde{c}}}$ is a constant, based on Proposition 4.1.1, we have

$$\frac{p^{2c}(1-p)^{2-c+\tilde{c}}(1-p^2)^{1-\tilde{c}}}{\tilde{c}^{\tilde{c}}(c-\tilde{c})^{2(c-\tilde{c})}(1-2c+\tilde{c})^{1-2c+\tilde{c}}}<1,$$

then the case of $p=\Theta(1)$ is proved.

When $a=o(n)$, note that

$$\frac{d\left(\exp\sqrt{a}\ln\frac{n}{a}\right)}{da}=\frac{1}{\sqrt{a}}\left(\frac{1}{2}\ln\frac{n}{a}-1\right)>0,\tag{4-26}$$

let $a=n^{1-\epsilon}$ where $\epsilon>0$ and is arbitrarily small, and we discover that when $\epsilon=O(\frac{1}{\log(n)})$ then $a=\Theta(n)$. Thus we only consider $\epsilon=\omega(\frac{1}{\log(n)})$.

Therefore we have

$$\begin{aligned}
f(a)&\le e^{(\epsilon n^{1-\frac{\epsilon}{2}})\ln n}(1-p^2)^n(1-p)^{\frac{3}{2}n-2n^{1-\frac{\epsilon}{2}}}p^{2n^{1-\frac{\epsilon}{2}}}\\
&=\left(e^{\epsilon n^{-\frac{\epsilon}{2}}\ln n}(1-p^2)(1-p)^{\frac{3}{2}-2n^{-\frac{\epsilon}{2}}}p^{2n^{-\epsilon}}\right)^n
\end{aligned}\tag{4-27}$$

For the case $\epsilon=\Omega(1)$, it is obvious that $p=\omega\left(\frac{1}{n}\right)$ can meet the requirement.

For the case $\epsilon = o(1)$, note that by setting

$$\frac{d}{da}\left(e^{\sqrt{an}\ln\frac{n}{a}}\left(\frac{p}{1-p}\right)^{2\sqrt{an}}\right) = 0, \tag{4-28}$$

we can derive

$$a^* = \frac{np}{e^2(1-p)}, \tag{4-29}$$

and verify that the function $f(a)$ is a monotonically increasing function when $a < a^*$ and a monotonically decreasing function when $a > a^*$. Therefore $a^* = \arg\max f(a)$.

Now we focus on a^*. First, if $\frac{np}{e^2(1-p)} = o(n)$, i.e. $p = o(1)$, then we have

$$\begin{aligned} f(a^*) &\le \left(e^{\frac{1}{e}\sqrt{\frac{p}{1-p}}\ln(e^2\frac{1-p}{p})}(1-p)^{\frac{3}{2}}\left(\frac{p}{1-p}\right)^{\frac{2}{e}\sqrt{\frac{p}{1-p}}}\right)^n \\ &\le \left(e^{\frac{2}{e}\sqrt{\frac{p}{1-p}}}(1-p)^{\frac{3}{2}}(1-p^2)\right)^n \end{aligned} \tag{4-30}$$

If $\frac{np}{e^2(1-p)} = \Theta(n)$, i.e., $p = \Theta(1)$, as $a = o(n)$, therefore since $f(a)$ is monotonically increasing when $a < \frac{np}{e^2(1-p)}$, we have

$$f(a) \le f\left(\frac{n}{4}\right) \le \left(2(1-p)^{\frac{1}{2}}p(1-p^2)\right)^n, \tag{4-31}$$

which shows that when $p = \Theta(1)$ $(p \le \frac{1}{2})$, then $f(a) \to 0$.

Then we can easily verify that if $p = \omega(\frac{1}{n})$, we can obtain $C \to 0$.

2. Upper Bound of D:

Now we focus on D, where

$$\begin{aligned} D &= \sum_{k=\frac{n}{2}}^{n-2}\sum_{t=2k-n+2}^{k} Z(n,p,k,t) \\ &= \sum_{k'=0}^{\frac{n}{2}-2}\sum_{t=n-2-2k'}^{n-2-k'}\binom{n-2}{t}\binom{n-2-t}{n-2-k'-t}\binom{k'}{t+2k'+2-n} \\ &\quad p^{2(n-2-k')}(1-p^2)^{n-2-t}(1-p)^{3k'+2+t-n} \end{aligned} \tag{4-32}$$

Set $t' = t - (n-2-2k')$, then

$$D = \sum_{k'=0}^{\frac{n}{2}-2} \sum_{t'=0}^{k'} \binom{n-2}{n-2-2k'+t'} \binom{2k'-t}{n-2-k'-t} \binom{k'}{t+2k'+2-n} \\ p^{2(n-2-k')}(1-p^2)^{n-2-t}(1-p)^{3k'+2+t-n} \tag{4-33}$$

Then similarly we can find that $t^* = \left\lfloor \frac{(k+1)^2}{n} \right\rfloor$ makes D achieve maximum among all feasible values of t. Thus by similar methods,

$$\begin{aligned}
D &\le \sum_{k'=0}^{\frac{n-2}{2}} \sum_{t'}^{k'} H'\left(\left\lfloor \frac{(k+1)^2}{n} \right) \rfloor\right) p^{2(n-2-k')}(1-p)^{3k'}(1+p)^{2k'-t'} \\
&\le \sum_{k'=0}^{\frac{n-2}{2}} H'\left(\left\lfloor \frac{(k+1)^2}{n} \right) \rfloor\right) p^{2(n-2-k')-1}(1-p)^{3k'}(1+p)^{2k'+1} \\
&\le \sum_{k'=0}^{\frac{n-2}{2}} H'\left(\left\lfloor \frac{(k+1)^2}{n} \right) \rfloor\right) p^{2(n-2-k')-1}(1-p)^{2k'}(1+p)^{\frac{n}{2}} \\
&\le \sum_{a=0}^{\lfloor \frac{n}{4} \rfloor} H'\left(\left\lfloor \frac{(k^*+1)^2}{n} \right\rfloor\right) p^{2n-5}(1+p)^{\frac{n}{2}} \sum_{\sqrt{an}-1 \le k' \sqrt{(a+1)n}-1} \left(\frac{1-p}{p}\right)^{2k} \\
&\le \sum_{a=0}^{\lfloor \frac{n}{4} \rfloor} H'\left(\left\lfloor \frac{(k^*+1)^2}{n} \right\rfloor\right) p^{2n-3}(1+p)^{\frac{n}{2}} \left(\frac{1-p}{p}\right)^{\sqrt{n}(\sqrt{a+1}-\sqrt{a})+1} \\
&\le \max_{0 \le a \le \lfloor \frac{n}{4} \rfloor} H' \sum_{a=0}^{\lfloor \frac{n}{4} \rfloor} p^{2n-4}(1-p)^{\sqrt{n}+1}(1+p)^{\frac{n}{2}} \frac{1}{1-2p}
\end{aligned} \tag{4-34}$$

Then we can verify that for $p \le \frac{1}{2}$ this equation always approaches 0, thus making $D \to 0$.

Combining $C \to 0$ and $D \to 0$, we complete the proof.

□

Hereinafter, we focus on the lower bound of successful de-anonymization under E-R graph model. We prove that when $p = O\left(\frac{1}{n}\right)$, it is with probability 0 that we can conduct de-anonymization.

Theorem 4.2. *For an E-R graph $G(n,p)$, when $p = O\left(\frac{1}{n}\right)$, then G_1 can be successfully*

de-anonymized with probability 0 *for the fully sampled situation.*

Proof. A crucial observation is that $P((i,j) \in \xi_{Sym}) \geq P(d(i) = d(j) = 0)$, since if the degree of node i and j are both 0, then they must be symmetric. So we can obtain

$$\begin{aligned} P((i,j) \in \xi_{Sym}) &\geq P(d(i) = d(j) = 0) \\ &= P(d(i) = 0)P(d(j) = 0) \\ &= (1-p)\left(\binom{n-2}{0} p^0 (1-p)^{n-2}\right)^2 \\ &= (1-p)^{2n-3} \end{aligned} \tag{4-35}$$

Therefore when $p = o\left(\frac{1}{n}\right)$, $P((i,j) \in \xi_{Sym}) \geq (1-p)^{2n-3} \to 1$. When $p = \Theta\left(\frac{1}{n}\right)$, according to the union bound, only a constant number of node pairs (i,j) can be asymmetric, which denotes that only a constant number of nodes can be de-anonymized. So we obtain our result. □

From Theorem 4.1 and 4.2, we obtain that there is a phase transition at $\Theta\left(\frac{1}{n}\right)$ in terms of de-anonymization in E-R graph in fully sampled E-R graph model, illustrated in Fig. 4.1.

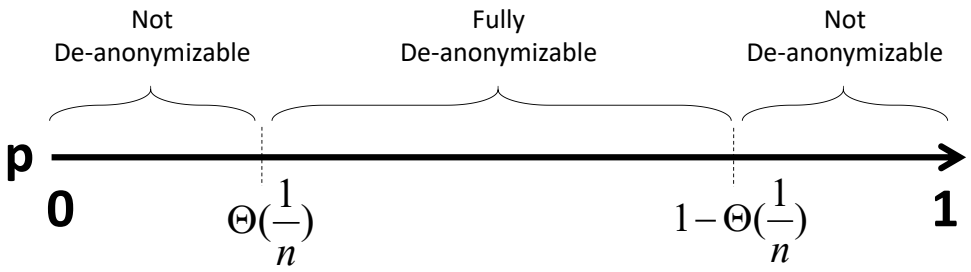


Figure 4.1 De-anonymizability for E-R graph when $s = 1$**.**

Remark: The threshold $p = \Theta(\frac{1}{n})$ indicates that the average degree $\mathbf{E}(d) = (n-1)p = \Theta(1)$ also acts as a threshold: If the average degree of a graph is constant when the network size n grows, then it is not de-anonymizable; If it grows as n with arbitrary rate, then as $n \to \infty$, the graph is de-anonymizable. For example, for a sparse graph, where $p = \Theta(\frac{1}{n})$, it is not de-anonymizable; for a graph a.s. connected where $p = \Theta(\frac{\log n}{n})$, then it is de-anonymizable.

4.1.2 Partly Sampled Situation: $s < 1$

Partly sampled situation, where relationship is not completely extracted from the underlying graph G to social networks G_1 and G_2, steps further towards the practical case than fully sampled situation: In real case, two users may have relationship, but they may not be connected in social networks; Even if they are connected in a social network, they are not necessarily connected in another. Therefore the sampling probability s counts much in the real application of de-anonymization. A natural question is that: Given p, which value of s will promise successful de-anonymization in large-scale social networks?

Though more practical, it exerts a crucial difficulty in theoretical aspect: The sampling process may damage the symmetry property of the underlying graph, disabling our attempt at studying the symmetry probability of nodes. If two nodes in the underlying graph are symmetric, they are not necessarily so after sampling since the edges may be discarded. Meanwhile, two nodes are symmetric in a social network does not lead to the symmetry in the underlying graph. Therefore, studying de-anonymization by fundamental exploration of symmetry does not make sense.

Take a step backward, we turn to the exploration of de-anonymization by graph matching. Graph matching zooms in on the discrepancy of edges between two graphs under the mapping Π. Specifically, set the adjacency matrix of G_1 and G_2 as $\mathbf{A}$ and $\mathbf{B}$ respectively, and the mapped adjacency matrix G_2 is $\mathbf{\Pi B \Pi}^T$. The discrepancy of edges can be characterized as

$$F(\mathbf{\Pi}) \triangleq ||\mathbf{A} - \mathbf{\Pi B \Pi}^T||_F^2, \tag{4-36}$$

where $||\cdot||_F$ denotes the Frobenius norm. This discrepancy measures how much error the mapping arouses. The minimizer of $F(\mathbf{\Pi})$, denoted as $\tilde{\mathbf{\Pi}}$, performs as an reasonable estimator of $\mathbf{\Pi_0}$.

Based on graph matching, we provide an upper and lower bound of the $||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2$. Both bounds, as we will show, are sensitive to the value of sampling probability s, indicating the tremendous effect of s on de-anonymization. Before that, we introduce several lemmas used in our main proof.

Given the underlying graph $G(n,p)$, permutation matrices $\tilde{\mathbf{\Pi}}$, $\mathbf{\Pi}_0$, and the adjacency matrix $\mathbf{B}$ of G_2. By taking the expectation over $\mathbf{B}$, we have

$$\frac{1}{2}\mathbf{E}(||\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0||_F^2)=\frac{\mathbf{E}(||(\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0)\mathbf{B}||_F^2)}{2nps(1-ps)} \tag{4-37}$$

Proof.

$$\begin{aligned}\mathbf{E}||(\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0)\mathbf{B}||_F^2&=\mathbf{E}\sum_{i=1}^{n}\sum_{j=1}^{n}[(\mathbf{\Pi_0}-\tilde{\mathbf{\Pi}})\mathbf{B}]_{ij}^2\\&=\mathbf{E}(\sum_{i=1}^{n}\mathbf{1}\{\pi_0(i)\neq\tilde{\pi}(i)\})2nps(1-ps)\\&=\mathbf{E}(\frac{1}{2}||\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0||_F^2)2nps(1-ps).\end{aligned} \tag{4-38}$$

□

Lemma 4.1.3. *Given permutation matrices $\tilde{\mathbf{\Pi}}$ and $\mathbf{\Pi}_0$, and adjacency matrices $\mathbf{A}$ and $\mathbf{B}$. We have*

$$||(\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0)\mathbf{B}||_F^2\leq 2(||\tilde{\mathbf{\Pi}}\mathbf{B}\mathbf{\Pi_0}^T-\mathbf{A}||_F^2+||\tilde{\mathbf{\Pi}}\mathbf{B}\tilde{\mathbf{\Pi}}^T-\mathbf{A}||_F^2). \tag{4-39}$$

Proof.

$$\begin{aligned}&||(\mathbf{\Pi_0}-\tilde{\mathbf{\Pi}})\hat{\mathbf{B}}||_F^2=||\hat{\mathbf{B}}(\mathbf{\Pi_0}-\tilde{\mathbf{\Pi}})^{\mathbf{T}}||_F^2\overset{\circ}{=}||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}(\mathbf{\Pi_0}-\tilde{\mathbf{\Pi}})^{\mathbf{T}}||_F^2\\&\leq(||(\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi}_0^T-\hat{\mathbf{A}})-(\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^T-\hat{\mathbf{A}})||_F)^2\\&\leq(||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}-\hat{\mathbf{A}}||_F+||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}}-\hat{\mathbf{A}}||_F)^2\\&\leq 2(||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}-\hat{\mathbf{A}}||_F^2+||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}}-\hat{\mathbf{A}}||_F^2),\end{aligned}$$

where $\overset{\circ}{=}$ holds because multiplying a permutation matrix keeps invariant of the value of Frobenius norm. □

Lemma 4.1.4. *Given permutation matrices $\tilde{\mathbf{\Pi}}$ and $\mathbf{\Pi}_0$, the adjacency matrices $\mathbf{A}$ and $\mathbf{B}$. We can obtain*

$$\begin{aligned}||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}}-\hat{\mathbf{A}}||_F^2\leq\frac{1}{2}(||\tilde{\mathbf{\Pi}}\hat{\mathbf{B}}\tilde{\mathbf{\Pi}}^{\mathbf{T}}-\hat{\mathbf{A}}||_F^2+||\mathbf{\Pi_0}\hat{\mathbf{B}}\mathbf{\Pi}_{\mathbf{0}}^{\mathbf{T}}-\hat{\mathbf{A}}||_F^2)\\+\boldsymbol{tr}((\tilde{\mathbf{\Pi}}-\mathbf{\Pi_0})\hat{\mathbf{B}}((\tilde{\mathbf{\Pi}}-\mathbf{\Pi_0})^{\mathbf{T}})\hat{\mathbf{A}}).\end{aligned} \tag{4-40}$$

Proof.

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

□

Lemma 4.1.5. *Given the underlying graph $G(n,p)$, sampling probability s, permutation matrices $\tilde{\mathbf{\Pi}}$ and $\mathbf{\Pi}_0$, and adjacency matrices $\mathbf{A}$ and $\mathbf{B}$. Taking the expectation over $\mathbf{A}$ and $\mathbf{B}$, we can derive*

$$
\mathbf{E}(\boldsymbol{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0)\mathbf{B}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0)^T\mathbf{A})) = \mathbf{E}(||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2)2np^2s^2(1-ps)^2. \tag{4-41}
$$

Proof. Set $\mathbf{Z} = (\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\mathbf{B}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}}\mathbf{A}$. Now we focus on $\mathbf{tr}(\mathbf{Z})$. Note that the i_{th} row of $\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0}$ is composed of either zeros if $\tilde{\mathbf{\Pi}}$ and $\mathbf{\Pi_0}$ map node i in G_2 to the same node in G_1, or zeros expect one 1 and one -1 if $\tilde{\mathbf{\Pi}}$ and $\mathbf{\Pi_0}$ map node i in G_2 to different nodes in G_1. 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 $s \neq t$. For simplicity, we define $\mathbf{Y} = (\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})\mathbf{B}$ and $\mathbf{X} = ((\tilde{\mathbf{\Pi}} - \mathbf{\Pi_0})^{\mathbf{T}})\mathbf{A}$, thus $\mathbf{Z} = \mathbf{YX}$. Then we can obtain the i_{th} row of $\mathbf{Y}$ as

$$
\mathbf{Y}_{i\cdot} = (\mathbf{B}_{s1} - \mathbf{B}_{t1}, \mathbf{B}_{s2} - \mathbf{B}_{t2}, \cdots, \mathbf{B}_{sn} - \mathbf{B}_{tn}).
$$

Similarly, we can obtain the i_{th} column of $\mathbf{X}$ as

$$
\mathbf{X}_{\cdot i} = (\mathbf{A}_{p_1 1} - \mathbf{A}_{q_1 1}, \mathbf{A}_{p_2 2} - \mathbf{A}_{q_2 2}, \cdots, \mathbf{A}_{p_n n} - \mathbf{A}_{q_n n})^{\mathbf{T}},
$$

where $p_i(q_i)$ means the row number of the $1(-1)$ in the i_{th} column of $\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0$, when $\tilde{\mathbf{\Pi}}$ and $\mathbf{\Pi}_0$ map node i in G_2 into different nodes in G_1. If they map node j in G_2 to the same node in G_1, then we set $\mathbf{X}_{ji} = 0$.

Then we have

$$\begin{aligned}\mathbf{E}(\mathbf{tr}(\mathbf{Z})) &= \mathbf{E}(\mathbf{tr}(\mathbf{YX})) = \mathbf{E}(\sum_{i=1}^{n}\sum_{k=1}^{n}\mathbf{Y}_{ik}\mathbf{X}_{ki}) \\ &= \mathbf{E}(\sum_{i=1}^{n}\sum_{k=1}^{n}(\mathbf{B}_{sk} - \mathbf{B}_{tk})(\mathbf{A}_{p_ik} - \mathbf{A}_{q_ik})) \\ &= 2nps(1-ps)\mathbf{E}(\sum_{k=1}^{n}(\mathbf{B}_{sk} - \mathbf{B}_{tk})) \\ &= 2np^2s^2(1-ps)^2\mathbf{E}(||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2).\end{aligned} \tag{4-42}$$

□

Lemma 4.1.6. *Given $G(n,p)$, the unique true mapping $\mathbf{\Pi}_0$ between G_1 and G_2, and adjacency matrices $\mathbf{A}$ and $\mathbf{B}$. By taking the expectation over $\mathbf{A}$ and $\mathbf{B}$, we have*

$$\mathbf{E}(||\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T - \mathbf{A}||_F^2) = 2psn(n-1)(1-s) \tag{4-43}$$

Proof. Note that $\mathbf{\Pi}_0$ is the true mapping, then for two nodes i_1 and j_1 in G_1, $\pi_0(i_1)$ and $\pi_0(i_2)$ in G_2 are connected if and only if i_1 and j_1 are connected. Therefore

$$\begin{aligned}\mathbf{E}(||\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T - \mathbf{A}||_F^2) &= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}(\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T - \mathbf{A})_{ij} \\ &= 2n(n-1)ps(1-s),\end{aligned} \tag{4-44}$$

where the result is not $2n^2ps(1-s)$ because the diagonal of $\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T - \mathbf{A}$ must be all zeros. □

Theorem 4.3. *Given $G_1 = (V, E_1, \mathbf{A})$ and $G_2 = (V, E_2, \mathbf{B})$ sampled from $G = (V, E, \mathbf{M})$ with probability $s < 1$. Suppose G is based on Erdos-Renyi model, i.e., $G(n,p)$. Set $\tilde{\mathbf{\Pi}} = \arg\min_{\mathbf{\Pi}} ||\mathbf{A} - \mathbf{\Pi}\mathbf{B}\mathbf{\Pi}^T||_F^2$ and $\mathbf{\Pi}_0$ as the true mapping. Then when $p = \omega(\frac{1}{n})$ (Suppose $p < \frac{1}{2}$), taking the expectation over $\mathbf{A}$ and $\mathbf{B}$, we obtain that the expected mapping error $\frac{1}{2}\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 = O(n(1-s))$.*

Proof. Denote the ratio:

$$\rho \triangleq \frac{||\tilde{\mathbf{\Pi}}\mathbf{B}\tilde{\mathbf{\Pi}}^T - \mathbf{A}||_F^2}{||\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T - \mathbf{A}||_F^2} \in [0,1]. \tag{4-45}$$

Based on Lemma 4.1.2 to Lemma 4.1.4, we can obtain

$$\begin{aligned}
&\frac{1}{2}\mathbf{E}(||\mathbf{\Pi}_0 - \tilde{\mathbf{\Pi}}||_F^2) \overset{1}{=} \frac{\mathbf{E}(||(\mathbf{\Pi}_0 - \tilde{\mathbf{\Pi}})\mathbf{B}||_F^2)}{2nps(1-ps)} \\
&\overset{2}{\leq} \frac{\mathbf{E}(||\tilde{\mathbf{\Pi}}\mathbf{B}\mathbf{\Pi_0}^T - \mathbf{A}||_F^2 + ||\tilde{\mathbf{\Pi}}\mathbf{B}\tilde{\mathbf{\Pi}}^T - \mathbf{A}||_F^2)}{nps(1-ps)} \\
&\overset{3}{\leq} \frac{1}{nps(1-ps)}\mathbf{E}(\frac{3}{2}||\tilde{\mathbf{\Pi}}\mathbf{B}\tilde{\mathbf{\Pi}}^T - \mathbf{A}||_F^2 + \frac{1}{2}||\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T - \mathbf{A}||_F^2 \\
&\quad + \mathbf{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0)\mathbf{B}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0)^T\mathbf{A})) \\
&\leq \frac{\mathbf{E}((3\rho+1)||\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T - \mathbf{A}||_F^2 + \mathbf{tr}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0)\mathbf{B}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0)^T\mathbf{A}))}{2nps(1-ps)} \\
&\overset{4}{=} \frac{\mathbf{E}((3\rho+1)||\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T - \mathbf{A}||_F^2 + 2np^2s^2(1-ps)^2||\mathbf{\Pi}_0 - \tilde{\mathbf{\Pi}}||_F^2)}{2nps(1-ps)} \\
&\overset{5}{\leq} ps(1-ps)\mathbf{E}||\mathbf{\Pi}_0 - \tilde{\mathbf{\Pi}}||_F^2 + \frac{3\rho+1}{(1-ps)}(n-1)(1-s),
\end{aligned}$$

where $\overset{1}{=}$ to $\overset{5}{\leq}$ are due to Lemma 4.1.2 to 4.1.4 respectively. Then we can bound the expected de-anonymization error $\frac{1}{2}\mathbf{E}||\mathbf{\Pi}_0 - \tilde{\mathbf{\Pi}}||_F^2$ as

$$\frac{1}{2}\mathbf{E}(||\mathbf{\Pi}_0 - \tilde{\mathbf{\Pi}}||_F^2) \leq \frac{(3\rho+1)n(1-s)}{(1-ps)(1-2ps(1-ps))}. \tag{4-46}$$

Therefore $\frac{1}{2}\mathbf{E}(||\mathbf{\Pi}_0 - \tilde{\mathbf{\Pi}}||_F^2) = O(n(1-s))$. □

Theorem 4.4. *With the same conditions as Theorem 4.3,*

$$\frac{1}{2}\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 = \Omega(n(1-s)^2). \tag{4-47}$$

Proof.

$$\begin{aligned}
&\mathbf{E}(||\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T - \mathbf{A}||_F^2 - ||\tilde{\mathbf{\Pi}}\mathbf{B}\tilde{\mathbf{\Pi}}^T - \mathbf{A}||_F^2) \\
&= 2\mathbf{E}(\mathbf{tr}(\mathbf{A}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0))\mathbf{B}(\tilde{\mathbf{\Pi}} + \mathbf{\Pi}_0)) \\
&= 2\mathbf{E}(\mathbf{tr}(\mathbf{A}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0))\mathbf{B}\tilde{\mathbf{\Pi}}) + \mathbf{E}(\mathbf{tr}(\mathbf{A}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0))\mathbf{B}\mathbf{\Pi}_0) \\
&\overset{1}{\leq} 4\mathbf{E}(||\mathbf{A}||_F||(\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0)\mathbf{B}||_F) \\
&\overset{2}{\leq} 4\sqrt{\mathbf{E}||\mathbf{A}||_F^2}\sqrt{\mathbf{E}||(\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0)\mathbf{B}||_F^2} \\
&= 4\sqrt{n(n-1)ps}\sqrt{\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 nps(1-ps)} \\
&\leq 4n^{\frac{3}{2}}ps\sqrt{1-ps}\sqrt{\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2},
\end{aligned} \tag{4-48}$$

where $\overset{1}{\leq}$ results from Cauchy-Schwarz Inequality: $|\mathbf{tr}(\mathbf{XY})| \leq \mathbf{tr}(\mathbf{X}^T\mathbf{X})\mathbf{tr}(\mathbf{Y}^T\mathbf{Y}) = ||\mathbf{X}||_F^2||\mathbf{Y}||_F^2$ and $\overset{2}{\leq}$ stems from Jensen's Inequality. Therefore

$$\begin{aligned}
\frac{1}{2}\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 &\geq \frac{(1-\rho)^2(\mathbf{E}||\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T - \mathbf{A}||_F^2)^2}{32n^3p^2s^2(1-ps)} \\
&= \frac{(1-\rho)^2n^2(n-1)^2p^2s^2(1-s)^2}{32n^3p^2s^2(1-ps)} \\
&\geq \frac{(1-\rho)^2(1-\epsilon)}{32(1-ps)}n(1-s)^2,
\end{aligned} \tag{4-49}$$

where $\epsilon > 0$ is an arbitrarily small constant. Therefore $\frac{1}{2}\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 = \Omega(n(1-s)^2)$. □

Remark: Theorem 4.3 and 4.4 illustrate that if s is closer to 1, then the expected mapping error will decrease. Specifically, some of the intuitive results are listed in Table 4.1. We can observe that when $s = 1 - o\left(\frac{1}{n}\right)$, then we can ensure that the de-anonymization based on graph matching can be conducted with arbitrarily small error. Intuitively, if there are $\Theta(n^2)$ edges in G, then only $o(n)$ edges can be sampled to ensure a.s. accurate de-anonymization. As s gets faraway from 1, the expected mapping error ascends, and finally when s is a constant, i.e., a constant portion of edges in G are discarded during the sampling, both upper and lower bounds are linear with n, which denotes that the de-anonymization error is uncontrollable: The absolute error $||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 = \Theta(n)$ while the relative error also $\frac{||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2}{||\mathbf{\Pi}_0||_F^2}$ does not vanish to 0 as

Table 4.1 Intuitive Results for Theorem 4.3 and 4.4

s	Upper Bound	Lower Bound	Accuracy
$1-o(\frac{1}{n})$	$o(1)$	$o(\frac{1}{n})$	Absolutely Bounded
$1-\Theta(\frac{1}{n})$	$\Theta(1)$	$\Theta(\frac{1}{n})$	Absolutely Bounded
$1-\Theta(\frac{\log n}{n})$	$\Theta(\log n)$	$\Theta(\frac{\log^2 n}{n})$	Relatively Bounded
$1-\Theta(\frac{1}{\sqrt{n}})$	$\Theta(\sqrt{n})$	$\Theta(1)$	Relatively Bounded
$1-\Theta(\sqrt{\frac{\log n}{n}})$	$\Theta(\sqrt{n\log n})$	$\Theta(\log n)$	Relatively Bounded
$\Theta(1)$	$\Theta(n)$	$\Theta(n)$	Unbounded

$n \to \infty$.

The fact that s cannot be any constant value in $(0,1)$ seems to be counter-intuitive, but it makes sense since for a connection in G to be characterized in both two social networks G_1 and G_2, the probability is s^2: If $s<1$ is a constant, then s^2 decays greatly away from 1 than s. However, if $s=1-o(1)$, then $s^2=1-o(1)$, which still does not deviate from 1 too much. This phase transition from $s=\Theta(1)$ to $s=1-o(1)$ determines whether there is a non-ignorable connection loss from G to G_1 and G_2, and we can draw that in partly sampled situation under E-R graph model, de-anonymization can be conducted with controllable mapping error if and only if the connection loss engendered by sampling is negligible.

For the completeness of discussing the case of $s<1$, another aspect is that $p=o(\frac{1}{n})$, i.e., there are symmetric nodes in G, hence for the adversary, there are multiple possibilities of the true mapping $\mathbf{\Pi}_0$ in front of him. Assume that there are k asymmetric nodes in G.This, reflected in our analysis, will lead Lemma 4.1.6 to be another version as

$$\begin{aligned}&\mathbf{E}(||\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T-\mathbf{A}||_F^2)\\&=2ps(k(k-1)(1-s)+(n-k)(n-k-1)(1-ps)),\end{aligned}\tag{4-50}$$

since for those $n-k$ symmetric nodes, the adversary's oracle estimation $\tilde{\mathbf{\Pi}}_0$ can vary and be different from $\mathbf{\Pi}_0$, so that no longer can it ensure the expectation of edge discrepancy $\mathbf{E}(\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T-\mathbf{A})_{ij}^2$ to be $2ps(1-s)$, but $2ps(1-ps)$.

Therefore, for the upper bound,

$$\begin{aligned}&\frac{1}{2}\mathbf{E}(||\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0||_F^2)\\&\leq\frac{(3\rho+1)(k(k-1)ps(1-s)+(n(n-1)-k(k-1))ps(1-ps))}{2nps(1-ps)(1-2ps(1-ps))}\\&\leq\frac{(3\rho+1)k(k-1)(1-s)}{2n(1-ps)(1-2ps(1-ps))}+\frac{(3\rho+1)(n^2-k^2)}{2n(1-2ps(1-ps))},\end{aligned}\tag{4-51}$$

and for the lower bound,

$$\begin{aligned}&\frac{1}{2}\mathbf{E}(||\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0||_F^2)\\&\geq\frac{(1-\rho)^2(k(k-1)ps(1-s)+(n(n-1)-k(k-1))ps(1-ps))^2}{32n^3p^2s^2(1-ps)}.\end{aligned}\tag{4-52}$$

We can find that $\frac{1}{2}\mathbf{E}(||\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0||_F^2)=O(\frac{k^2}{n}(1-s)+\frac{n^2-k^2}{n})$ and $\frac{1}{2}\mathbf{E}(||\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0||_F^2)=\Omega(\frac{k^4}{n^3}(1-s)^2+\frac{k^2(n^2-k^2)}{n^3}(1-s)+\frac{(n^2-k^2)^2}{n^3})$. Note that for both bounds, there is a term irrelevant with s: For the upper bound is $\frac{n^2-k^2}{n}$ while for the lower bound is $\frac{(n^2-k^2)^2}{n^3}$. This means that even if $s\to 1$, the expected mapping error is not necessarily controllable. For instance, even if there are $k=\Theta(n)$ asymmetric nodes, the upper and lower bound turn to be $O(n)$ and $\Omega(n)$, showing that no matter how s is close to 1, the expected mapping error is linear with n. However, when $k=n-o(\sqrt{n})$, the term irrelevant to s will be negligible and both bounds become the same as the case of a.s. no symmetric nodes in G. This indicates that the symmetry property in G, predominating the success of de-anonymization in fully sampled situation, also performs as an additive but sometimes still predominant factor in determining the accuracy of de-anonymization in partly sampled situation.

4.2 De-anonymizability of Stochastic Block Model

Similar to but incorporating community structure based on E-R graph model, the stochastic block model (SBM) performs as a more lifelike depiction of real social networks. However, more complexity is induced as it needs to characterize two probabilities: inner-connected probability p and inter-connected probability q. Nevertheless,

due to our assumption that the connection existence in G is independent, the essence of de-anonymization in SBM does not deviate much from that in E-R graph model, hence the main approach follows that in E-R graph model with some affinity.

We also propose an upper bound and a lower bound of symmetry probability and expected mapping error respectively for both fully and partly sampled situations.

4.2.1 Fully Sampled Situation: $s = 1$

Theorem 4.5. *Given $G(n, p, q)$ and K communities, where the i^{th} community is of size c_i. Then if $\forall c_i$, $i \in \{1, 2, \cdots, K\}$, $(1-p)^{c_i}(1-q)^{n-c_i} < 1$, or equivalently, $p = \omega(C_i) \vee q = \omega(n - C_i)$, then G_1 can be successfully de-anonymized with probability 1 for the fully sampled situation.*

Proof. The proof follows a similar procedure but turns to be more complex since it involves inner-connection and inter-connection. We also characterize the probability of symmetric nodes in G.

Note that if two nodes are symmetric in SBM, they must exist in the same community. Therefore to characterize the symmetry probability, we should consider not only

degree equality, but also the *inner-degree* and *inter-degree* equality, specified as

$$
\begin{aligned}
&P((i,j)\in\xi_{sym}) \le P(d(i)=d(j)) = \sum_{k=0}^{n-2} P(d(i)=d(j)=k) \\
&= \sum_{k=0}^{n-2}\sum_{t=0}^{k} P(d(i)=d(j)=k, d(i,j)=t) \\
&= \sum_{k=0}^{n-2}\sum_{t=0}^{k}\sum_{x=0}^{\min\{k,a-2\}} \\
&\quad P(d(i)=d(j)=k, d(i,j)=t, d_{in}(i)=d_{in}(j)=x) \\
&= \sum_{k=0}^{n-2}\sum_{t=0}^{k}\sum_{x=0}^{\min\{k,a-2\}}\sum_{y=0}^{\min\{t,x\}} \\
&\quad P(d(i)=d(j)=k, d(i,j)=t, d_{in}(i)=d_{in}(j)=x, d_{in}(i,j)=y) \\
&= \sum_{k=0}^{n-2}\sum_{t=0}^{k}\sum_{x=0}^{\min\{k,a-2\}}\sum_{y=0}^{\min\{t,x\}} \\
&\quad \binom{a-2}{y}(p^2)^y(1-p^2)^{a-2-y}\binom{a-2-y}{x-y}p^{x-y}(1-p)^{a-2-x} \\
&\quad \binom{a-2-x}{x-y}p^{x-y}(1-p)^{a-2-2x+y}\binom{n-a}{t-y}(q^2)^{t-y}(1-q^2)^{n-a-(t-y)} \\
&\quad \binom{n-a-(t-y)}{k-x-(t-y)}q^{k-x-(t-y)}(1-q)^{n-a-(k-x)} \\
&\quad \binom{n-a-(k-x)}{k-x-(t-y)}q^{k-x-(t-y)}(1-q)^{n-a-2(k-x)+t-y} \\
&= \sum_{k=0}^{n-2}\sum_{t=0}^{k}\sum_{x=0}^{\min\{k,a-2\}}\sum_{y=0}^{\min\{t,x\}} \binom{a-2}{y}\binom{a-2-y}{x-y}\binom{a-2-x}{x-y} \\
&\quad \binom{n-a}{t-y}\binom{n-a-(t-y)}{k-x-(t-y)}\binom{n-a-(k-x)}{k-x-(t-y)}p^{2x}(1-p^2)^{a-2-y} \\
&\quad (1-p)^{2a-4-3x+y}q^{2(k-x)}(1-q^2)^{n-a-(t-y)}(1-q)^{2(n-a)-3(k-x)+t-y} \\
&\overset{\circ}{\le} \sum_{k=0}^{n-2}\sum_{t=0}^{k} K(p,a)H(q,n-a),
\end{aligned}
\tag{4-53}
$$

where

$$
K(p,a) = \sum_{x=0}^{\min\{k,a-2\}}\sum_{y=0}^{\min\{t,x\}} \tilde{K}(a)p^{2x}(1-p^2)^{a-2-y}(1-p)^{2a-4-3x+y} \tag{4-54}
$$

$$H(q,n-a)=\sum_{x=0}^{\min\{k,a-2\}}\sum_{y=0}^{\min\{t,x\}}\tilde{H}(n-a)q^{2(k-x)}(1-q^2)^{n-a-(t-y)} (1-q)^{2(n-a)-3(k-x)+t-y}. \tag{4-55}$$

$\overset{\circ}{\leq}$ since $\tilde{K}(a)p^{2x}(1-p^2)^{a-2-y}(1-p)^{2a-4-3x+y}<1$ and $\tilde{H}(n-a)q^{2(k-x)}(1-q^2)^{n-a-(t-y)}(1-q)^{2(n-a)-3(k-x)+t-y}<1$, then based on $xy<x+y$ when $0<x,y<1$ we obtain $\overset{\circ}{\leq}$.

For $K(p,a)$ and $H(q,n-a)$, similar to the proof in Theorem 4.1, we can obtain that

$$K(p,a)\leq\frac{(1-p)^{2a}(1-p^2)^{a-1}}{p^2(1-2p)^2} \max_{0\leq x\leq\min\{k,a\}}\binom{a-2}{y^*}\binom{a-2-y^*}{x-y^*}\binom{a-2-x}{x-y^*}, \tag{4-56}$$

where $y*=\lfloor\frac{(x+1)^2}{a}\rfloor$ and

$$H(q,n-a)\leq\frac{(1-q)^{2(n-a)-k+2}(1-q^2)^{n-a}}{q^2(1-2q)^2} \max_{0\leq x\leq\min\{k,a\}}\binom{n-a}{t-y^*}\binom{n-a-(t-y^*)}{k-x-(t-y^*)}\binom{n-a-(k-x)}{k-x-(t-y^*)}, \tag{4-57}$$

where $(t-y)^*=\lfloor\frac{(k-x+1)^2}{n-a+2}\rfloor$. Then following the derivation of Eqn. (4-27), we can obtain that for any $\epsilon_1=\omega(\frac{1}{\log a})$ and $\epsilon_2=\omega(\frac{1}{\log(n-a)})$.

$$\begin{aligned}K(p,a)H(q,n-a)&\leq(e^{\epsilon_1 a^{-\frac{\epsilon_1}{2}}}(1-p)^{\frac{3}{2}-2a^{-\frac{\epsilon_1}{2}}}p^{2a^{-\frac{\epsilon_1}{2}}})^a\\&\quad(e^{\epsilon_2(n-a)^{-\frac{\epsilon_2}{2}}}(1-p)^{\frac{3}{2}-2(n-a)^{-\frac{\epsilon_2}{2}}}p^{2(n-a)^{-\frac{\epsilon_2}{2}}})^{n-a}\\&\sim(1-p)^{\frac{3}{2}a}(1-q)^{\frac{3}{2}(n-a)}\end{aligned} \tag{4-58}$$

Then we discuss the community size a in different cases:

1. $a=\Theta(1)$: the exponent of p is constant while that of q is $\Theta(n)$, hence if $q=\omega(\frac{1}{n-a})$, then we can control the symmetry probability by an upper bound vanishing to 0 as $n\to\infty$.

2. $a=o(n)$: the exponent of p is $o(n)$ while that of q is $\Theta(n)$, hence if $p=\omega(\frac{1}{a})$ or $q=\omega(\frac{1}{n-a})$, then the symmetry probability can be controlled.

Table 4.2 Summarization for Theorem 4.5

a	p	q
$\Theta(1)$	/	$\omega(\frac{1}{n})$
$o(n)$	$\omega(\frac{1}{a})$	$\omega(\frac{1}{n})$
$\Theta(n)$	$\omega(\frac{1}{n})$	$\omega(\frac{1}{n})$
$n-o(n)$	$\omega(\frac{1}{n})$	$\omega(\frac{1}{n-a})$
$n-\Theta(1)$	$\omega(\frac{1}{n})$	/

3. $a=\Theta(n)$: the exponent of p is $\Theta(n)$ while that of q is $\Theta(n)$, hence if $p=\omega(\frac{1}{n})$ or $q=\omega(\frac{1}{n})$, then the symmetry probability can be controlled.

Table 4.2 summarizes all possible cases of a. Additionally, as $a,n\to 1$, $(1-p)^{\frac{3}{2}a}\leq 1$ and $(1-q)^{\frac{3}{2}(n-a)}\leq 1$, therefore $(1-p)^{\frac{3}{2}a}(1-q)^{\frac{3}{2}(n-a)}<1$ if and only if $(1-p)^{\frac{3}{2}a}<1$ or $(1-q)^{\frac{3}{2}(n-a)}<1$. So we complete our proof.

□

Theorem 4.6. *For SBM $G(n,p,q)$, if there exists a community with size b, such that $p=O(\frac{1}{b})$ and $q=O(\frac{1}{n-b})$, then G_1 can be successfully de-anonymized with probability 0 for the fully sampled situation.*

Proof. Similarly, we use the probability of two nodes with degree 0 as the lower bound of symmetric nodes in G. Note that node i and j should be in the same community. Specifically,

$$\begin{aligned}P((i,j)\in\xi_{sym}) &\geq P(d(i)=d(j)=0)\\ &=(1-p)(\binom{b-2}{0}p^0(1-p)^{b-2}\binom{n-b}{0}q^0(1-q)^{n-b})^2\\ &=(1-p)^{2(b-1)}(1-q)^{2(n-b)}.\end{aligned}\tag{4-59}$$

Then when $p=O(\frac{1}{b})$ and $q=O(\frac{1}{n-b})$, $P((i,j)\in\xi_{sym})$ does not vanish to 0, indicating uncontrollable de-anonymization error. □

4.2.2 Partly Sampled Situation: $s < 1$

In SBM, the partly sampled situation also follows the style as E-R graph model. More consideration occurs in the communities limiting the available choice of mapping. Concretely, two nodes in G_1 and G_2 respectively representing the same user must exist in the same community. Predicated on this unique attribute compared with E-R graph, there are some different and enlightening results, shown below, about the effect of communities in de-anonymization.

Recall Eqn. (4-36) in E-R graph model. Naturally we can transplant it in SBM, but the restriction of mapped nodes in identical community, mentioned above, is indispensable. So we need the following extra constraint attached to Eqn. (4-36):

$$C(i) = C(\tilde{\boldsymbol{\pi}}(i)), \forall i \in \{1, 2, \cdots, K\}, \tag{4-60}$$

where $C(i)$ denotes the community node i belongs to and $\tilde{\boldsymbol{\pi}}$ corresponds equally to $\tilde{\boldsymbol{\Pi}}$.

Lemma 4.2.1. *Given $G(n, p, q)$, G_1 and G_2, adjacency matrices $\mathbf{A}$ and $\mathbf{B}$ respectively for G_1 and G_2, sampling probability $s < 1$, K communities with size c_i for the i^{th} community ($i \in \{1, 2, \cdots, K\}$), and the true mapping $\boldsymbol{\Pi}_0$. We have*

$$\begin{aligned} &\mathbf{E}||\mathbf{A} - \boldsymbol{\Pi}_0\mathbf{B}\boldsymbol{\Pi}_0^T||_F^2 \\ &= 2s(1-s)(p\sum_{i=1}^{K} c_i(c_i-1) + q(n(n-1) - \sum_{i=1}^{K} c_i(c_i-1))). \end{aligned} \tag{4-61}$$

Proof. For two nodes within the same community, the probability of connection in G is p, and there are totally $\sum_{i=1}^{K} \frac{c_i(c_i-1)}{2}$ node pairs. For two nodes in two different communities, the probability of connection in G is q, and there are totally $\frac{n(n-1)}{2} - \sum_{i=1}^{K} \frac{c_i(c_i-1)}{2}$ node pairs. As shown in Lemma 4.1.6, the unique true mapping will cause $\mathbf{E}(\mathbf{A} - \boldsymbol{\Pi}_0\mathbf{B}\boldsymbol{\Pi}_0^T)_{ij} = 2ps(1-s)$ when i and j belong to the same community, while $\mathbf{E}(\mathbf{A} - \boldsymbol{\Pi}_0\mathbf{B}\boldsymbol{\Pi}_0^T)_{ij} = 2qs(1-s)$ otherwise. Thus we complete our proof. □

Theorem 4.7. *Given $G(n, p, q)$, G_1 and G_2, adjacency matrices $\mathbf{A}$ and $\mathbf{B}$ respectively for G_1 and G_2, sampling probability $s < 1$, K communities with size c_i for the i^{th}*

community ($i \in \{1, 2, \cdots, K\}$)*, and the true mapping* $\mathbf{\Pi}_0$ *and estimation by minimizing Eqn. (4-36)* $\tilde{\mathbf{\Pi}}$ *under constraint (4-60). We have*

$$\frac{1}{2}\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 = O\left(\left(\frac{q}{p}n + \left(1 - \frac{q}{p}\right)\frac{\sum_{i=1}^K c_i^2}{n}\right)(1-s)\right). \tag{4-62}$$

Proof. Based on Lemma 4.2.1, we can obtain

$$\begin{aligned}
&\frac{1}{2}\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 \\
&\leq \frac{(3\rho+1)\mathbf{E}||\mathbf{A} - \mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T||_F^2}{2nps(1-ps)(1-2ps(1-ps))} \\
&= \frac{(3\rho+1)(1-s)}{2nps(1-ps)(1-2ps(1-ps))} \\
&\quad (2ps(1-s)\sum_{i=1}^K c_i(c_i-1) + 2qs(1-s)(n(n-1) - \sum_{i=1}^K c_i(c_i-1))) \\
&\sim \frac{(3\rho+1)(1-s)}{(1-ps)(1-2ps(1-ps))}\left(\frac{q}{p}n + \frac{\sum_{i=1}^K c_i^2}{n}\left(1-\frac{q}{p}\right)\right).
\end{aligned} \tag{4-63}$$

Then we complete the proof. □

Theorem 4.8. *With the same conditions as Theorem 4.7,*

$$\frac{1}{2}\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 = \Omega\left(\left(\frac{q}{p}n + \left(1 - \frac{q}{p}\right)\frac{\sum_{i=1}^K c_i^2}{n}\right)(1-s)^2\right). \tag{4-64}$$

Proof. The proof is similar to Theorem 4.4.

$$\begin{aligned}
&\mathbf{E}(||\mathbf{A} - \mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T||_F^2 - ||\mathbf{A} - \tilde{\mathbf{\Pi}}\mathbf{B}\tilde{\mathbf{\Pi}}^T||_F^2) \\
&\leq 4\sqrt{\mathbf{E}||\mathbf{A}||_F^2}\sqrt{\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2} \\
&= 4\sqrt{\sum_{i=1}^K c_i(c_i-1)ps + (n(n-1) - \sum_{i=1}^K c_i(c_i-1))qs} \\
&\quad \sqrt{\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 nps(1-ps)} \\
&= 4s\sqrt{np(1-ps)(n(n-1)q + \sum_{i=1}^K c_i(c_i-1)(p-q))}\sqrt{\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2}
\end{aligned} \tag{4-65}$$

Then we can upper bound $\frac{1}{2}\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2$ as

$$\begin{aligned}
&\frac{1}{2}\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 \\
&\geq \frac{(1-\rho)^2(\mathbf{E}||\mathbf{A} - \mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T||_F^2)}{32s^2np(1-ps)(n(n-1)q + (\sum_{i=1}^K c_i(c_i-1))(p-q))} \\
&= \frac{(1-\rho)^2(1-s)^2(p\sum_{i=1}^K c_i(c_i-1) + q(n(n-1) - \sum_{i=1}^K c_i(c_i-1)))^2}{8np(1-ps)(n(n-1)q + (\sum_{i=1}^K c_i(c_i-1))(p-q))} \\
&= \frac{(1-\rho)^2(1-s)^2}{8(1-ps)}\left(\frac{q}{p}n + \left(1-\frac{q}{p}\right)\frac{\sum_{i=1}^K c_i^2}{n}\right).
\end{aligned} \tag{4-66}$$

Then we complete the proof. □

As we can see from Theorem 4.7 and 4.8, besides the sampling probability, the number of communities, the size of each community, and the ratio between inter-connection probability q and inner-connection probability p all play important roles in bounding the expected mapping error. Specifically,

Probability Ratio $\frac{q}{p}$: This ratio reflects the relative difference of connection density within communities or not. If $q > p$ then comparatively more connections consist in bridging nodes in different communities, while $p < q$ shows the opposite case. If $p = q$, the SBM performs as an E-R graph model.

From Theorem 4.7 and 4.8, we can observe that $\sum_{i=1}^K c_i^2 \leq n^2$ conditioned upon $\sum_{i=1}^K c_i = n$, therefore $\frac{q}{p}$ positively correlates to the expected mapping error since its coefficient $(n - \frac{\sum_{i=1}^K c_i^2}{n})$ is non-negative. Hence we can draw the conclusion that $p > q$ benefits more in higher de-anonymization accuracy.

Concretely its effect is shown in Table 4.3: When $\frac{q}{p} = \Theta(1)$, the case is the same as E-R graph model. When $\frac{q}{p} = \Theta(n)$, the bound in terms of n is of order n^2 unless there is only one community, indicating that $q > p$ will lead de-anonymization effect inferior to that under E-R graph model. When $\frac{q}{p} = \Theta\left(\frac{1}{n}\right)$, the bound in terms of n is of order $\frac{\sum_{i=1}^K c_i^2}{n}$, offering chances that relax the constraint for s to achieve vanishing expected mapping error: For example assume $K = \Theta(n)$, then if all K commnuities are of the same size $\frac{n}{K}$, the bounds turn to be $O(1-s)$ and $\Omega((1-s)^2)$, meaning that $s = 1 - o(1)$ is enough for absolutely bounded the error close to 0, rather than $s = 1 - o(\frac{1}{n})$ in E-R

Table 4.3 Effect of $\frac{q}{p}$ in Theorem 4.7 and 4.8

$\frac{q}{p}$	Upper Bound	Lower Bound
n	$(1-s)(n^2-\sum_{i=1}^{K}c_i^2)$	$(1-s)^2(n^2-\sum_{i=1}^{K}c_i^2)$
1	$(1-s)n$	$(1-s)^2n$
$\frac{1}{n}$	$(1-s)\frac{\sum_{i=1}^{K}c_i^2}{n}$	$(1-s)^2\frac{\sum_{i=1}^{K}c_i^2}{n}$

model, and even $s=\Theta(1)$ can achieve constant mapping error, unlike uncontrollable in E-R model. Therefore $p>q$ is conducive to de-anonymization, which corresponds to the realistic social network communities.

Homogenity of Community Size Distribution: This homogenity measures how variant the size of communities is. If all communities are of similar size, it means the community size distribution is homogenous, otherwise it is heterogenous.

To bring an intuitive view of its effect, we consider two extreme cases: completely homogenous and heterogenous. For completely homogenous case, all communities are of same size $\frac{n}{K}$, then $\frac{\sum_{i=1}^{K}c_i^2}{n}$ reaches its minimum $\frac{n}{K}$; For completely heterogenous case, one community contains all nodes, then $\frac{\sum_{i=1}^{K}c_i^2}{n}$ reaches its maximum n.

Meanwhile, we notice that the coefficient of $\frac{\sum_{i=1}^{K}c_i^2}{n}$ is $\left(1-\frac{q}{p}\right)$, so the general guiding direction of homogenity depends on whether the network is assortative or dissortative. If assortative $(p>q)$, then more homogeneity promises higher accuracy; If dissortative $(q<p)$, then more heterogeneity does better. This can be explained as follows: High heterogeneity means a slew of nodes exist in a large community. They act as if the same status if only considering inner-connection. Hence the higher interconnection will bring higher distinguishability for nodes inside this large set, because it may connect to other smaller communities which nodes inside have higher distinctiveness due to their finer granularity.

4.3 De-anonymizability of Power Law Model

Power law model is shown to exist in many real social networks. Unlike E-R graph model and SBM pinpointing on the probability of connection between any node pair, power law characterizes the degree distribution of each node. Specifically, a power law

based graph model can be denoted as $\Gamma(n, \gamma, k_{\min})$, with the degree distribution

$$P(d(i) = k) = \frac{k^{-\gamma}}{\sum_{t=t_{\min}}^{n-1} t^{-\gamma}}, \tag{4-67}$$

where $\gamma \in [0, +\infty)$ is the exponent adjusting the homogenity of degree distribution, and $k_{\min}$ is the minimum degree in the graph.

For γ, if γ is small, then the probability for every possible degree value from $k_{\min}$ to $n-1$ does not vary much, and for the extreme case where $\gamma = 0$, all degree values are of the same probability; if γ is large, then the probability for low degree values will largely outweighs that for high ones, and for the extreme case where $\gamma \to \infty$, the node degree is $k_{\min}$ with probability 1, showing the ultimate imbalance of degree values.

For $k_{\min}$, the minimum degree of nodes in Γ, its function is to lead power law model to nestle up to practical cases. According to survey on real social networks with power law property, it is discovered that only when the node degree exceeds some value do power law distribution holds elegantly. Therefore, we affiliate our power law model with this $k_{\min}$.

4.3.1 Fully Sampled Situation: $s = 1$

We still consider the fully sampled situation by characterizing the symmetry property. However, due to the parameter $k_{\min}$, it is hard to provide the lower bound of symmetry nodes as that in E-R model and SBM since it does not contain nodes with degree 0. Therefore, we only provide the upper bound result, while leaving the lower bound as our future work.

Theorem 4.9. *Given $\Gamma(n, \gamma, k_{\min})$, then for two nodes i and j, the probability they are symmetric is*

$$\begin{aligned} P((i,j) \in \xi_{sym}) &\leq UB(n, \gamma, k_{\min}) \\ &= \frac{(1-\gamma)^2}{1-2\gamma} \frac{(n-1)^{1-2\gamma} - (k_{\min}-1)^{1-2\gamma}}{(n^{1-\gamma} - k_{\min}^{1-\gamma})^2}. \end{aligned} \tag{4-68}$$

Then the adversary can de-anonymize successfully with probability 1 if $UB(n, \gamma, k_{\min}) \to$

0 *when* $n \to \infty$ *in fully sampled situation.*

Proof. For nodes i and j, we still upper bound $P((i,j) \in \xi_{sym})$ by $P(d(i) = d(j))$, i.e.,

$$\begin{aligned} P((i,j) \in \xi_{sym}) &\leq P(d(i) = d(j)) \\ &= \sum_{k=k_{\min}}^{n-1} P(d(i) = d(j) = k) \\ &= \frac{\sum_{k=k_{\min}}^{n-1} k^{-2\gamma}}{(\sum_{k=k_{\min}}^{n-1} k^{-\gamma})^2} \end{aligned} \tag{4-69}$$

Then we discuss different cases of γ.

1. $\gamma > 1$**:** Note that

$$\sum_{k=k_{\min}}^{n-1} k^{-2\gamma} \leq \int_{k_{\min}-1}^{n-1} k^{-2\gamma} dk = \frac{(n-1)^{1-2\gamma} - (k_{\min}-1)^{1-2\gamma}}{1-2\gamma}, \tag{4-70}$$

and

$$(\sum_{k=k_{\min}}^{n-1} k^{-\gamma})^2 \geq \left(\int_{k_{\min}}^{n} k^{-\gamma} dk\right)^2 = \frac{(n^{1-\gamma} - k_{\min}^{1-\gamma})^2}{(1-\gamma)^2}. \tag{4-71}$$

We then have upper bound

$$\begin{aligned} P((i,j) \in \xi_{sym}) &\leq \frac{(1-\gamma)^2}{1-2\gamma} \frac{(n-1)^{1-2\gamma} - (k_{\min}-1)^{1-2\gamma}}{(n^{1-\gamma} - k_{\min}^{1-\gamma})^2} \\ &= \frac{(1-\gamma)^2}{1-2\gamma} \frac{1}{k_{\min} n} \frac{k_{\min}^{2\gamma-1} - n^{2\gamma-1}}{(k_{\min}^{\gamma-1} - n^{\gamma-1})^2} \\ &\sim \frac{1}{k_{\min}} \end{aligned} \tag{4-72}$$

2. $\gamma = 1$**:**

$$\sum_{k=k_{\min}}^{n-1} k^{-2\gamma} \leq \int_{k_{\min}-1}^{n-1} k^{-2} dk = \frac{1}{k_{\min}-1} - \frac{1}{n-1}, \tag{4-73}$$

and

$$(\sum_{k=k_{\min}}^{n-1} k^{-\gamma})^2 \geq (\int_{k_{\min}}^{n} k^{-1} dk)^2 = \log^2 \frac{n}{k_{\min}}. \tag{4-74}$$

Then

$$\begin{aligned} P((i,j) \in \xi_{sym}) &\leq \left(\frac{1}{k_{\min}-1} - \frac{1}{n-1}\right) \frac{1}{\log^2 \frac{n}{k_{\min}}} \\ &\sim \frac{1}{k_{\min} \log^2 \frac{n}{k_{\min}}}. \end{aligned} \tag{4-75}$$

3. $\gamma \in \left(\frac{1}{2}, 1\right)$ Similar to case 1,

$$\begin{aligned} P((i,j) \in \xi_{sym}) &\leq \frac{(1-\gamma)^2}{1-2\gamma} \frac{(n-1)^{1-2\gamma} - (k_{\min}-1)^{1-2\gamma}}{(n^{1-\gamma} - k_{\min}^{1-\gamma})^2} \\ &\sim \frac{1}{k_{\min}^{2\gamma-1} n^{2-2\gamma}} \end{aligned} \tag{4-76}$$

4. $\gamma = \frac{1}{2}$

$$\begin{aligned} P((i,j) \in \xi_{sym}) &\leq \frac{\int_{k_{\min}-1}^{n-1} k^{-1} dk}{\left(\int_{k_{\min}}^{n} k^{-\frac{1}{2}} dk\right)^2} \\ &= \frac{\log \frac{n-1}{k_{\min}-1}}{4(\sqrt{n} - \sqrt{k_{\min}})^2} \\ &\sim \frac{1}{n} \log \frac{n}{k_{\min}}. \end{aligned} \tag{4-77}$$

5. $\gamma < \frac{1}{2}$. Similar to case 1.

$$P((i,j) \in \xi_{sym}) \leq \frac{(1-\gamma)^2}{1-2\gamma} \frac{(n-1)^{1-2\gamma} - (k_{\min}-1)^{1-2\gamma}}{(n^{1-\gamma} - k_{\min}^{1-\gamma})^2} \sim \frac{1}{n}. \tag{4-78}$$

Thus we complete our proof.

□

Remark: From Theorem 4.9, we can observe that lower γ and higher $k_{\min}$ makes for better guarantee of de-anonymization. This observation is in line with practice: When γ is small, most nodes are of low degree, and consider the extreme case where $k_{\min} = 1$, then a slew of nodes have degree 1, which turns out to be many 'spider' subgraphs, where one node is connected to a bunch of nodes all with degree value 1. The 'spider' structure results in a large number of symmetric nodes with degree 1. Additionally, when $k_{\min} = \Theta(n)$, the value of γ does not function in the bound, which corresponds that when $k_{\min}$ exceeds a certain value, there will not be such 'spider' structure prone to induce symmetry nodes, thus γ does not make a difference.

Table 4.4 Summarization for Theorem 4.9

γ	**Upper Bound**	**Upper Bound** $(k_{\min} = \Theta(n))$
$\left[0, \frac{1}{2}\right)$	$O\left(\frac{1}{n}\right)$	$O\left(\frac{1}{n}\right)$
$\frac{1}{2}$	$O\left(\frac{1}{n}\log\frac{n}{k_{\min}}\right)$	$O\left(\frac{1}{n}\right)$
$\left(\frac{1}{2}, 1\right)$	$O\left(\frac{1}{k_{\min}^{2\gamma-1}n^{2-2\gamma}}\right)$	$O\left(\frac{1}{n}\right)$
1	$O\left(\frac{1}{k_{\min}\log^2\frac{n}{k_{\min}}}\right)$	$O\left(\frac{1}{n}\right)$
$(1, +\infty)$	$O\left(\frac{1}{k_{\min}}\right)$	$O\left(\frac{1}{n}\right)$

4.3.2 Partly Sampled Situation: $s < 1$

In partly sampled situation, we also consider the upper bound of expected mapping error based on graph matching based approach. Before the main proof, we propose several subsidiary lemmas leveraged.

Lemma 4.3.1. *Given* $\tilde{\mathbf{\Pi}}$ *and* $\mathbf{\Pi}_0$, $G_2 = (V, E_2, \mathbf{B})$, *we have*

$$\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 = \frac{\mathbf{E}||(\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0)\mathbf{B}||_F^2}{2\mathbf{E}(d)}. \tag{4-79}$$

Proof. We can obtain

$$\begin{aligned}
&\mathbf{E}||(\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0)\mathbf{B}||_F^2 = \mathbf{E}\sum_{i=1}^{n}\sum_{j=1}^{n}((\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0)\mathbf{B})_{ij}^2 \\
&= \mathbf{E}(\sum_{i=1}^{n}\mathbf{1}\{\pi_0(i) \neq \tilde{\pi}(i)\}\sum_{j=1}^{n}(B_{\tilde{\pi}(i),j} - B_{\pi_0(i),j})^2) \\
&= \mathbf{E}(\sum_{i=1}^{n}\mathbf{1}\{\pi_0(i) \neq \tilde{\pi}(i)\}\sum_{j=1}^{n}(B_{\tilde{\pi}(i),j}^2 + B_{\pi_0(i),j}^2)) \\
&= 2\mathbf{E}(||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2)\mathbf{E}(d).
\end{aligned} \tag{4-80}$$

Then we complete our proof. □

Lemma 4.3.2. *Given* $\tilde{\mathbf{\Pi}}$ *and* $\mathbf{\Pi}_0$ *and adjacency matrices* $\mathbf{A}$ *and* $\mathbf{B}$, *we have*

$$\mathbf{E}tr((\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0)\mathbf{B}(\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0)^T\mathbf{A}) \leq \frac{1}{2}\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2\mathbf{E}|d(t) - d(s)|, \tag{4-81}$$

where $\mathbf{E}|d(t)-d(s)|$ *denotes the expectation of degree difference between two arbitrary nodes.*

Proof. Set $\mathbf{Y}=(\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0)\mathbf{B}$ and $\mathbf{X}=\mathbf{A}(\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0)$. Then we can obtain

$$\begin{aligned}&\mathbf{Etr}((\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0)\mathbf{B}(\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0)^T\mathbf{A})=\sum_{i=1}^{n}\sum_{k=1}^{n}Y_{ik}X_{ik}\\&=\sum_{i=1}^{n}\sum_{k=1}^{n}(\mathbf{B}_{t_ik}-\mathbf{B}_{s_ik})(\mathbf{A}_{it_k}-\mathbf{A}_{is_k})\\&\leq\frac{1}{2}\mathbf{E}||\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0||_F^2\mathbf{E}|d(t)-d(s)|.\end{aligned}\tag{4-82}$$

□

Theorem 4.10. *Given* $\Gamma(n,\gamma,k_{\min})$*, sampling probability* s*, the true mapping* $\mathbf{\Pi}_0$ *and the estimation* $\tilde{\mathbf{\Pi}}$*, then we can upper bound* $\mathbf{E}||\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0||_F^2$ *as Table 4.5 shows in terms of different values of* γ*.*

Proof. Recall that in power law model, the characterization focuses on the degree, rather than the probability of connection, so the proof is different from Theorem 4.3 and 4.7.

Based on Lemma 4.3.1 and 4.3.2, we can obtain

$$\begin{aligned}\mathbf{E}||\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0||_F^2&\leq\frac{\mathbf{E}||(\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0)\mathbf{B}||_F^2}{2\mathbf{E}(d)}\\&\leq\frac{2((1+\rho)\mathbf{E}||\mathbf{A}-\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T||_F^2+\mathbf{tr}((\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0)\mathbf{B}(\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0)^T\mathbf{A}))}{2\mathbf{E}(d)}\\&\leq\frac{2(1+\rho)\mathbf{E}||\mathbf{A}-\mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T||_F^2+\mathbf{E}|d(t)-d(s)|\mathbf{E}||\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0||_F^2}{2\mathbf{E}(d)}\end{aligned}\tag{4-83}$$

For $\mathbf{E}|d(t)-d(s)|$, we can obtain

$$\begin{aligned}\mathbf{E}|d(t)-d(s)|&=\sum_{x=k_{\min}}^{n-1}\sum_{y=k_{\min}}^{n-1}|x-y|\frac{(xy)^{-\gamma}}{(\sum_{k_{\min}}^{n-1}k^{-\gamma})^2}\\&=2\sum_{x=k_{\min}}^{n-1}\sum_{y=x+1}^{n-1}(y-x)\frac{(xy)^{-\gamma}}{(\sum_{k_{\min}}^{n-1}k^{-\gamma})^2}\\&=2\sum_{x=k_{\min}}^{n-1}\frac{x^{-\gamma}}{(\sum_{k_{\min}}^{n-1}k^{-\gamma})^2}\left(\sum_{y=x+1}^{n-1}y^{-\gamma+1}-x\sum_{y=x+1}^{n-1}y^{-\gamma}\right).\end{aligned}\tag{4-84}$$

We then discuss different cases of γ. Before that, we denote $C = \sum_{k=k_{\min}}^{n-1} k^{-\gamma}$ for convenience.

1. $\gamma > 2$**:**

$$\begin{aligned}
&\mathbf{E}|d(t) - d(s)| \\
&\leq \frac{2}{C^2}\left(\sum_{x=k_{\min}}^{n-1} x^{-\gamma}\left(\int_x^{n-1} y^{-\gamma+1}dy - x\int_{x+1}^{n} y^{-\gamma}dy\right)\right) \\
&= \frac{2}{C^2}\sum_{x=k_{\min}}^{n-1}\left(\frac{x^{-\gamma}}{2-\gamma}(n-1)^{2-\gamma} - \frac{x^{2-2\gamma}}{2-\gamma} - \frac{x^{1-\gamma}}{1-\gamma}n^{1-\gamma} + \frac{x^{1-\gamma}(x+1)^{1-\gamma}}{1-\gamma}\right) \\
&\leq \frac{2}{C^2}\left(\frac{k_{\min}^{2-\gamma} - (n-1)^{2-\gamma}}{\gamma-2}C - \sum_{x=k_{\min}}^{n-1}\frac{x^{1-\gamma}}{\gamma-1}(n^{1-\gamma} - (x+1)^{1-\gamma})\right) \\
&\leq \frac{2}{C}\frac{k_{\min}^{2-\gamma} - (n-1)^{2-\gamma}}{\gamma-2},
\end{aligned} \tag{4-85}$$

and

$$\mathbf{E}(d) = \frac{\sum_{k=k_{\min}}^{n-1} k^{-\gamma+1}}{C} \geq \frac{1}{C}\int_{k_{\min}}^{n} k^{-\gamma+1}dk = \frac{k_{\min}^{2-\gamma} - n^{2-\gamma}}{C(\gamma-2)}. \tag{4-86}$$

Therefore we can obtain

$$\frac{\mathbf{E}|d(t) - d(s)|}{2\mathbf{E}(d)} \leq \frac{k_{\min}^{2-\gamma} - (n-1)^{2-\gamma}}{k_{\min}^{2-\gamma} - n^{2-\gamma}} \tag{4-87}$$

Then with Eqn. (4-87), we have

$$\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 \leq \frac{(1+\rho)\mathbf{E}||\mathbf{A} - \mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T||_F^2}{\mathbf{E}(d)}\frac{k_{\min}^{2-\gamma} - n^{2-\gamma}}{(n-1)^{2-\gamma} - n^{2-\gamma}} \tag{4-88}$$

To continue our derivation, we assume the probability of connection between nodes i and j is p_{ij} in G. Therefore we can easily find that

$$\mathbf{E}||\mathbf{A} - \mathbf{\Pi}_0\mathbf{B}\mathbf{\Pi}_0^T||_F^2 = 2\sum_{1\leq i<j\leq n} p_{ij}s(1-s), \tag{4-89}$$

and

$$\mathbf{E}(d) = 2s\frac{\sum_{1\leq i<j\leq n} p_{ij}}{n}. \tag{4-90}$$

Therefore

$$\begin{aligned}\mathbf{E}||\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0||_F^2 &\le (1+\rho)n(1-s)\frac{k_{\min}^{2-\gamma}-n^{2-\gamma}}{(n-1)^{2-\gamma}-n^{2-\gamma}}\\ &= (1+\rho)n(1-s)\frac{n^{\gamma-2}}{k_{\min}^{\gamma-2}}\frac{n^{\gamma-2}-k_{\min}^{\gamma-2}}{n^{\gamma-2}-(n-1)^{\gamma-2}}\\ &\sim (1-s)\frac{n^{\gamma}}{k_{\min}^{\gamma-2}}.\end{aligned} \tag{4-91}$$

2. $\gamma = 2$**:**

$$\begin{aligned}&\mathbf{E}|d(t)-d(s)|\\ &\le \frac{2}{C^2}\left(\sum_{x=k_{\min}}^{n-1} x^{-\gamma}\log\frac{n-1}{x}-\sum_{x=k_{\min}}^{n-1}\frac{x^{1-\gamma}}{1-\gamma}(n^{1-\gamma}-(x+1)^{1-\gamma})\right)\\ &\le \frac{2}{C}\log\frac{n-1}{k_{\min}},\end{aligned} \tag{4-92}$$

and

$$\mathbf{E}(d) \ge \frac{1}{C}\int_{k_{\min}}^{n} k^{-\gamma+1}dk = \frac{1}{C}\log\frac{n}{k_{\min}}. \tag{4-93}$$

Then we can obtain

$$\frac{\mathbf{E}|d(t)-d(s)|}{2\mathbf{E}(d)} \le \frac{\log(n-1)-\log k_{\min}}{\log n-\log k_{\min}}. \tag{4-94}$$

Then based on Eqn. (4-88) we have

$$\begin{aligned}\mathbf{E}||\tilde{\mathbf{\Pi}}-\mathbf{\Pi}_0||_F^2 &\le (1+\rho)n(1-s)\frac{\log(n-1)-\log k_{\min}}{\log n-\log k_{\min}}\\ &\sim (1-s)n^2\log\frac{n}{k_{\min}}\end{aligned} \tag{4-95}$$

3. $\gamma \in (1,2)$

Similar to case 1,

$$\mathbf{E}|d(t)-d(s)| \le \frac{2}{C(2-\gamma)}((n-1)^{2-\gamma}-k_{\min}^{2-\gamma}), \tag{4-96}$$

and

$$\mathbf{E}(d) \geq \frac{1}{C(2-\gamma)}(n^{2-\gamma} - k_{\min}^{2-\gamma}). \tag{4-97}$$

We can derive

$$\begin{aligned} \mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 &\leq (1+\rho)n(1-s)\frac{n^{2-\gamma} - k_{\min}^{2-\gamma}}{n^{2-\gamma} - (n-1)^{2-\gamma}} \\ &\sim (1-s)n^2 \end{aligned} \tag{4-98}$$

4. $\gamma = 1$

$$\begin{aligned} \mathbf{E}|d(t) - d(s)| &= \frac{2}{C^2}\left(\sum_{x=k_{\min}}^{n-1} x^{-\gamma}((n-x-1) - x\sum_{x+1}^{n-1} y^{-1})\right) \\ &\leq \frac{2}{C^2}\left(\sum_{x=k_{\min}}^{n-1} x^{-\gamma}((n-x-1) - x\int_{x+1}^{n-1} y^{-1}dy)\right) \\ &= \frac{2}{C^2}\left(\sum_{x=k_{\min}}^{n-1} x^{-\gamma}\left(n-x-1-x\log\frac{n}{x+1}\right)\right) \\ &\overset{\circ}{\leq} \frac{2}{C}(n - k_{\min} - 1 - k_{\min}\log\frac{n}{k_{\min}+1}), \end{aligned} \tag{4-99}$$

where $\overset{\circ}{\leq}$ is due to the monotonically decreasing property of $\left(n-x-1-x\log\frac{n}{x+1}\right)$. Meanwhile we note that

$$\mathbf{E}(d) = \frac{n-1-k_{\min}}{C}. \tag{4-100}$$

Therefore

$$\frac{\mathbf{E}|d(t) - d(s)|}{2\mathbf{E}(d)} \leq 1 - \frac{k_{\min}\log\frac{n}{k_{\min}+1}}{n-1-k_{\min}}, \tag{4-101}$$

and thus

$$\begin{aligned} \mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 &\leq (1+\rho)n(1-s)\frac{n-1-k_{\min}}{k_{\min}\log\frac{n}{k_{\min}+1}} \\ &\sim \frac{n^2}{k_{\min}\log\frac{n}{k_{\min}}}(1-s) \end{aligned} \tag{4-102}$$

Table 4.5 Summarization for Theorem 4.10

γ	**Upper Bound**	**Upper Bound** ($k_{\min} = \Theta(n)$)
$[0, 1)$	$O(n(1-s))$	$O(n(1-s))$
1	$O\left(\frac{n^2}{k_{\min} \log \frac{n}{k_{\min}}}(1-s)\right)$	$O(n(1-s))$
$(1, 2)$	$O(n(1-s)^2)$	$O(n(1-s)^2)$
2	$O\left((1-s)n^2 \log \frac{n}{k_{\min}}\right)$	$O\left((1-s)n^2\right)$
$(2, +\infty)$	$O\left((1-s)\frac{n^\gamma}{k_{\min}^{\gamma-2}}\right)$	$O\left((1-s)n^2\right)$

5. $\gamma \in [0, 1)$

$$\begin{aligned}
&\mathbf{E}|d(t) - d(s)| \\
&\leq \frac{2}{C^2}\left(\sum_{x=k_{\min}}^{n-1} x^{-\gamma}\left(\int_{x+1}^{n} y^{-\gamma+1}dy - x\int_{x+1}^{n} y^{-\gamma}dy\right)\right) \\
&= \frac{2}{C^2}\sum_{x=k_{\min}}^{n-1} x^{-\gamma}\left(\frac{n^{2-\gamma} - (x+1)^{2-\gamma}}{2-\gamma} - \frac{x()}{1-\gamma}\right) \\
&\leq \frac{2}{C^2}\left(\sum_{x=k_{\min}}^{n-1} x^{-\gamma}\left(\frac{(x+1)^{2-\gamma}}{(1-\gamma)(2-\gamma)} + \frac{n^{2-\gamma}}{2-\gamma} - \frac{xn^{1-\gamma}}{1-\gamma}\right)\right) \\
&\leq \frac{2}{C}\frac{n^{1-\gamma}(n-k_{\min})}{1-\gamma},
\end{aligned} \tag{4-103}$$

and

$$\mathbf{E}(d) \geq \frac{(n-1)^{2-\gamma} - (k_{\min}-1)^{2-\gamma}}{C(2-\gamma)}. \tag{4-104}$$

Then

$$\frac{\mathbf{E}|d(t) - d(s)|}{2\mathbf{E}(d)} \leq \frac{2-\gamma}{1-\gamma}\frac{n^{1-\gamma}(n-k_{\min})}{(n-1)^{2-\gamma} - (k_{\min}-1)^{2-\gamma}}, \tag{4-105}$$

and we can obtain

$$\begin{aligned}
\mathbf{E}||\tilde{\mathbf{\Pi}} - \mathbf{\Pi}_0||_F^2 &\leq (1+\rho)n(1-s)\frac{1}{1 - \frac{2-\gamma}{1-\gamma}\frac{n^{1-\gamma}(n-k_{\min})}{(n-1)^{2-\gamma}-(k_{\min}-1)^{2-\gamma}}} \\
&\sim n(1-s)
\end{aligned} \tag{4-106}$$

To present the result more clearly, we summarize the result in Table 4.5. Thus we complete the proof.

□

Remark: Shown in Theorem 4.5, we can discover that γ plays negatively on the de-anonymization result when γ rises up, while $k_{\min}$ functions more positively as it increases. Both results correspond nicely with fully sampled situation. Additionally, we can also discover that the performance of Power-Law based model does not outperform E-R graph model, with the upper bound as $O(n(1-s))$ and lower bound $\Omega(n(1-s)^2)$. This is because Power-Law graph reduces the freedom of the original totally random E-R graph by restriction on degree distribution, which arouses more topological symmetries and thus induces more hardness to distinguish the identity of nodes with high symmetric property.

4.4 Algorithm Design

From the theoretical analysis above, we gain a fundamental view of de-anonymizability in three basic social network models, under both fully and partly sampled situations. All results show that parameters in each model mutually decide whether the adversary can successfully identify anonymized users. However, even if parameter values make successful de-anonymization available, the adversary has to contrive of an available scheme which can literally manage the de-anonymization. This means that we need to bridge the gap between theory and practice via designing algorithms to verify that theoretical bounds in different models for successful de-anonymization are **achievable** practically. Meanwhile, the adversary is faced with large-scale social networks nowadays under the social network explosion, so this algorithm should be **efficient** enough, i.e., the time complexity is not prohibitively high.

Several algorithms have been proposed for de-anonymization. However, they do not measure up to the requirement of achievability and efficiency as we consider in our work: For achievability, prior algorithms shed light on a specific model, for example (overlapping) stochastic block model, and only give algorithm guarantee under a series of conditions not so fundamental, that is, these conditions do not explicitly unveil the effect of parameters but perform as if a paving stone for mathematical derivation in

a form not so intuitive. For efficiency, these algorithms only achieve $O(n^6)$, which, although polynomial, inhibits its practical use in large-scale social networks.

Unlike the previous work which generally forms this problem as an optimization and design algorithms above to solve it with outputting the de-anonymization result in a batch, we take a novel perspective to explore the property of each node specifically and map the nodes in two networks in an online manner: starting from mapping one node with high distinguishability, and mapping other nodes based on previous mapping results.

The reason we consider specific exploration in lieu of the general optimization is also based on our consideration of **achievability** and **efficiency**: Firstly, specific exploration can leverage the attribute of each node, for example degree and adjacent nodes, in a more explicit way, rather than general optimization mingling the information of the whole network together which makes the individual information implicit. This novel idea corresponds to our focus on the issue of achievability since we derive theoretical bounds largely relying on extracting node and connection properties including node symmetry, node degree, connection probability, etc. Secondly, specific exploration microscopically studies each node and its neighbors in each round, eliminating the huge cost aroused by macroscopical operations on the whole network in general optimization, for example the large cost to deal with the whole adjacency matrix of G_1 and G_2 in optimization. Concretely, these panoramic operations in general optimization arouses great redundancy, like in sparse graph, there are a slew of useless 0s but still melted into consideration.

To implement our specific exploration idea, we principally concentrate on two node attributes: *node degree* and *adjacent degree series*. Node degree is intuitive, while adjacent degree series means the sorted array containing degree value of each neighbor of our target node. These two attributes provide us useful and easily available information to map nodes in G_1 and G_2: If node $v_1 \in G_1$ and $v_2 \in G_2$ are of similar node degree and similar adjacent degree series, then chances are that they represent the same network user. The validty of focusing on these two attributes echoes the fact that in most cases, sampling probability could not deviate much from 1: Almost negligible

connections are discarded compared with the whole connections in G during sampling, so most information about degree and adjacent nodes in G_1 and G_2 are preserved from G.

Specifically, as studying node degree costs less than adjacent degree series, we firstly sort the nodes by degree value nonincreasingly in G_1 and G_2 respectively, and pick up the node with largest degree in G_1, denoted as v_1^1. Correspondingly, we build a *candidate* set of nodes in G_2 consisting of K largest node degree, and we will select the value of K such that the correct mapping of v_1^1 in G_2 lies in the candidate set. After that, we check the adjacency degree series for all K nodes in the candidate set, and select one with the closest 'distance' to that of v_1^1 by our defined metric as our mapping of v_1^1, denoted as v_1^2, and thus a round of mapping ends. We mark those mapped nodes and continue to do this iteratively until all nodes have been mapped. We implement this two-step macro-to-micro idea in Algorithm 4.1.

In Algorithm 4.1, there are some steps that need explanations:

- Step 2: Initialize the mapping;
- Step $3, 4$: Initialize the mark status of nodes in G_1 and G_2;
- Step 8: For the first round, we build the K-size candidate set, while in following rounds, we only need to update one node each round to fill the vacancy aroused by the mapped node last round.
- Step 9: $CurrentDist$ records the current minimum distance between our v_j and v_i^1 based on our defined metric $F(v_i^1, v_j)$.
- Step 12: $F(v_i^1, v_j)$ is our defined metric to evaluate the difference of adjacent degree series between two nodes in two networks. The first term $||Adj(v_i^1) - Adj(v_j)||_1$ denotes the 1-norm of the difference between the (sorted) adjacent degree series of v_i^1 and v_j. The 1-norm means the sum of absolute value of each element. However, it is possible that $Adj(v_i^1)$ and $Adj(v_j)$ are of different length as the degree of v_i^1 and v_j may be different. To tackle this, we can supplement the shorter vector with 0 s to make two vectors have the same length.

Algorithm 4.1: Candidate Set based De-anonymization Algorithm (CASDA)

Input: $G_1 = (V, E_1)$, $G_2 = (V, E_2)$; Parameter μ.
Output: Estimated mapping $\hat{\mathbf{\Pi}}$.

Sort all nodes in V in G_1 and G_2 by their degree values non-increasingly, and denote the sorted node set as $\tilde{V}_1$ and $\tilde{V}_2$ respectively;
$\hat{\mathbf{\Pi}} \leftarrow \mathbf{0}_{n\times n}$;
$\mathbf{Mark}_1 \leftarrow \mathbf{0}_{n\times 1}$;
$\mathbf{Mark}_2 \leftarrow \mathbf{0}_{n\times 1}$;
for $i = 1$ *to* n **do**
    Pick unmarked node with largest degree in $\tilde{V}_1$, denoted as v_i^1;
    Calculate and sort $Adj(v_i^1)$ non-increasingly;
    Build/Update the K-size candidate set in $\tilde{V}_2$, denoted as $CAND$;
    Set $CurrentDist \leftarrow \infty$;
    for *every node* v_j *in* $CAND$ *and* $\mathbf{Mark}_2(v_j) = 0$ **do**
        Calculate and sort $Adj(v_j)$ non-increasingly;
        Calculate $F(v_i^1, v_j) = ||Adj(v_i^1) - Adj(v_j)||_1 + \mu||\#_{adj}(v_i^1, v_j)||$;
        if $F(v_i^1, v_j) < CurrentDist$ **then**
            $CurrentDist \leftarrow F(v_i^1, v_j)$;
            Set $v_i^2 \leftarrow v_j$;
    $\hat{\mathbf{\Pi}}(v_i^1, v_i^2) \leftarrow 1$;
    $\mathbf{Mark}_1(v_i^1) \leftarrow 1$;
    $\mathbf{Mark}_2(v_i^2) \leftarrow 1$;
return $\hat{\mathbf{\Pi}}$.

We can randomly insert these 0 inside the short vector since we do not know exactly which connection have gotten lost during the sampling. The second term $\mu||\#_{adj}(v_i^1, v_j)||$ is a penalty, where $\#_{adj}(v_i^1, v_j)$ denotes the number of different *mapped* nodes in the neighbors of v_i^1 and v_j. For example, in the neighbors of v_i^1 $(i \geq 6)$, there are three mapped nodes: v_1^1, v_3^1, v_5^1, while in the neighbors of v_j $(j \geq 6)$, there are two: v_1^2, v_4^2, then $||\#_{adj}(v_i^1, v_j)|| = 3$ since there are three differently mapped nodes v_3^1, v_4^2, v_5^1. Only v_1^1 and v_1^2 are mapped to be estimated to represent the same user. μ is a weight controller. The adoption of this penalty form helps the algorithm leverage results in previous round properly to guide the decision in current round, which is in an online manner. Certainly, this guidance works if the previous mapped nodes are correct, which we will show later.

- Step $13, 14, 15$: Update $CurrentDist$ and our estimated v_i^2.

- Step $16, 17, 18$: Mark the mapped nodes in $\hat{\mathbf{\Pi}}$ and record the mapped nodes in $\mathbf{Mark}_1$ and $\mathbf{Mark}_2$.

Time Complexity: $O(Kn^2 \log n)$. Specifically,

- Sorting the nodes based on degree: $O(n \log n)$;
- Select the node with largest degree: $O(1)$;
- Build/Update the candidate set: $O(K)$/$O(1)$;
- Sort $Adj(v_j)$: $O(n \log n)$;
- Calculate $F(v_i^1, v_j)$: $O(n)$;

The main cost is sorting $Adj(v_j)$ inside two outer loops with $O(nK)$ rounds, so the time complexity of Algorithm 4.1 os $O(Kn^2 \log n)$, which shows the *efficiency* of our proposed algorithm. Note that different ways to calculate $||Adj(v_i^1) - Adj(v_j)||_1$ will arouse difference in time complexity. For example, if inserting 0s in different slots of the short vector multiple times and take the average, then it costs more time.

Performance Guarantee: Hereinafter we prove the *achievability* of our theoretical results by showing that Algorithm 4.1 can output the correct de-anonymization result when parameters in different models take values that lies in our derived intervals for successful de-anonymization.

Theorem 4.11. *If Algorithm 4.1 can ensure that in every iteration, the size of candidate set $K > 2$, then it works as a bound-achievable de-anonymization algorithm for E-R graph model, stochastic block model, and power-law model.*

Proof. The main thread of our proof follows 2 major parts in Algorithm 4.1 in order: (i) The K-size candidate set needs to include the correct mapping of v_i^1 in each iteration; (ii) Minimizing our defined distance metric $F(v_i^1, v_j)$ enables us to obtain the correct mapping of v_i^1 with probability arbitrarily close to 1.

Note that in Algorithm 4.1 if in the t^{th} step we can ensure that the previous $t-1$ steps output correct mapping with probability 1, then the t^{th} step can also do that as each

step is of nearly same operations except that the term $\#_{adj}(v_i^1, v_j)$ is updated. However since previous $t-1$ steps are a.s. correct, it can ensure that $\#_{adj}(v_i^1, v_j)$ will a.s. not lead us to a wrong direction. Then in the following, we simply consider the first step and prove that it maps v_i^1 to the correct node in G_2 a.s..

Step 1: Building the candidate set.

The key attribute involved in building the candidate set is *node degree*. Note that due to the sampling process, G_1 and G_2 are not entirely the same when $s < 1$. If we consider G_1 and G_2 separately, we will fail to characterize the degree relationship between G_1 and G_2, which is what de-anonymization depends on. Therefore we bridge them by the underlying graph G in our proof (Note that in real case the adversary does not know G.).

We sort the nodes in G, G_1, G_2 respectively. Specifically, in each graph, we reorder the indexes of each node as

- G: $d_G^1 \geq d_G^2 \geq \cdots \geq d_G^n$;
- G_1: $d_{G_1}^1 \geq d_{G_1}^2 \geq \cdots \geq d_{G_1}^n$;
- G_2: $d_{G_2}^1 \geq d_{G_2}^2 \geq \cdots \geq d_{G_2}^n$.

Then we study the probability that the correct mapping of v_i^1 lies in the candidate set. Specifically, we use $d_{G_1}^1 \Leftrightarrow d_{G_2}^i$ to denote that the nodes corresponded to $d_{G_1}^1$ and $d_{G_2}^i$ are mapped.

$$\begin{aligned} P(\exists k \in CAND, s.t. d_{G_1}^1 \Leftrightarrow d_{G_2}^k) &= \sum_{i=1}^{|CAND|} P(d_{G_1}^1 \Leftrightarrow d_{G_2}^i) \\ &= 1 - \sum_{|CAND|+1}^{n} P(d_{G_1}^1 \Leftrightarrow d_{G_2}^i) \\ &= 1 - \sum_{|CAND|+1}^{n} \sum_{j=1}^{n} P(d_{G_1}^1 \Leftrightarrow d_G^j, d_{G_2}^i \Leftrightarrow d_G^j) \\ &= 1 - \sum_{|CAND|+1}^{n} \sum_{j=1}^{n} P(d_{G_1}^1 \Leftrightarrow d_G^j) P(d_{G_2}^i \Leftrightarrow d_G^j). \end{aligned} \tag{4-107}$$

Meanwhile, note that for a node v in G with degree d_0, then after sampling, we can obtain that in G_1 the degree of v, denoted as d_0^1, is of distribution $P(d_0^1 = d_0 - x) =$

$\binom{d_0}{x}(1-s)^x s^{d_0-x}$, where we set x as the loss of degree during the sampling. Similarly in G_2, the degree of v, denoted as d_0^2, is of distribution $P(d_0^2 = d_0 - y) = \binom{d_0}{y}(1-s)^y s^{d_0-y}$.

Meanwhile, note that for $d_{G_1}^1$ in G_1 ($d_{G_2}^1$ in G_2 similarly), if the case $d_{G_1}^1 \Leftarrow d_G^j$ happens, it must ensure that the difference of degree loss of v_1^1 and v_j must be larger than $d_G^1 - d_G^j$. We use $del(d_G^1)$ to denote the degree loss of the node corresponding to d_G^1.

Therefore we can obtain

$$\begin{aligned}
&P(d_{G_1}^1 \Leftrightarrow d_G^j) \le P(del(d_G^1) - del(d_G^j) \ge d_G^1 - d_G^j) \\
&= \sum_{k=0}^{d_G^j} P(del(d_G^j) = k) P(del(d_G^1) \ge k + d_G^1 - d_G^j) \\
&= \sum_{k=0}^{d_G^j} \binom{d_G^j}{k}(1-s)^k s^{d_G^j - k} \left(1 - \sum_{t=0}^{k+d_G^1-d_G^j-1} \binom{d_G^1}{t} s^t (1-s)^{d_G^1 - t}\right) \\
&= 1 - \sum_{k=0}^{d_G^j} \left(\binom{d_G^j}{k}(1-s)^k s^{d_G^j - k} \sum_{t=0}^{k+d_G^1-d_G^j-1} \binom{d_G^1}{t} s^t (1-s)^{d_G^1 - t}\right).
\end{aligned} \tag{4-108}$$

Similarly, for the case of $d_{G_2}^i \Leftrightarrow d_G^j$, we can derive

$$\begin{aligned}
&P(d_{G_2}^i \Leftrightarrow d_G^j) \\
&\le 1 - \sum_{k=0}^{d_G^j} \left(\binom{d_G^j}{k}(1-s)^k s^{d_G^j - k} \sum_{t=0}^{k+d_G^i-d_G^j-1} \binom{d_G^i}{t} s^t (1-s)^{d_G^i - t}\right).
\end{aligned} \tag{4-109}$$

Then based on Eqn. (4-107), (4-108), and (4-109), we have

$$\begin{aligned}
&P(\exists k \in CAND, s.t. d^1_{G_1} \Leftrightarrow d^k_{G_2}) \\
&\geq 1 - \sum_{i=|CAND|+1}^{n} \sum_{j=i}^{n} \\
&\left(1 - \sum_{k=0}^{d^j_G} \left(\binom{d^j_G}{k} (1-s)^k s^{d^j_G - k} \sum_{t=0}^{k+d^1_G - d^j_G - 1} \binom{d^1_G}{t} s^t (1-s)^{d^1_G - t} \right) \right) \\
&\left(1 - \sum_{k=0}^{d^j_G} \left(\binom{d^j_G}{k} (1-s)^k s^{d^j_G - k} \sum_{t=0}^{k+d^i_G - d^j_G - 1} \binom{d^i_G}{t} s^t (1-s)^{d^i_G - t} \right) \right),
\end{aligned} \tag{4-110}$$

where $\sum_{j=i}^{n}$ instead of $\sum_{j=0}^{n}$ since for $j < i$, $P(d^i_{G_2} \Leftrightarrow d^j_G) = 0$.

Note that we can observe that $P(d^1_{G_1} \Leftrightarrow d^j_G) < P(d^i_{G_2} \Leftrightarrow d^j_G)$ since $s = 1 - o(1)$, and we can discover that $k = 0$ determines the order of $\sum_{i=|CAND|+1}^{n} \sum_{j=i}^{n} P(d^1_{G_1} \Leftrightarrow d^j_G) P(d^i_{G_2} \Leftrightarrow d^j_G)$. Therefore we have when $s = 1 - o(1)$, if $|d^1_G - d^{|CAND|+1}_G| > 2$, we can ensure that when $n \to \infty$, $P(\exists k \in CAND, s.t. d^1_{G_1} \Leftrightarrow d^k_{G_2})$ is arbitrarily close to 1 a.s..

Step 2: Finding the mapping of v^1_1 in the candidate set

In this step, we are to show that the way we find the mapping of v^1_1 in Algorithm 4.1 will output the mapping correctly. Specifically, we use i to denote v^1_1, and $\pi_0(i)$ to denote the correct mapping of v^1_1 in G_2. Then we need to show $P(\pi_0(i) = \arg\min_{j \in K} ||Adj(i) - Adj(j)||_1) \to 1$. To prove that, we have

$$\begin{aligned}
&P(\pi_0(i) = \arg\min_{j \in K} ||Adj(i) - Adj(j)||_1) \\
&= P(\forall j : ||Adj(i) - Adj(\pi_0(i))||_1 \leq ||Adj(i) - Adj(j)||_1) \\
&= P(\cap_{j \in K \setminus \{\pi_0(i)\}} (||Adj(i) - Adj(\pi_0(i))||_1 \leq ||Adj(i) - Adj(j)||_1)) \\
&= 1 - P(\cup_{j \in K \setminus \{\pi_0(i)\}} (||Adj(i) - Adj(\pi_0(i))||_1 > ||Adj(i) - Adj(j)||_1)).
\end{aligned} \tag{4-111}$$

Then we just need to show $P(\cup_{j \in K \setminus \{\pi_0(i)\}} (||Adj(i) - Adj(\pi_0(i))||_1 > ||Adj(i) -$

$Adj(j)||_1)) \to 0$. Notice that

$$\begin{aligned} &P(||Adj(i) - Adj(\pi_0(i))||_1 > ||Adj(i) - Adj(j)||_1) \\ &= \sum_{t>0} P(||Adj(i) - Adj(\pi_0(i))||_1 = t) P(||Adj(i) - Adj(j)||_1 < t). \end{aligned} \tag{4-112}$$

Then we define the following three parameters:

- $d(i)$: The degree of node i;
- $b(i)$: The number of edges among the adjacent nodes of i;
- $c(i)$: The number of edges connected to other nodes (except i and all adjacent nodes of i) from the adjacent nodes.

Then consider when $s = 1 - o\left(\frac{1}{n}\right)$, by eliminating all the higher order terms approaching 0, we can show that

$$\begin{aligned} &P(||Adj(i) - Adj(\pi_0(i))||_1 = 1) \\ &\sim \left(\binom{d(i)}{0} (1-s)^0 s^{d(i)} \binom{b(i)}{0} (1-s)^0 s^{b(i)} \right)^2 \binom{2c(i)}{1} (1-s)^1 s^{2c(i)-1}; \end{aligned} \tag{4-113}$$

$$\begin{aligned} &P(||Adj(i) - Adj(\pi_0(i))||_1 = 2) \\ &\sim \left(\binom{d(i)}{0} (1-s)^0 s^{d(i)} \binom{c(i)}{0} (1-s)^0 s^{c(i)} \right)^2 \binom{2b(i)}{1} (1-s)^1 s^{2b(i)-1}; \end{aligned} \tag{4-114}$$

$$\begin{aligned} &P(||Adj(i) - Adj(\pi_0(i))||_1 = d(j)) \\ &\sim \left(\binom{c(i)}{0} (1-s)^0 s^{c(i)} \binom{b(i)}{0} (1-s)^0 s^{b(i)} \right)^2 2(1-s)^1 s^{2d(i)-1}, \end{aligned} \tag{4-115}$$

where j is an adjacent node of i. Therefore

$$\begin{aligned} &P(||Adj(i) - Adj(\pi_0(i))||_1 > ||Adj(i) - Adj(j)||_1) \\ &\sim 2c(i)(1-s)s^{2(b(i)+c(i)+d(i))-1}P(||Adj(i) - Adj(j)||_1 < 1) \\ &\quad + 2b(i)(1-s)s^{2(b(i)+c(i)+d(i))-1}P(||Adj(i) - Adj(j)||_1 < 2) \\ &\quad + 2(1-s)s^{2(b(i)+c(i)+d(i))-1}\sum_{t=1}^{d(i)} P(||Adj(i) - Adj(j)||_1 < d_t). \end{aligned} \tag{4-116}$$

In terms of $P(||Adj(i) - Adj(j)||_1 < 1)$, we note that there are two possibilities for node i and j in the underlying graph G.

- $d(i) = d(j)$;
- $|d(i) - d(j)| = 1$ (Set $d(i) < d(j)$).

The probability of the case $d(i) = d(j)$ is $P(||Adj(i) - Adj(j)||_1 < 1 \mid d(i) = d(j)) = \binom{d(i)}{0}(1-s)^0 s^{d(i)}\binom{d(j)}{0}(1-s)^0 s^{d(j)} = s^{d(i)+d(j)}$. The probability of the case $|d(i) - d(j)| = 1$ is $P(||Adj(i) - Adj(j)||_1 < 1 \mid d(j) - d(i) = 1) = \binom{d(i)}{0}(1-s)^0 s^{d(i)}\binom{d(j)}{1}(1-s)^1 s^{d(j)} = d(j)(1-s)s^{d(i)+d(j)-1} = o(s^{d(i)+d(j)})$ as $s = 1 - o\left(\frac{1}{n}\right)$. Therefore the main term we need to tackle is $P(||Adj(i) - Adj(j)||_1 < 1 \mid d(i) = d(j)))$.

For E-R graph model, we can show that

$$P(d(i) = d(j)) = \sum_{k=0}^{n-2}\left(\binom{n-2}{k}p^k(1-p)^{n-2-k}\right)^2 = O\left(\frac{1}{\sqrt{n}}\right). \tag{4-117}$$

Therefore

$$P(d(i) = d(j))P(||Adj(i) - Adj(j)||_1 < 1 \mid d(i) = d(j))) \sim \frac{1}{\sqrt{n}}s^{d(i)+d(j)}. \tag{4-118}$$

As for $d(i)$, $b(i)$, and $c(i)$, we have $\mathbf{E}(d(i)) = (n-1)p$, $\mathbf{E}(b(i)) = \frac{n(n-1)}{2}p$,

$\mathbf{E}(c(i)) = (n-1)^2 p$. Therefore

$$\begin{aligned} &P(||Adj(i) - Adj(\pi_0(i))||_1 > ||Adj(i) - Adj(j)||_1) \\ &= 2c(i)(1-s)s^{2(b(i)+c(i)+d(i))-1} \sum_{k=0}^{n-2} \left(\binom{n-2}{k} p^k (1-p)^{n-2-k} \right)^2 s^{d(i)+d(j)}. \end{aligned} \tag{4-119}$$

Then we need to show that when $p = \omega\left(\frac{1}{n}\right)$, $\sum_{k=0}^{n-2} \left(\binom{n-2}{k} p^k (1-p)^{n-2-k} \right)^2$ is not of constant order. However, this is very easily by asymptotic analysis by setting $p = \frac{1}{n^{1-\epsilon}}$ where $\epsilon > 0$, we can show that

$$\begin{aligned} \sum_{k=0}^{n-2} \left(\binom{n-2}{k} p^k (1-p)^{n-2-k} \right)^2 &= \sum_{k=0}^{n-2} \left(\binom{n-2}{k} \frac{1}{n^{(1-\epsilon)k} \left(1 - \frac{1}{n^{1-\epsilon}}\right)^{n-2-k}} \right)^2 \\ &\sim \sum_{k=0}^{n-2} \frac{n^{k\epsilon+1}}{e^{n\epsilon}} \to 0. \end{aligned} \tag{4-120}$$

Similarly we can show that $P(||Adj(i) - Adj(j)||_1 < 2)$ is not of constant order, and thus base on Eqn. (4-116), we complete the proof of E-R graph model.

□

Chapter 5 Conclusion

In this thesis, we target at social network de-anonymization problem in a longitudinal and lateral manner respectively:

Longitudinally, we tackle de-anonymization in overlapping stochastic block model, a more practical model than used in prior 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.

Laterally, we take a panoramic view of de-anonymizability in different common models, including Erdos-Renyi model, stochastic block model, and power law model. We theoretically unveil that theoretical bounds exist for each model distinguishing whether the network can be successfully de-anonymized or not in both fully and partly sampled situation. For E-R model $G(n,p)$, we show that $p = \Theta\left(\frac{1}{n}\right)$ acts as a phase transition point in fully sampled case, while in partly sampled case, when sampling probability $s = 1 - o\left(\frac{1}{n}\right)$ then expected mapping error will vanish to 0 as network size goes to infinity. For stochastic block model $SBM(n,p,q)$, the de-anonymization result is mutually determined by p and q in fully sampled case, while in partly sampled case, higher p benefits more than higher q, and when $p > q$, higher homogeneity of community size distribution will make for higher de-anonymization accuracy. For power law model $\Gamma(n,\gamma,k_{\min})$, lower γ and higher $k_{\min}$ promise better de-anonymization result. Moreover, generally with given sampling probability s, de-anonymizability: SBM>ER>Power Law. Furthermore, we propose the Candidate Set based De-anonymization Algorithm (CASDA) to show that our derived thresholds are achievable.

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] A. Decelle, F. Krzakala, C. Moore and L. Zdeborov, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications" in *Physical Review E*, No. 84, Vol. 6, pp. 066106, 2011.

[20] G. H. Hardy, J. E. Littlewood and G. Plya, "Inequalities. Reprint of the 1952 edition." in *Cambridge Mathematical Library*, 1988

[21] 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.

[22] 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.

[23] 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.

[24] 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.

Acknowledgements

Firstly, I would like to give my greatest thanks to my advisor to Professor Xinbing Wang for his patient and insightful instruction on my research and also my future career. I am deeply impressed by Professor Wang's farsighted perspective in distinguishing what is a good topic, a topic worth paying efforts on. I am also very appreciated that Professor Wang has offered me many opportunities other than research, for example, attending the ACM Turc in Shanghai twice, going to Hawaii to attend the IEEE Internation Conference on Computer Communications (INFOCOM) in 2018, and leading international gurus around SJTU.

Meanwhile, I am very grateful for two research advisors, Professor Xiaohua Tian and Professor Luoyi Fu, who had helped me a lot in doing my research work. I learned so much on how to select topics, design algorithms, and write papers with strong logic from them. I can never forget those nights during which they helped me revise my papers. I really appreciate their altruistic care for me and for all students in our group.

I would also express my thanks to all seniors and peers in the group of Intelligent Internet of Things (IIoT). When I started to get into IIoT, the seniors helped me a lot in getting accustomed to the environment in the lab, and becoming more versed in reading and writing academic papers and reports. Specially, I would like to greatly thank Mr. Shitao Li, Mr. Wenxin Li, Mr. Xinzhe Fu, and Mr. Zhongzhao Hu, who had brought many suggestions on my research.

I take this opportunity to express gratitude to the all my instructors in Shanghai Jiao Tong University, especially the faculties in the Department of Electronic Engineering (1st major) and Mathematics (2nd major). I can never build such solid and comprehensive mathematical and programming background without your inculcation in different courses.

Also, I would express my gratitude to the School of Electronic information and Electrical Engineering, and Shanghai Jiao Tong University, for providing me with such stable and excellent education resources, diverse opportunities, and spiritual replenish-

ment.

I also would like to show my gratefulness to my family, especially my parents. They always support me when I am down, and offer me instructive advice making for my smooth development.

Moreover, I would like to greatly thank the supports from all the projects and funds. This work was supported by NSF China (No. 61532012, 61325012, 61521062, 61602303 and 91438115).

Papers Published During the Study for Bachelor's Degree

[1] **Xinyu Wu**, Zhongzhao Hu, Xinzhe Fu, Luoyi Fu, Xinbing Wang, and Songwu Lu, "Social Network De-anonymization with Overlapping Communities: Analysis, Algorithm and Experiment," in *IEEE International Conference on Computer Communications (INFOCOM)*, 2018.

[2] **Xinyu Wu**, Xiaohua Tian, and Xinbing Wang, "Large-scale Wireless Fingerprints Prediction for Cellular Network Positioning," in *IEEE International Conference on Computer Communications (INFOCOM)*, 2018.

[3] Xiaohua Tian, Wencan Zhang, Jingchao Wang, Wenxin Li, Shitao Li, **Xinyu Wu**, and Yucheng Yang, "Online Pricing Crowdsensed Fingerprints for Accurate Indoor Localization", *The* 86^{th} *IEEE Vehicular Technology Conference (VTC)*, 2017.