De-anonymizability of Social Network: Through the Lens of Symmetry

Benjie Miao, Shuaiqi Wang, Luoyi Fu
Shanghai Jiao Tong University
[bjmiao,wangshuaiqi,yiluofu]@sjtu.edu.cn

Xiaojun Lin
Purdue University
linx@ecn.purdue.edu

ABSTRACT

Social network de-anonymization, which refers to re-identifying users by mapping their anonymized network to a correlated network, is an important problem that has received intensive study in network science. However, it remains less understood how network structural features intrinsically affect whether or not the network can be successfully de-anonymized. To find the answer, this paper offers the first general study on the relation between de-anonymizability and network symmetry. To this end, we propose to capture the symmetry of a graph by the concept of graph bijective homomorphism. By defining the matching probability matrix, we are able to characterize the de-anonymizability, i.e., the expected number of correctly matched nodes. Specifically, we show that for a graph pair with arbitrary topology, the de-anonymizability is equal to the maximal diagonal sum of the matching probability matrix generated from homomorphisms. Due to the prohibitive cost of enumerating all possible homomorphisms, we further obtain an upper bound of such de-anonymizability by counting the orbits of each of the two graphs, which significantly reduces the computational cost. Such a general result allows us to theoretically obtain the de-anonymizability of any networks with more specific topology structure. For example, for any classic Erdős-Rènyi graph with designated n and p, we can represent its de-anonymizability numerically by calculating the local symmetric structure that it contains. Extensive experiments are performed to validated our findings.

CCS CONCEPTS

• **Security and privacy** → *Privacy protections*; • **Networks** → **Network privacy and anonymity**; • **Theory of computation** → *Random network models.*

KEYWORDS

Social De-anonymization, De-anonymizability, Symmetry, Graph Automorphism, Graph Bijective Homomorphism

ACM Reference Format:
Benjie Miao, Shuaiqi Wang, Luoyi Fu and Xiaojun Lin. 2020. De-anonymizability of Social Network: Through the Lens of Symmetry. In *Proceedings of International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (Mobihoc '20).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Mobihoc '20, October 11-14, 2020, Online

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As the popularity of social networks increases, the privacy of personal information in social networks becomes an issue of great concern. Concealing personal identity in social network is one of the most common methods to protect personal information, but it is insufficient for privacy protection since adversaries may use correlated side information across multiple networks to uncover the identity of anonymous user. Such re-identification process using auxiliary correlated information is called *Social Network De-anonymization*. The problem is initially proposed by Narayanan and Shimatikov [16]. In the past decades, a large number of works [16][19][14][11][3][13][9][18][24][4][20] have emerged focusing on the de-anonymization problem with different aspect.

In this work, our focus is on an important branch of de-anonymization problem: *seedless de-anonymization*. In seedless de-anonymization, the attacker needs to re-identify the user identity in a published network using an auxiliary network with full identity information. **The published network is completely anonymous, where no pre-identified nodes (i.e. the so-called seeds) are given.** The correlation between two networks only lies in the similarity in their topology, since these two networks are supposed to be from the same underlying relationship network. The attacker aims to uncover the identities in this published anonymous network by matching the users in the published network to those in the auxiliary network.

Various algorithms for seedless de-anonymization have been proposed [16][19][14][13][9][18][24][4]. Unfortunately, such algorithms may occasionally fail to perfectly de-anonymize the published network due to the natural characteristics of the network itself. Further, many works [3][11][9] have discussed that under some circumstances, **no algorithm** can successfully re-identify the users in the network. We thus use the term **de-anonymizability** to describe the accuracy with which a de-anonymization attack can (at most) achieve upon certain network.

However, it has not yet been well understood how network structural features intrinsically affect the de-anonymizability of the graph. Most previous works focused on proposing certain algorithms to solve de-anonymization problems in the context of some network model, and even analyzed their performance for certain classes of networks. However, to the best of our knowledge, no previous work gives comprehensive analysis on the phenomenon that some kinds of networks cannot be de-anonymized by any algorithm. First, all of the previous works were based on the assumption that the graphs are generated by classical network models, e.g. classical Erdős-Rènyi network [16][19][13][9], correlated Erdős-Rènyi network model [18][24][4] and power law model [14], which may not represent real networks. Second, almost all of the previous studies [16][19][14][13][9][18][24][4] only focus on the

asymptotic regime, i.e. the regime where the probability of successfully matching all nodes approaches either 1 or 0, as the number of nodes approaches infinity, while there are no non-asymptotic results on de-anonymizability yet. In short, there is a need for a more systematical, quantitative and non-asymptotic analysis on de-anonymizability.

Therefore, in this paper, we provide the first study that systematically analyzes how graph structure characteristics will affect the de-anonymizability, without restricting our assumption to any specific network model. We will obtain quantitative, non-asymptotic results on de-anonymizability, which is defined as the maximum number of nodes that one can expect to correctly match in the given network for any de-anonymization algorithms.

In particular, we are interested in how the *symmetry* of a graph can affect the accuracy of de-anonymization. The idea of studying symmetry is intuitive, since attackers have no way to re-identify the symmetric nodes using only structural information. However, a thorough understanding on the relationship between symmetry and de-anonymizability remains elusive. In particular, *how should we measure and describe the degree of symmetry of an arbitrary graph? What is the exact quantitative relationship between symmetry and de-anonymizability?* In this paper, we aim to answer these two problems.

Specifically, in this paper we define the degree of symmetry of a graph by generating a matching probability matrix using the concept of graph bijective homomorphism. It then allows us to build the relationship between symmetry and de-anonymizability in general graphs. This result enables us to predict the maximum expected number of correctly-matched nodes for any algorithm, given any instance of a de-anonymization problem. Although the exact algorithm implementing such predictions incurs exponential complexity, we develop a practical algorithm to obtain an upper bound of such de-anonymization by finding graph automorphisms and counting the number of orbits of each graph, which overcomes the exponential time complexity of the original exact algorithm. Further, we conduct a case study on Erdős-Rènyi network model to apply such general results to a more specific situation. We also conduct experiments to verify our result.

Our main contributions are:

(1) We conduct the first theoretical study on de-anonymizability through the lens of symmetry. We precisely capture the structural similarity between two social networks by generating a matching probability transition matrix, using the concept of graph bijective homomorphisms.

(2) Based on these concepts, we proposed a method that quantitatively determines the de-anonymizability of given networks. Our method can find the maximum expectation number of correctly matched nodes, which is equal to the maximum diagonal sum of the matching probability matrix that we define. Then, by defining homomorphism transition matrix, we build the relationship between symmetry and de-anonymizability. Our method is systematic, general, and non-asymptotic. It can be applied to any general de-anonymization problem without depending on any specific network model.

(3) To overcome the exponential time complexity of finding all homomorphisms, we propose an method to obtain an upper bound of the de-anonymizability of any given de-anonymization problem. We prove that the number of orbits in each graph can serve as an upper bound of de-anonymizability, and we propose such algorithm accordingly which finds all the graph automorphisms and counts the number of orbits in each graph. This method is also general and can be exerted within practical time consumption.

(4) We apply the general method to the analysis of classical network model. As a case study, we analyzed the de-anonymizability of Erdős-Rènyi graph with any given parameters n and p. By enumerating the local symmetric structure in Erdős-Rènyi model, we obtain a numerical upper bound on de-anonymizability in Erdős-Rènyi graph. We also gave proof on the correctness by illustrating the fact that in the giant component of a supercritical Erdős-Rènyi graph, the number of symmetric nodes is of the order $o(1)$. All the results above are verified by extensive experiment results.

The remainder of the paper is organized as follows: In Section 2, we survey previous works on the topic of de-anonymization, de-anonymizability and symmetry. In Section 3, we introduce the model for de-anonymization and problem formulation for de-anonymizability, and also introduce some symmetry-related concepts. Section 4 demonstrates our main result, i.e., the method of obtaining the de-anonymizability on general graphs. In Section 5, we conduct a case study in the context of Erdős-Rènyi graph, in which we extend our general method to special prior network model conditions. Section 6 contains the experiment verification and result. We conclude with some discussion in Section 7.

2 RELATED WORK

2.1 De-anonymization Algorithms

Narayanan and Shimatikove [16] first proposed de-anonymization problem. They formulated this problem and proposed a generic algorithm based on network structure information with the help of seed nodes, i.e. pre-identified node pairs that are known to be correctly matched. However, in many situations, it is difficult to obtain such seed nodes due to the limited access to user profiles [9] [24]. Pedarsani and Grossglauer [19] first studied the seedless de-anonymization problem in the context of Erdős-Rènyi model, and they took the number of mismatched edge as the objective function. A different cost function based on Maximum a Posterior (MAP) was proposed in [17] and also used in [9] [24]. Recent works for correlated Erdős-Rènyi networks were reported in [20] [4]; Nitish and Silvio also proposed algorithm in [14] for the preferential attachment (PA) model.

2.2 De-anonymizability

Some networks are difficult to de-anonymize due to their inherent topoligical structure. Along with their problem formulation and algorithm, Pedarsani and Grossglauser [19] also approached the problem of finding theoretic conditions for successful de-anonymization. Cullina and Kiyavash [3] further investigated the conditions under which a pair of correlated Erdős-Rènyi graphs can be correctly matched. However, most of these studies focus on the asymptotic regime, i.e., when the probability of correctly matching all nodes goes to either 1 or 0, as the number of nodes approaches infinity. Further, they mostly base their studies on the assumption of classical network models such as Erdős-Rènyi [17][9] and preferential attachment[14].

The concept of de-anonymizability was also proposed by Ji et al. [10] [11] [12], which is a metric to describe the accuracy that a de-anonymization attack can achieve. Although the original intention of our proposing de-anonymizability is the same, the de-anoymizability in our paper has a different definition from theirs. The metric that we consider here is the performance of de-anonymization algorithm in **non-asymptotic** situation. We aim to provide a quantitative characterization of de-anonymizability by obtaining the maximum expected number of correctly mapped nodes of a de-anonymization problem.

2.3 Symmetry and De-anonymization

Symmetry is a widely discussed topic in mathematics, especially in abstract algebra. Many concepts, like isomorphism, automorphism, homomorphism, etc., are used to describe different types of symmetry in algebra structure. Related contents can be found in any textbook on abstract algebra, and are beyond the scope of this work. In the context of graph, Graph Isomorphism, Graph Automorphism and Graph Homomorphism are also classical topics[2][6], which show potential to describe the degree of symmetry of graphs.

[25][15] leveraged symmetry to anonymize the network. They proposed techniques to add symmetry to a network in order to protect personal information from structural attack. However, no previous work has applied symmetry to the theoretical analysis of de-anonymization problem. To the best of our knowledge, this paper is the first to build the relationship between symmetry and de-anonymizability in general graphs.

3 PRELIMINARY DEFINITION AND CONCEPT

3.1 De-anonymization Problem

Let $G = (V, E)$ be the underlying social network, where V is the set of nodes, and E is the set of edges. The underlying network indicates the true relationship among all users in V, but the true relationship is invisible to the attacker. We further define $G_1 = (V_1, E_1)$ as the published network and $G_2 = (V_2, E_2)$ as the auxiliary network. The published network is completely anonymous, meaning that no identity information is given of any node in G_1. In contrast, in auxiliary network, each node in G_2 may have a name label which is available to the attacker. Similar to most previous work [11][9], we suppose that both G_1 and G_2 have the same node set with G. Further, we denote the vertex in G as $V = \{1, 2, \ldots, n\}$ [3].

We suppose the G_1 and G_2 are generated via independent *edge-sampling* and random *vertex permutation* from G. By independent edge-sampling we mean that for graph $G_i (i = 1, 2)$, each existing edge in G is sampled to G_i i.i.d. with a sampling rate s_i. That is, for each edge $e \in E$, we have

$$P(e \in E_i) = \begin{cases} s_i & \text{if } e \in E \\ 0 & \text{if } e \notin E \end{cases} \tag{1}$$

In this sense, the underlying network G is the only bridge between G_1 and G_2, though it is invisible to the adversaries. We then randomly shuffle all the nodes in V_1 and V_2 independently and uniformly across all possible permutations.

Given the published network G_1 and the auxiliary network G_2, the problem of *social network de-anonymization* aims to match the node in G_1 to the nodes in G_2 using only the structural information of G_1 and G_2 as side information. Formally, we need to find a permutation $\sigma : V_1 \mapsto V_2$. For $v_1 \in V_1$ and $v_2 \in V_2$, $\sigma(v_1) = v_2$ means the node $v_1 \in V_1$ and the $v_2 \in V_2$ derive from the same node in the underlying network (and since we have the name label of v_2 in V_2, we can then deduce the name label of v_1 in V_1). The random shuffling of all nodes in V_1 and V_2, respectively, ensures that the attacker can only use structural information to find σ.

Unlike most of the previous work, in this paper we do not assume that G is generated by some specific network model. We only assume the sampling rates s_1, s_2 to be known. This assumption is reasonable since they can be obtained from statistical methods.

The parameters defined above can be simply denoted by a parameter set $\theta = (G_1, G_2, s_1, s_2)$, which we will use to state a de-anonymization problem. In the rest of the paper we may simply refer to a de-anonymization problem as *a de-anonymization problem with parameter* $\theta = (G_1, G_2, s_1, s_2)$ without ambiguity.

3.2 De-anonymizability

Predicated on Section 3.1, in the sequel we propose the concrete quantification of de-anonymization accuracy, denoted as **de-anonymizability** formally. Note that de-anonymizability will be the principal metric discussed in this paper, which measures the potential accuracy of a de-anonymization problem. To this end, we denote the true permutation between G_1 and G_2 as σ_0. Note that σ_0 **is a random variable due to the lack of ground truth information**. In other words, although there is a unique true mapping between G_1 and G_2, this true mapping is unknown to the attacker, who can only hypothetically choose among many similar seemingly true mappings. Counter-intuitive at the first sight, this claim can be illustrated by the following two examples:

1. Suppose $G = K_N$(complete graph) and $G_1 = G_2 = G$ (i.e. $s_1 = s_2 = 1$). Given G_1 and G_2, any permutation σ could be the true mapping assuming that adversaries have no information other than their topology. More concretely, σ_0 is a random variable satisfying $P(\sigma_0 = \sigma) = \frac{1}{N!}$ for any permutation σ from V_1 to V_2.

2. Suppose G_1, G_2 are given in Figure 1. Two possible underlying networks G and G' are shown, and we cannot tell which to be the true underlying network. As a result, different ways of matching from $G_i (i = 1, 2)$ to G (or G') exist, and it is uncertain which one is true. We can see that even G_1 itself is asymmetric[7], the process of sampling will still bring uncertainty to the de-anonymization.

To be more concrete, for any de-anonymization problem with parameter set $\theta = (G_1, G_2, s_1, s_2)$, let $\Pi = \{\pi\}$ denote all the permutations from V_1 to V_2. For each de-anonymization problem, there exists a probability distribution of the true mapping σ_0, denoted as $P(\sigma_0 = \pi|\theta)$ for all π in Π. One of the main focuses in the rest of this paper is to study ways to calculate such probability distribution.

For any two permutations π_1,π_2 from V_1 to V_2, we denote $N_{(\pi_1,\pi_2)}$ as the number of nodes in V_1, each of which, under π_1 and π_2, has the same image in V_2. Formally, we have [1]

$$N_{(\pi_1,\pi_2)} = \sum_{v \in V, \pi_1(v) = \pi_2(v)} 1 \tag{2}$$

[1]The notation $v \in V$ in the formula is equivalent to $v = 1, 2, \ldots, n$. At times, we interchange these two notations in this paper, especially in the subscript of summation notation.

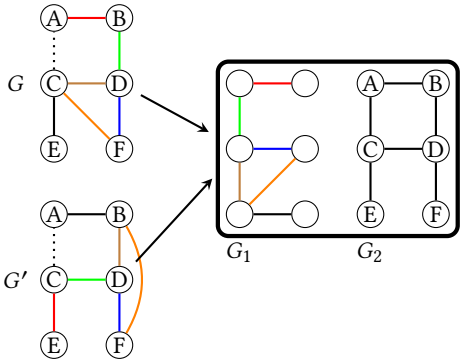


Figure 1: Uncertainty From Sampling. G_1 and G_2 are known, but multiple underlying networks (including G and G') is feasible, and can be the true underlying network.

Then, for any permutation (as a possible solution to the de-anonymization problem), the expectation of the number of correctly matched nodes σ, denoted as $E_{\sigma|\theta}$, can be calculated by

$$E_{\sigma|\theta} = \sum_{\pi\in\Pi} P(\sigma_0 = \pi|\theta)N_{(\sigma,\pi)}$$

Under the circumstance without ambiguity, we use E_σ to refer to the expectation. Intuitively, E_σ is the expectation of the number of correctly mapped nodes when σ is exerted to the anonymous network G_1.

Among all possible permutations σ, a best permutation σ^* for a de-anonymization problem is such a permutation that maximizes the expectation of the number of correctly-mapped nodes, which can be expressed as $\sigma^* = \arg\max_{\sigma\in\Pi} E_\sigma$. And the expectation E_{σ^*} is the maximum expectation of the number of successfully de-anonymized nodes. We define de-anonymizability of a de-anonymization problem as E_{σ^*}, which is a performance upper bound of any de-anonymization algorithm on this de-anonymization problem. To be concrete, we have the following definition:

Definition 3.1 (De-anonymizability). Given a de-anonymization problem with parameter $\theta = (G_1, G_2, s_1, s_2)$, the true mapping σ_0 is a random variable with probability distribution $P(\sigma_0 = \pi|\theta)$ for any π in Π, all the permutations from V_1 to V_2. (One of) the best permutation σ^* is

$$\sigma^* = \arg\max_\sigma E_\sigma = \arg\max_\sigma \sum_{\pi\in\Pi} P(\sigma_0 = \pi|\theta)N_{(\pi,\sigma_0)}$$

where σ is any permutation on V. The de-anonymizability of this problem is defined as E_{σ^*}. It reaches the maximum expectation of the number of correctly matched nodes over all permutations on V.

As mentioned, de-anonymizability will be our primary focus throughout the rest of the paper.

3.3 Symmetry

Intuitively, symmetry property determines de-anonymizability of the networks fundamentally. To better demonstrate this intuition, we can first study the *fully sampled case* where the sampling rate $s_1 = s_2 = 1$. In this case, $G_1 = G_2 = G$. As long as the attacker has known that the underlying graph is fully sampled, he can attack this network, i.e. mapping G_1 to G_2 by relabeling G_1 (we denote the graph after relabeling as G_1') such that each node pair, as long as they have the same labels in G_1' and G_2, keeps the existence or absence of edge between them. It is easy to see that this mapping process is equivalent to finding a *graph isomorphism*[2] from G_1 to G_2. However, multiple isomorphisms from G_1 to G_2 may exist, and the attacker cannot judge which one to be the ground truth. The best thing he could do is to 'make a guess'; in other words, any isomorphism has a possibility to be the true mapping from G_1 to G_2. Therefore, whether multiple isomorphisms exist from G_1 to G_2 determines the de-anonymizability of this fully sampled de-anonymization problem.

Since G_1 and G_2 are the same in structure in fully sample case, finding isomorphisms between G_1 to G_2 is then equivalent to finding *graph automorphisms*[2] of G (also G_1 or G_2). Interestingly, the number of automorphisms of a graph is indeed an indicator of symmetry [21]. Therefore we can come to the conclusion that symmetry can affect the de-anonymizability of a given problem.

To dive more deeply into its essence, the reason why existence of multiple automorphisms affects the de-anonymizability lies in that, to some of the node in G_1, it has a probability distribution to be mapped to more than one node in G_2 by the true mapping. For example, if there are two isomorphisms from G_1 to G_2, and the node v_i in G_1 is mapped to v_j and v_k in G_2 by these two isomorphisms respectively, since both isomorphisms have the possibility to be the true mapping σ_0, whether v_i is mapped to v_j or v_k is also not deterministic, and thus there is a probability distribution of the node in G_2 that v_i is mapped to. To this end, a good indicator of graph symmetry can be the probability distribution of the node image in G_2 of each node in G_1.

Similarly, we can generalize this concept of symmetry to any de-anonymization problem, as summarized in the following definition:

Definition 3.2 (Symmetry of a de-anonymization problem). We quantify the symmetry of a de-anonymization problem $\theta = (G_1, G_2, s_1, s_2)$ by the probability distribution that each node in G_1 will be mapped to the each node in G_2 by the true mapping σ_0. Concretely, the symmetry of a problem can be organized into a n-by-n matrix (where $n = |V_1| = |V_2|$) $M = M_{ij}$, where $M_{ij} = \sum_{\pi\in\Pi} P(\sigma_0 = \pi|\theta)\mathbb{1}\{\pi(i) = j\}$. We denote the matrix M as the *matching probability matrix*. Intuitively, the element M_{ij} is equal to the probability that node i in V_1 is mapped (by the true mapping) to node j in V_2.

We claim that there is direct link between de-anonymizability and this matching probability matrix. In order to prove this claim, the concept of doubly stochastic matrix needs to be introduced. A doubly stochastic matrix [22] is a square matrix with nonnegative real entries and the sum of the elements in each row and each column is equal to 1.

Proposition 3.3. *A matching probability matrix is a doubly stochastic matrix.*

Proof. Obviously each element in M is nonnegative. Also,

$$\sum_{j=1}^{n} M_{ij} = \sum_{j=1}^{n}\sum_{\pi\in\Pi} P(\sigma_0 = \pi|\theta)\mathbb{1}\{\pi(i) = j\}$$

$$= \sum_{\pi\in\Pi|\theta} P(\sigma_0 = \pi)\sum_{j=1}^{n}\mathbb{1}\{\pi(i) = j\} = \sum_{\pi\in\Pi|\theta} P(\sigma_0 = \pi) = 1$$

Therefore, the summation of the elements in each row is equal to 1. Similarly, the summation of the elements in each row sum is also equal to 1. The result follows. □

The diagonal sum of M corresponding to a permutation σ on $\{1, 2, \ldots, n\}$ of a doubly stochastic matrix M is defined as $\sum_{i=1}^{n} M_{i\sigma(i)}$. We now demonstrate that the expected number of correctly matched nodes of a permutation σ is equal to the diagonal sum of M corresponding to σ.

Proposition 3.4. *Given a de-anonymization problem with θ, with matching probability matrix M. For any permutation σ from V_1 to V_2, the expectation of the number of correctly matched nodes of a permutation σ is equal to the diagonal sum of M corresponding to σ.*

Proof. By definition of the expectation of the number of correctly matched nodes of a permutation σ, we have

$$\begin{aligned} E_{\sigma|\theta} &= \sum_{\pi\in\Pi} P(\sigma_0 = \pi|\theta) \sum_{v\in V, \sigma(v)=\pi(v)} 1 \\ &= \sum_{\pi\in\Pi} P(\sigma_0 = \pi|\theta) \sum_{i=1}^{n} \mathbb{1}\{\sigma(i) = \pi(i)\} \\ &= \sum_{i=1}^{n} \sum_{\pi\in\Pi} P(\sigma_0 = \pi|\theta)\mathbb{1}\{\pi(i) = \sigma(i)\} \\ &= \sum_{i=1}^{n} M_{i\sigma(i)} \end{aligned}$$

□

Corollary 3.5. *Obtaining the maximum expectation (i.e. obtaining the de-anonymizability) is equivalent to finding the maximum diagonal sum [22] of the matching probability matrix M.*

We denote $h(M)$ as the maximum diagonal sum of a doubly stochastic matrix M. Corollary 3.5 thus suggests that de-anonymization has close relationship with the symmetry of a given de-anonymization problem through the quantity $h(M)$.

3.4 The Concept of Graph Bijective Homomorphism

Earlier we have pointed out that Graph Automorphism, which determines the symmetry of a single graph, affects de-anonymizability. In addition, in this paper we also use the concept of *Graph Bijective Homomorphism*[6] to capture the symmetry between any pair of given graphs. The Graph Bijective Homomorphism is defined as follows:

Definition 3.6 (Graph Bijective Homomorphism). For two graph $G_1 = (V_2, E_2)$ and $G_2 = (V_2, E_2)$, if there is a bijection $h : V_1 \to V_2$ such that $\forall(v_i, v_j) \in E_1 \to (f(v_i), f(v_j)) \in E_2$, then we say that h is a *bijective homomorphism* from G_1 to G_2. In this paper, we slightly abuse the term *homomorphism* to refer to graph bijective homomorphism.

4 DE-ANONYMIZABILITY IN GENERAL GRAPHS

In Section 3 we have already built the link between symmetry and de-anonymizability. Precisely, we can calculate de-anonymizability directly after obtaining the matching probability matrix. Therefore, in this section we aim to obtain the probability transition matrix of a given problem θ.

4.1 Probability Distribution of Underlying Network

Since the underlying graph G is the mere link between G_1 and G_2, we in this section derive the probability distribution of G from a Bayesian's perspective. As a necessity of Bayes' Rule, we assume a prior probability of G, denoted as $P(G)$. Notice that, this prior probability is a generalization of previous model-based assumption of de-anonymization problem. We then derive the posterior probability distribution of the underlying network G when $\theta = (G_1, G_2, s_1, s_2)$ is given.

Proposition 4.1. *Given parameter $\theta = (G_1, G_2, s_1, s_2)$ of a de-anonymization problem, for any graph $G = (V, E)$, a prior probability of G is given, denoted as $P(G)$. The probability of its being the ground truth underlying network is propositional to $P(G)\hom(G_1, G)\hom(G_2, G)((1-s_1)(1-s_2))^{|E|}$, where $\hom(F, G)$ is the number of graph bijective homomorphisms from F to G. More precisely,*

$$P(G|\theta) = \frac{1}{H} P(G)\hom(G_1, G)\hom(G_2, G)((1-s_1)(1-s_2))^{|E|}$$

where $H = \sum_{G\in\mathcal{G}} P(G)\hom(G_1, G)\hom(G_2, G)((1-s_1)(1-s_2))^{|E|}$ is a normalization parameter.

Proof. Determining which graph to be the underlying network is like an inferencing process, so it is reasonable to use Bayes' Rule to deal with the probability. By Bayes' Rule we can write the probability of G given G_1,G_2 as:

$$P(G|G_1, G_2) = \frac{P(G_1, G_2|G)P(G)}{P(G_1, G_2)}$$

On the right side, $P(G_1, G_2)$ is a normalized factor and is identical among different G. Thus we have

$$\begin{aligned} P(G|G_1, G_2) \propto& P(G)P(G_1, G_2|G) \\ \overset{(0)}{=}& P(G)P(G_1|G)(G_2|G) \\ \overset{(1)}{=}& P(G)\hom(G_1, G)s_1^{|E_1|}(1-s_1)^{|E|-|E_1|} \\ & \hom(G_2, G)s_2^{|E_2|}(1-s_2)^{|E|-|E_2|} \\ \propto& P(G)\hom(G_1, G)\hom(G_2, G)((1-s_1)(1-s_2))^{|E|} \end{aligned}$$

In formula, (0) holds because G_1 and G_2 are independent samplings from G. (1) holds because there are $\hom(G_i, G)$ ways to produce G_i from G by sampling and vertex permutation. □

The probability distribution of the true mapping σ_0 can be expressed by a total probability formula with known probability of G, which is

$$P(\sigma_0 = \pi|\theta) = \sum_{G\in\mathcal{G}} P(G|\theta)P(\sigma_0 = \pi|\theta, G)$$

Therefore in the following section we focus on $P(\sigma_0 = \pi|\theta, G)$, the probability distribution of the true mapping when both the problem and the G is given.

4.2 Probability Distribution of True Mapping with Known Underlying Network

Due to the existence of the underlying network G, the permutation from G_1 to G_2 is not enough for our analysis. Therefore, we analyze the problem via matching G_1 and G_2, respectively, to G. The final mapping from G_1 to G_2 is a **composition** of these two mappings. Since $G_i (i = 1, 2)$ is sampled from G, a feasible mapping (from G_i to G) only needs to keep the edge existence, but not the non-existence of the edge. In other words, the feasible mapping from $G_i (i = 1, 2)$ to G should be a *graph bijective homomorphism* from $G_i (i = 1, 2)$ to G.

Recall that when we construct G_1 and G_2, we randomly permute all vertices. Therefore, there exists a true mapping from G_1 and G_2, respectively, to G. We denote the true mapping from G_1 to G as f_0, and the mapping from G_2 to G as h_0. Then we have $\sigma_0 = f_0 \circ h_0^{-1}$.[2] For the same reason with σ_0, f_0 and h_0 are all random variables. Also, We define $F_G = \{f_1, f_2, \ldots, f_{k1}\}$ as all the homomorphisms from G_1 to G, and $H_G = \{h_1, h_2, \ldots, h_{k2}\}$ the homomorphisms from G_2 to G. Here $k1 = \hom(G_1, G), k2 = \hom(G_2, G)$ represent the number of homomorphisms from G_1 and G_2, respectively, to G. Notice that here F_G, H_G, $k1$, $k2$ are variant among different topological realizations of G. Proposition 4.2 claims that each homomorphism has the same probability to be the true permutation from G_1 and G_2, respectively, to G .

Proposition 4.2. *Given underlying network G and parameter θ of a de-anonymization problem, each homomorphism mapping f_i from G_1 to G has the probability of $\frac{1}{\hom(G_1,G)}$ to be the true mapping from G_1 to G. Similarly, each h_i has the probability of $\frac{1}{\hom(G_2,G)}$ to be the true mapping from G_2 to G.*

Proof. We only prove the proposition for G_1. For G_2 the proof is the same.

Given G_1 and G, if a permutation f_i is proved to be the true mapping f_0, then: (1) f_i is a (graph bijective) homomorphism from G_1 to G (otherwise G_1 cannot be sampled from G) (2) In the sampling process, for the edges in G, all the edges that exist in G_1 are 'sampled in', while all other edges that are absent from G_1 are 'sampled out'.

By Bayes' Law we can write that for any homomorphism f_i from G_1 to G, the probability that f is the true mapping f_0 is:

$$P(f_0 = f_i | G_1, G) = \frac{P(G_1 | f_0 = f_i, G) P(f_0 = f_i | G)}{P(G_1 | G)}$$

here, (a) $P(f_0 = f_i | G)$ is a constant since we have no preference for any specific mapping; (b) $P(G_1|G)$ is a normalized factor and is identical among different f_i; (c) $P(G_1 | f_0 = f_i, G) = (1 - s_1)^{|E| - |E_1|} s_1^{|E_1|}$ is constant since the edge number of G and both G_i are determined. Therefore, each homomorphism f_i has the same probability $\frac{1}{\hom(G_1,G)}$ to be the true mapping when G is given. Similar for G_2. □

4.3 Main Result

The previous two sections provide all the evidence that we need to obtain the de-anonymizability. In this section we combine previous results to obtain the de-anonymizability of a given de-anonymization problem.

We define the homomorphism transition matrix from G_1 to G as follows:

Definition 4.3 (Homomorphism Transition Matrix). For a de-anonymization problem θ and a given underlying network G, the homomorphism transition matrix from G_1 to G is

$$C_G = \frac{1}{hom(G_1, G)} \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1n} \\ c_{21} & c_{22} & \ldots & c_{2n} \\ \ldots & \ldots & & \\ c_{n1} & c_{n2} & \ldots & c_{nn} \end{bmatrix}$$

where c_{ij} is the number of homomorphisms from G_1 to G that matches the node i in G_1 to the node j in G. Formally, $(C_G)_{ij} = c_{ij} = \sum_{f \in F_G} \mathbb{1}(f(i) = j)$. Similarly, for G_2, the homomorphism transition matrix from G_2 to G is

$$D_G = \frac{1}{hom(G_2, G)} \begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1n} \\ d_{21} & d_{22} & \ldots & d_{2n} \\ \ldots & \ldots & & \\ d_{n1} & d_{n2} & \ldots & d_{nn} \end{bmatrix}$$

where $(D_G)_{ij} = d_{ij} = \sum_{h \in H_G} \mathbb{1}(h(i) = j)$.

Intuitively, the element C_{ij} (resp. D_{ij}) in homomorphism transition matrix indicates the probability that node i in G_1 (resp. G_2) is mapped to node j in G by the true mapping.

Now we can calculate the probability distribution of the true mapping σ_0 in terms of C_G and D_G along with the probability of underlying graph $P(G|\theta)$.

Theorem 4.4. *For a de-anonymization problem θ with a given underlying network G, $M = \sum_{G \in \mathcal{G}} P(G|\theta) C_G D_G^T$*

Proof. For each possible underlying network G, we define $M_G = C_G D_G^T$. For each element in M_G, we have

$$\begin{aligned}
(M_G)_{ij} =& (C_G D_G^T)_{ij} = \sum_{k=1}^{n} (C_G)_{ik} (D_G)_{jk} \\
\overset{(0)}{=}& \sum_{k=1}^{n} P(f_0(i) = k | \theta, G) P(h_0(j) = k | \theta, G) \\
\overset{(1)}{=}& \sum_{k=1}^{n} P(f_0(i) = k, h_0(j) = k | \theta, G) \\
\overset{(2)}{=}& P(\sigma_0(i) = j | \theta, G) \\
\overset{(3)}{=}& \mathbb{E}(P(\sigma_0 = \pi | \theta, G) \mathbb{1}\{\pi(i) = j\}) \\
\overset{(4)}{=}& \sum_{\pi \in \Pi} P(\sigma_0 = \pi | \theta, G) \mathbb{1}\{\pi(i) = j\}
\end{aligned}$$

In this formula, (0) holds due to the probability distribution we obtained in Proposition 4.2, (1) holds due to the fact that G_1 and G_2 are independent samplings from G, (2) holds due to the fact that σ_0 is the composition of f_0 and h_0^{-1}, (3) holds due to the fact that the expectation of an indicator function is equal to its probability, (4) holds due to the definition of expectation.

Therefore, each element in M_G, $(M_G)_{ij}$ is equal to the probability that node i in G_1 to be matched to node j in G_2 when G is given.

[2] For a permutation f, the inverse of f, denoted as f^{-1}, is a permutation that satisfies : for each v, $f^{-1}(f(v)) = v$

Applying a total formula, each element in M_{ij} is the probability that node i in G_1 to be matched to node j in G_2. □

4.4 An Upper Bound of De-anonymizability

So far, we have already proposed our method to determine the matching probability matrix of a de-anonymization problem. However, this method is costly in terms of time complexity due to two reasons: (1) the method involves enumerating common supergraphs G, which is exponentially expensive; (2) finding graph bijective homomorphisms is proved to be NP-Complete in general case[8]. The prohibitive cost drives the necessity of proposing more efficient approximate algorithms. To this end, in this section we want to bound the de-anonymization using the structural information of only either G_1 or G_2.

Definition 4.5 (Orbit). For a graph $G = (V, E)$, two nodes v_1, v_2 are symmetric (automorphically equivalent) [23] (denoted as $v_1 \sim v_2$) if there exists an automorphism f of G such that $f(v_1) = v_2$. An orbit is a subset of nodes. The orbit that a certain node v belongs to contains all the nodes that are symmetric to v. Precisely, an orbit $\mathcal{O} = \{v_1, v_2, \ldots, v_i\}$ satisfies: if $v \in \mathcal{O}$,then for any $v' \sim v$, $v' \in \mathcal{O}$,

According to Definition 4.5, intuitively, an orbit is a subset of nodes, in which all nodes are internally symmetric.

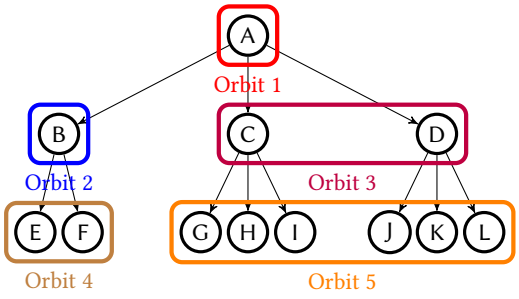


Figure 2: An Illustration of Orbit

Lemmas 4.6 present some further properties regarding the orbit.

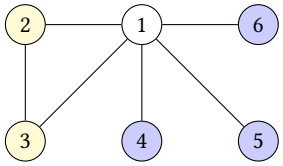
Lemma 4.6. *[5] A graph G can be partitioned into several orbits. That is, there exists a partition $\{\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_k\}$ of V such that for each $i = 1, 2, \ldots k$, $\mathcal{O}_i$ is an orbit in V.*

We then define $|Orb(G)| = k$ as the number of orbits contained in G (i.e. the number of partitions in Lemma 4.6). We now use the concept of orbit to define the *automorphism transition matrix*, which captures the symmetry of a single graph.

Definition 4.7 (Automorphism transition matrix). For a graph G, the *automorphism transition matrix* A(G) is defined as

$$A_{ij} = \begin{cases} \frac{1}{|Orb(i)|} & j \in Orb(i) \\ 0 & otherwise \end{cases}$$

Note that A is a symmetric matrix, since for any two nodes $i, j \in V$, $i \in Orb(j)$ is equivalent to $j \in Orb(i)$, which implies $|Orb(i)| = |Orb(j)|$. Particularly, we denote $A_1 = A(G_1)$, $A_2 = A(G_2)$ as the automorphism transition matrix of G_1 and G_2, respectively. The following Figure 3 is an illustration of automorphism matrix.



(a) A graph with three orbits, each color represents an orbit

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

(b) The automorphism transition matrix of the graph on the left

Figure 3: An illustration of automorphism transition matrix

Automorphism transition matrix is used to capture the symmetry within a graph. The following proposition shows that automorphism transition matrix can be used to obtain an upper bound of the de-anonymizability.

Proposition 4.8. *Given problem θ and an presumed underlying network G, let A_1,A_2 be their automorphism transition matrices of G_1,G_2, respectively. Let C_G, D_G be the homomorphism transition matrix from G_1 and G_2, respectively, to G. Then each homomorphism transition matrix keeps invariant under the multiplication of corresponding automorphism transition matrix. Precisely, $C_G = A_1 C_G, D_G = A_2 D_G$*

Proof. We only prove the result of G_1, i.e. $C_G = A_1 C_G$. The proof for G_2 is completely the same. Let $C' = A_1 C_G$. Then

$$C'_{ij} = \sum_{k \in V} (A_1)_{ik} (C_G)_{kj} = \sum_{k \in V}^{k \in Orb_{G_1}(i)} \frac{1}{|Orb_{G_1}(i)|} C_{kj}$$

Here $Orb_{G_1}(i)$ represents the orbit in G_1 that contains i. We can see from the formula that the theorem holds if $(C_G)_{kj} = (C_G)_{ij}$ for any $k \in Orb_{G_1}(i)$. In fact, the latter can be proved as follows:

Suppose i and k are in the same orbit (of G_1). That indicates that there exists an automorphism f (on G_1) that maps i to k ($f(i) = k$).

Then for each homomorphism σ from i (in G_1) to j (in G), there exists a permutation $\sigma' = f \circ \sigma$. On one hand, σ' is a homomorphism, since G_1 keeps invariant under the action of f. On the other hand, σ' maps k to j. This suggests that for each homomorphism that maps i(in G_1) to j(in G), there also exists a homomorphism mapping k(in G_1) to j(in G), and vice versa. Therefore, the number of homomorphisms that map i to j and map k to j is equal.

Recall that $(C_G)_{ij} = \frac{1}{k_1} c_{ij}$, which is determined by the number of homomorphisms that match i (in G_1) to j (in G). Thus $(C_G)_{ij} = (C_G)_{kj}$ for any i,k in the same orbit. Therefore, $C_G = A_1 C_G$. Similarly $D_G = A_2 D_G$. □

Corollary 4.9. *The matching probability matrix M can be represented as: $M = A_1 M A_2$.*

Proof. M is a linear combination of M_G, and for each M_G, $M_G = C_G D_G^T = A_1 C_G D_G^T A_2$. Since A_1 and A_2 are constant matrices within a de-anonymization problem, the result follows. □

Notice that A_1, M, A_2 are all doubly stochastic matrices[3]. Then Theorem 4.10 shows that the maximum diagonal sum of the product

[3]We have proved previously that M is a doubly stochastic matrix. For A_1 and A_2, the result can be proved easily by the definition of automorphism transition matrix

of two doubly stochastic matrices is no greater than that of any one of them.

Theorem 4.10 (Theorem 4.1 in [22]). *For two n-×-n doubly stochastic matrices A and B, $h(AB) \leq \min(h(A), h(B))$.*

Corollary 4.11. *For any G, $h(M) = h(A_1 M A_2) \leq h(A_1) = |Orb(G_1)|$. Similarly $h(M) \leq |Orb(G_2)|$.*

Corollary 4.11 indicates that automorphisms in G_1 or G_2 can determine the upper bound of the de-anonymizability of the de-anonymization problem. Given G_1, G_2, the de-anonymizability of the problem can not exceed the orbit numbers of G_1 and G_2. In other words, as long as either G_1 G_2 are highly symmetric, we cannot expect too many nodes to be correctly de-anonymized.

5 A CASE STUDY: ERDŐS-RÈNYI MODEL AS THE UNDERLYING NETWORK

Next, we conduct a case study of calculating de-anonymizability in the context of Erdős-Rènyi network. Since most previous works make this assumption, this part can be seen as an application of our main theorem.

An Erdős-Rènyi network is characterized by two parameters n and p, where n represents the number of nodes, and p represents the probability that any node pair has an edge in between, independently of other node pairs. The de-anonymizability here refers to **the expectation of de-anonymizability of all graph instances** generated by the Erdős-Rènyi model. Since our method focuses on the automorphism and homomorphism of a graph, the key is to theoretically analyze the automorphism and homomorphism of an instance generated by Erdős-Rènyi model. We will study the **fully sampled case** in detail, and then briefly demonstrate how we deal with the partially sampled case based on the result from the fully sample case.

As [15] has mentioned, the symmetric structure in real-world network is more likely to be 'local'. Inspired by this phenomenon, we enumerate some locally symmetric structures in Erdős-Rènyi graph, calculate the expected number of times of the appearance of each of them, and count the number of nodes that are mutually symmetric (so that they are on the same orbit).

5.1 Fully Sampled Case

In fully sampled cases we only need to count the orbits of the graph to get the de-anonymizability. In doing so, we introduce the concept of motif to express the locally symmetric structure.

Definition 5.1. For a graph $G = (V, E)$, a motif is denoted by $V_s \subset V$, and we define $T(V_s) = |V_s| - |Orb(V_s)|$, where $|Orb(V_s)|$ is the number of orbits in the subgraph of G that contains all the nodes in V_s. Notice that $T(V_s)$ captures how many nodes in the original graph G are collapsed into orbits formed by the motif V_s. Thus, below we often refer to $T(V_s)$ as the contraction of graph G by the motif V_s.

Here we focus on two most common kinds of motif (illustrated in Figure 4) that have been verified to exist in many complex networks [15]:

(1) Fruits in cherry-like structure: A set of k nodes forms a 'k-fruit cherry-like' motif (in short 'k-fruit') *iff*: (1) the degree of each of them is one; (2) they are connected to the same node. Proposition 5.2 shows the effect of all cherry-like structures S in terms of $T(V_s)$, $V_s \in S$.
(2) Small isolated components: A small isolated component in a graph is simply a connected component whose size is less than a threshold k. To reduce the number of components that we have to enumerate, we use the fact that almost all the small components in Erdős-Rènyi graph are trees [1]. Here we choose the threshold $k = 7$, of which the detailed illustration of all these 13 types of tree components are available in Figure 5. Proposition 5.3 shows the effect of small components C.

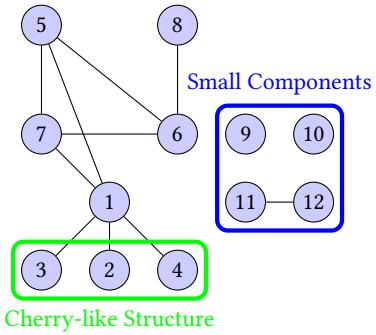


Figure 4: Two kinds of motifs we consider here.

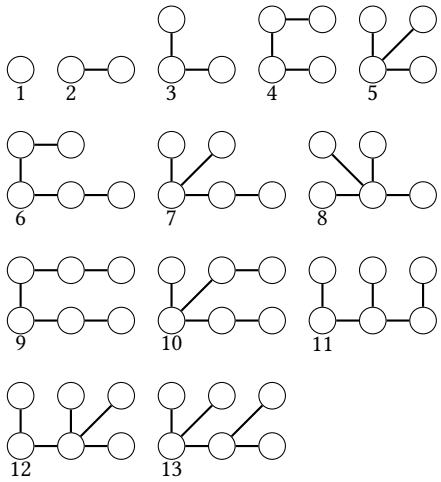


Figure 5: Different types of small components.

Proposition 5.2. *The expected total contraction by cherry-like structures is:*

$$T(S) = \sum_{V_s \in S} T(V_s) = \sum_{k=2}^{n} (-1)^k \binom{n}{k} (n-k) p^k (1-p)^{k(n-k-1)} * (1-p)^{\binom{k}{2}}$$

Proof. In an Erdős-Rènyi graph $G(n, p)$, for k nodes, the probability that all of them are connected only to the same node is

$$P_{V_{s_k}} = (n-k) p^k (1-p)^{k(n-k-1)} * (1-p)^{\binom{k}{2}}$$

The expected number of 'k-fruits' over the whole graph is

$$E|S_k| = \binom{n}{k} P_{s_k} = \binom{n}{k}(n-k)p^k(1-p)^{k(n-k-1)} * (1-p)^{\binom{k}{2}}$$

It is easy to see that the contraction by a k-fruit' structure is $k-1$, since all the 'fruits' in this structure are on the same orbit. However, note that a 'k-fruit' structure also contains $\binom{k}{i}$ 'i-fruit' substructure for $1 \le i \le k$. Due to this repetitive counting, we can not simply add up $(k-1)E|S_k|$ for all k to be the expected total contraction $T(S)$. Instead, we denote A_k as the number of '*pure* k-fruit' structures. Here a '*pure* k-fruit' structure means that it is not a substructure of a 'K-fruit' structure where $k < K$. Let $a_k = E|S_k|$. Then, we have

$$A_k = a_k - \sum_{i=k+1}^{n} \binom{i}{k} A_i \tag{3}$$

when $k = 2, 3, \ldots, n-1$. Apparently $A_n = a_n$. Then we have $T(S) = \sum_{k=2}^{n}(k-1)A_k$.

Next we prove the following equality:

$$\sum_{k=2}^{n}(k-1)A_k = \sum_{k=2}^{n}(-1)^k a_k. \tag{4}$$

To see this, note that by Equation (3), we have

$$a_k = A_k + \sum_{i=k+1}^{n} \binom{i}{k} A_i = \sum_{i=k}^{n} \binom{i}{k} A_i.$$

Therefore,

$$\sum_{k=2}^{n}(-1)^k a_k = \sum_{k=2}^{n}(-1)^k \left[\sum_{i=k}^{n} \binom{i}{k} A_i\right]$$

$$= \sum_{i=2}^{n}\left(\sum_{k=2}^{i}(-1)^k \binom{i}{k} A_i\right) \quad \text{(by switching the order of summation)}$$

$$= \sum_{i=2}^{n}((1-1)^i - 1 + i)A_i \quad \text{(by binomial expansion)}$$

$$= \sum_{i=2}^{n}(i-1)A_i = \sum_{k=2}^{n}(k-1)A_k.$$

Now that Equation (4) is proven, we then have $T(S) = \sum_{k=2}^{n}(-1)^k a_k$. Substituting a_k back with $E|S_k|$ we can get the result. □

Next we count the contraction due to small isolated components. Note that some of the small isolated components may contain a 'k-fruit' structure, whose contraction has been counted in Prop. 5.2. Thus, below we focus on the additional contraction contributed by the small components, excluding the contraction by any sub 'cherry-like' structure.

Proposition 5.3. *The additional contraction by small isolated components are $T(C) = \sum_i T(c_i)$, where $T(c_i)$ for each type of components are listed in Table 1.*

Proof. For each type, the expected number of its appearance is calculated in order to obtain the additional contraction.

As an example, we show in detail how to calculate the expectation of C_1 (i.e. an isolated node). The probability that a node is connected to none of the other nodes is $(1-p)^{n-1}$.

Table 1: T(c_i) for each type of motifs

i	$E(c_i)$	$T(c_i)$
1	$n(1-p)^{n-1}$	$E(c_1) - 1$
2	$\binom{n}{2}p(1-p)^{2(n-2)}$	$2 * E(c_2) - 1$
3	$\binom{n}{3}\frac{3!}{2}p^2(1-p)^{3(n-3)+1}$	$2 * E(c_3) - 1$
4	$\binom{n}{4}\frac{4!}{2}p^3(1-p)^{4(n-4)+3}$	$2 * E(c_4) - 2$
5	$\binom{n}{4}\frac{4!}{3}p^3(1-p)^{4(n-4)+3}$	$4 * E(c_5) - 2$
6	$\binom{n}{5}\frac{5!}{2}p^4(1-p)^{5(n-5)+6}$	$5 * E(c_6) - 3$
7	$\binom{n}{5}\frac{5!}{2}p^4(1-p)^{5(n-5)+6}$	$4 * E(c_7) - 4$
8	$\binom{n}{5}\frac{5!}{4!}p^4(1-p)^{5(n-5)+6}$	$2 * E(c_8) - 1$
9	$\binom{n}{6}\frac{6!}{2}p^5(1-p)^{6(n-6)+10}$	$6 * E(c_9) - 2$
10	$\binom{n}{6}\frac{6!}{2}p^5(1-p)^{6(n-6)+10}$	$6 * E(c_9) - 4$
11	$\binom{n}{6}\frac{6!}{2}p^5(1-p)^{6(n-6)+10}$	$6 * E(c_10) - 4$
12	$\binom{n}{6}\frac{6!}{3!}p^5(1-p)^{6(n-6)+10}$	$4 * E(c_11) - 4$
13	$\binom{n}{6}\frac{6!}{8}p^5(1-p)^{6(n-6)+10}$	$4 * E(c_12) - 2$

Then the expectation over the whole graph is equal to

$$E(c_1) = n(1-p)^{n-1}$$

which shows the expected number of C_1 motifs (i.e. isolated nodes) in G (in expectation). For type 3,5,7,8,12,13, cherry-like structure exists in those components. We now take type 3 as an example to show how to calculate the additional contraction by those motifs. We can show that expected number of the appearance of type 3 components is:

$$E(c_3) = \binom{n}{3}\frac{3!}{Aut(c_3)}p^2(1-p)^{3(n-3)+1},$$

where $Aut(c_i)$ is the number of automorphisms of the c_i motif, which is 2 in the case of c_3. Note that all type-3 components together form 2 orbits. Thus, the total contraction is $3E(c_3) - 2$. However, in each type-3 component, the two 'fruit' nodes have already contributed a contraction of 1 in the analysis in Proposition 5.2. Therefore, the additional contration is $2 * E(c_3) - 2$. For other types of components, the analysis is similar. The results are listed in the Table 1. □

Considering the overall effect of both types of motifs, Proposition 5.4 characterizes the de-anonymizability of a Erdős-Rènyi model.

Proposition 5.4. *An upper bound of the de-anonymizability of a graph generated from Erdős-Rènyi model $G(n,p)$ is $N - T(S) - T(C)$.*

For ease of understanding, let us now take an Erdős-Rènyi graph with $n = 1000, p = 1/500$ as an example. After calculation we get

$$T(S) = 33.69, T(C) = 181.49$$

$$E_{\sigma^*} = n - T(S) - T(C) = 784.82$$

which means in an Erdős-Rènyi graph generated by $G(1000, 1/500)$, at most 3/4 nodes can be (expected to be) de-anonymized.

5.2 Partially Sampled Case

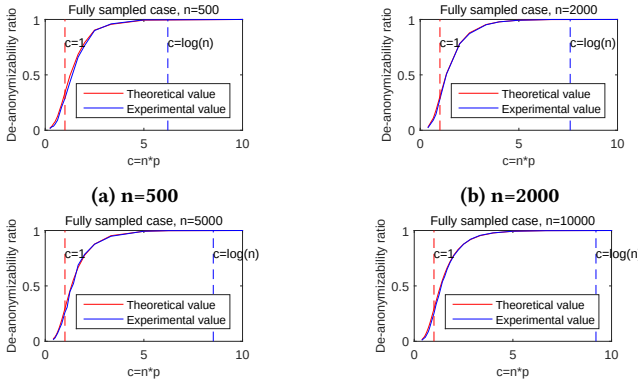
In partially sampled case, for the problem $\theta = (G_1, G_2, s_1, s_2)$ where the underlying network is generated by Erdős-Rènyi model $G(n,p)$, we calculate the de-anonymizability of $G(n, ps_1)$ and $G(n, ps_2)$ respectively, using the method proposed in Section 4.4. Then we take

the smaller one as the upper bound of the de-anonymizability of the problem.

6 EXPERIMENT EVALUATION

To verify our result in the case study, we conduct experiments on Erdős-Rènyi graph to testify our theoretical results.

We choose $N = 500, 2000, 5000, 10000$ as the representative of small-size network and large-size network, respectively. Using the method in Section 5.1 , we calculate the expected de-anonymizability of the Erdős-Rènyi graph with different parameter. To compare the theoretical result with the experimental one, we generate a number of graphs generated by the given model, and count the orbit number as the experimental result. In this experiment, we use *nauty* [15], an efficient automorphism-related toolkit, to obtain the orbit number of a graph. For each model $G(n, p)$, we generate 10 independent samples of Erdős-Rènyi graph and take the average of their orbit numbers. Figure 6 shows the results of our experiments. In this plot, x-axis is the (asymptotic) average node degree $c = np$, and y-axis is the ratio of de-anonymizability to the number of nodes n. The result demonstrates the high consistency of our theoretical result with the experimental result. Also, the result accords with a well-known classic conclusion [1] that the Erdős-Rènyi network tends to be asymmetric when $c = np$ exceeds the threshold $\log(n)$.



(a) n=500 **(b) n=2000**

(c) n=5000 **(d) n=10000**

Figure 6: Experiments on fully sampled Erdős-Rènyi graph

7 CONCLUSIONS

The past decades witnessed the advancement of the study on de-anonymization problem problem. Many algorithms are proposed, but systematic analysis on the accuracy of de-anonymization remains limited. We proposed a quantitative method to determine the de-anonymizability of a non-asymptotic de-anonymization problems through the lens of symmetry. To the best of our knowledge, our work is the first to study the exact relationship between de-anonymizability and symmetry.

We believe that such a study of symmetry will be of great value to not just social network de-anonymizability, but also to other types of *Graph Matching* problem, such as field pattern recognition, chemistry molecular reconstruction, etc. Therefore, our analysis has the potential to be applied to a large number of real-world applications.

ACKNOWLEDGEMENT

This work was supported by NSF China under Grant (No. 61822206, 61960206002, 61832013, 62041205, 61532012), Medical Engineering Cross Fund of Shanghai Jiaotong University (YG2020YQ16), and by an NSF sub-award via Duke University (NSF IIS-1932630).

REFERENCES

[1] Béla Bollobás. 1998. *Random Graphs.* Springer New York, New York, NY, 215–252. https://doi.org/10.1007/978-1-4612-0619-4_7
[2] Peter J Cameron et al. 2004. Automorphisms of graphs. *Topics in algebraic graph theory* 102 (2004), 137–155.
[3] Daniel Cullina and Negar Kiyavash. 2016. Improved Achievability and Converse Bounds for Erdos-Renyi Graph Matching. *Acm Sigmetrics Performance Evaluation Review* 44, 1 (2016), 63–72.
[4] Osman Emre Dai, Daniel Cullina, Negar Kiyavash, and Matthias Grossglauser. 2019. Analysis of a Canonical Labeling Algorithm for the Alignment of Correlated Erdos-Rényi Graphs. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 3, 2 (2019), 36.
[5] John D. Dixon and Brian Mortimer. 1996. *Permutation Groups.*
[6] Martin Dyer and Catherine Greenhill. 2000. The complexity of counting graph homomorphisms. *Random Structures & Algorithms* 17, 3-4 (2000), 260–289.
[7] Paul Erdős and Alfréd Rényi. 1963. Asymmetric graphs. *Acta Mathematica Hungarica* 14, 3-4 (1963), 295–315.
[8] Jiří Fiala and Jan Kratochvíl. 2008. Locally constrained graph homomorphisms—structure, complexity, and applications. *Computer Science Review* 2, 2 (2008), 97 – 111. https://doi.org/10.1016/j.cosrev.2008.06.001
[9] Luoyi Fu, Xinzhe Fu, Zhongzhao Hu, Zhiying Xu, and Xinbing Wang. 2017. De-anonymization of Social Networks with Communities: When Quantifications Meet Algorithms. *arXiv preprint arXiv:1703.09028* (2017).
[10] S. Ji, W. Li, N. Z. Gong, P. Mittal, and R. Beyah. 2016. Seed-Based De-Anonymizability Quantification of Social Networks. *IEEE Transactions on Information Forensics and Security* 11, 7 (July 2016), 1398–1411. https://doi.org/10.1109/TIFS.2016.2529591
[11] S. Ji, W. Li, S. Yang, P. Mittal, and R. Beyah. 2016. On the relative de-anonymizability of graph data: Quantification and evaluation. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications.* 1–9. https://doi.org/10.1109/INFOCOM.2016.7524585
[12] Shouling Ji, Prateek Mittal, and Raheem Beyah. 2016. Graph Data Anonymization, De-Anonymization Attacks, and De-Anonymizability Quantification: A Survey. *IEEE Communications Surveys & Tutorials* PP (12 2016), 1–1. https://doi.org/10.1109/COMST.2016.2633620
[13] S. Ji, P. Mittal, and R. Beyah. 2017. Graph Data Anonymization, De-Anonymization Attacks, and De-Anonymizability Quantification: A Survey. *IEEE Communications Surveys Tutorials* 19, 2 (Secondquarter 2017), 1305–1326. https://doi.org/10.1109/COMST.2016.2633620
[14] Nitish Korula and Silvio Lattanzi. 2014. An efficient reconciliation algorithm for social networks. *Proceedings of the VLDB Endowment* 7, 5 (2014), 377–388.
[15] Ben D. Macarthur, Rubén J. Sánchez-García, and James W. Anderson. 2008. Symmetry in complex networks. *Discrete Applied Mathematics* 156, 18 (2008), 3525–3531.
[16] Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing Social Networks. In *IEEE Symposium on Security & Privacy.*
[17] Efe Onaran, Siddharth Garg, and Elza Erkip. 2016. Optimal de-anonymization in random graphs with community structure. In *Sarnoff Symposium, 2016 IEEE 37th.* IEEE, 1–2.
[18] Efe Onaran, Siddharth Garg, and Elza Erkip. 2017. Optimal De-Anonymization in Random Graphs with Community Structure. In *IEEE Sarnoff Symposium.*
[19] Pedram Pedarsani and Matthias Grossglauser. 2011. On the privacy of anonymized networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 1235–1243.
[20] Yingxia Shao, Jialin Liu, Shuyang Shi, Yuemei Zhang, and Bin Cui. 2019. Fast De-anonymization of Social Networks with Structural Information. *Data Science and Engineering* 4, 1 (01 Mar 2019), 76–92. https://doi.org/10.1007/s41019-019-0086-8
[21] Nenad Trinajstic. 2018. *Chemical graph theory.* Vol. Vol. 1. Routledge.
[22] Tzu Hsia Wang. 1974. Maximum and minimum diagonal sums of doubly stochastic matrices. *Linear Algebra & Its Applications* 8, 6 (1974), 483–505.
[23] Wentao Wu, Yanghua Xiao, Wei Wang, Zhenying He, and Zhihui Wang. 2010. K-symmetry model for identity anonymization in social networks. In *Proceedings of the 13th international conference on extending database technology.* ACM, 111–122.
[24] Xinyu Wu, Zhongzhao Hu, Xinzhe Fu, Luoyi Fu, Xinbing Wang, and Songwu Lu. 2018. Social network de-anonymization with overlapping communities: Analysis, algorithm and experiments. In *Proc. IEEE INFOCOM.*
[25] Lei Zou, Lei Chen, and M. Tamer Özsu. 2009. K-Automorphism: A General Framework For Privacy Preserving Network Publication. *PVLDB* 2 (2009), 946–957.