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, unanonymized cross-domain 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 homomorphism from abstract algebra. 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 propose an equivalent characterization of de-anonymizability, and accordingly design a sampling algorithm for approximately estimating the de-anonymizability, 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 also performed to validated all our findings. To our best knowledge, this is the first work that rigorously quantifies the relationship between the de-anonymizability and the symmetry property of general networks in a non-asymptotic manner, and thus sheds light on enhancing privacy for real networks design.

1 INTRODUCTION

As the popularity of social networks increases, the privacy of personal information in social networks becomes an issue of great concern. Concealing the personal identity in social network is one of the commonest 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] [4] [13][9] [18] [24] [5] [20] have emerged focusing on the de-anonymization problem with different aspect.

In this work, our focus is on a significant 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. In this sense, seedless de-anonymization is analog to the classical *graph matching* problem, which aims to match the nodes between two graphs with a maximum number of edges overlap.

Various algorithms for seedless de-anonymization have been proposed [16] [19][14] [13] [9] [18] [24] [5]. 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 [4] [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, but these work paid little attention to the performance of those proposed algorithms in networks. Furthermore, 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] [5] 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] [5] 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 the assumption on any network model. We will obtain the quantitative, non-asymptotic result on de-anonymizability, which is defined as the maximum number of nodes that one can expect to correctly match in the given network.

In particular, we are interested in how *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.

This paper defines the degree of symmetry of a graph by generating a matching probability matrix using the concept of graph homomorphisms. It then allows us to build the relationship between symmetry and de-anonymizability in general graphs. This result enables us to predict the maximum expectation of the number of correctly-matched nodes when a de-anonymization problem is given. We implement a practical algorithm based on sampling, which overcomes the exponential time complexity of the original exact algorithm. Besides, 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.

It should be pointed out that the focus of this work is different from that of most previous de-anonymization works, which are striving for a practical de-anonymization algorithm. This paper, on the other hand, attempts to find the theoretical upper bound of the de-anonymization performance of any algorithms on a specific de-anonymization problem. As a result, we will not present any de-anonymization algorithm in this paper. Instead, we will propose algorithms that analyzes the network symmetry to obtain the de-anonymizability of given problem based on our theoretical analysis. Due to its intractability, we also propose approximate algorithms to make the method practical.

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 bijective graph 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 equivalent characterization of de-anonymizability by enumerating permutations instead of enumerating possible underlying networks, and accordingly design a practical approximating algorithm via sampling techniques, which overcomes the exponential time complexity, and can approximately estimate de-anonymizability by virtue of a well-designed sample selection strategy. Other techniques like orbit contracting are also used to reduce the complexity.

(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 two case studies in which we extend our general method to special prior network model conditions. In Section 6 we consider the algorithm aspect of de-anonymizability. Section 7 contains the experiment verification and result. We conclude with some discussion in Section 8. Due to space limitation, some technical details are not included in this paper, and can be found in the online technical report[1].

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] [5]; 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 [4] 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[3][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 has 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 $V = \{1, 2, \ldots, n\}$ for the sake of convenience [4].

We suppose G_1 and G_2 are generated via independent *edge-samplings* of 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.

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).

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 manners.

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:

(a). 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 can be the true mapping since 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.

(b). 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.

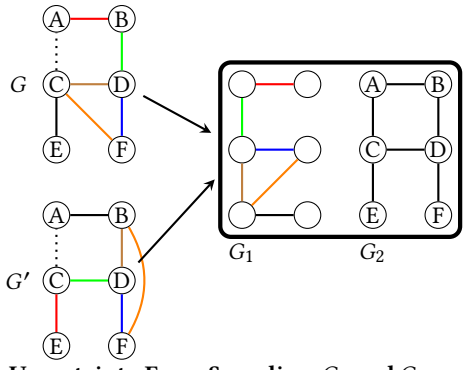


Figure 1: Uncertainty From Sampling. G_1 **and** G_2 **are known, but multiple underlying networks (including** G **and** G'**) is feasible, and can be the true underlying network.**

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 on V. 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. Note that one of the main focuses in the body part of this paper is how to obtain 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]

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

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 on V. (One of) the best permutation σ^* is

$$\sigma^* = \arg\max_{\sigma} E_\sigma = \arg\sigma \sum_{\pi\in\Pi} P(\sigma_0 = \pi|\theta) \sum_{v\in V,\sigma_0(v)=\pi(v)} 1$$

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* 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* 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 there is also 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). The symmetry of a de-anonymization problem $\theta = (G_1, G_2, s_1, s_2)$ is 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,

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

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. □

[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.

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

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

□

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. This proves the claim that de-anonymization has close relationship with the symmetry of a given de-anonymization problem.

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 the main part 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 bijective homomorphism counting from F to G. More precisely,*

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

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. The proof is based on Bayes' Rule. The detail is given in Section 1 of the technical report[1]. □

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.

We define the true mapping from G_1 to G as f_0, and the mapping from G_2 to G 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.*

The proof is based on Bayes' Law and the detail can be seen in Section 2 of the technical report [1].

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

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

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

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

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

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, $M = \sum_{G \in \mathscr{G}} P(G|\theta) C_G D_G^T$*

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. 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 supergraph, 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 $\mathscr{O} = \{v_1, v_2, \dots, v_i\}$ satisfies: if $v \in \mathscr{O}$,then for any $v' \sim v$, $v' \in \mathscr{O}$,

Based on Definition 4.5, intuitively, an orbit is a subset of node set, in which all nodes are internally symmetric.

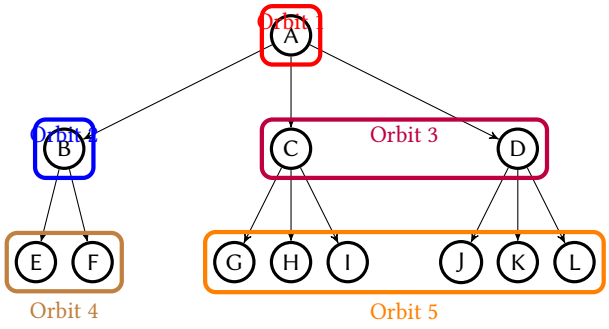


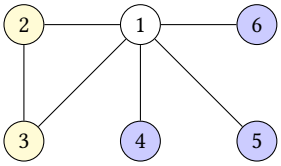
Figure 2: An Illustration of Orbit

Definition 4.6 (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 A_1 and A_2, respectively. The following Figure 3 is an illustration of automorphism matrix.

Automorphism transition matrix is used to capture the symmetry within a graph. The following proposition shows that automorphism transition matrix can used to obtain an upper bound of the de-anonymizability. The prove can be seen in the Section 3 of the technical report [1].



(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

Proposition 4.7. *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, $C_G = A_1 C_G, D_G = A_2 D_G$*

Corollary 4.8. *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 4.9 shows that the maximum diagonal sum of the product of two doubly stochastic matrices is no greater than that of any one of them.

Theorem 4.9 (Theorem 4.1 in [22]). *For two n-×-n doubly stochastic matrices A and B, $h(AB) \leq \min(h(A), h(B))$.*

Corollary 4.10. *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 4.10 indicates that intra-symmetry 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 TWO CASE STUDIES

In this section, we present two case studies to show the validation of our theoretical analysis.

5.1 Uniform Prior Probability on Underlying Network

In this case study, we assume the prior probability distribution is uniform among all the graph in $\mathscr{G}$. In this case we can obtain a quantitative result on the de-anonymization of any given de-anonymization problem. We propose a different approach to work out the probability distribution of the true mapping. This approach can reach the equivalent result to the de-anonymizability presented in our main result in Section, but it is much more useful in the approximating process. We refer to this approach as *the equivalent approach* hereinafter with contrast to the approach proposed in Section 4.3 (referred to as original approach). The equivalent approach contains the following process: in the equivalent method, instead of enumerating all possible underlying network G, we alternatively **enumerate all permutation** π from G_1 to G_2, and **calculate the probability** of each π becoming the true mapping.

For a certain permutation π, we define the union graph of G_1 and G_2 under permutation π as the minimum graph which guarantees the existence of edges from both the G_1' and G_2, where G_1' is the relabeled G_1 by permutation π.

Definition 5.1 (Union graph under given permutation). For two graphs $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$ with $V_1 = V_2 = V$ under given permutation π is construct as follow: first relabel G_1 with π so that the node i in G_1 is labeled by $\pi(i)$. We denote this relabeled graph as $G_1' = (V_1, E_1')$. Then union graph $G = (V, E_1' \cup E_2)$.

The union graph of G_1 and G_2 under given permutation π is the minimal[4] feasible underlying network, since it must keep the existence of edges from both G_1 and G_2. All the spanning supergraphs of the union graph can be candidates of feasible underlying network. The probability of a permutation to be the true mapping, consequently, is proportional to the sum of probability of all possible underlying networks of π. Theorem 5.2, which is the core of *equivalent approach*, states this idea in detail. The proof is given in Section 4 of the technical report [1].

Theorem 5.2. *Given a de-anonymization problem with parameter $\theta = (G_1, G_2, s_1, s_2)$, let Π be all the permutations from V_1 to V_2(i.e. all the permutations on V). Then matching probability matrix M can be calculated by*

$$M_{ij} = Z \sum_{\pi \in \Pi} \mathbb{1}\{\pi(i) = j\} \left(\frac{\bar{s_1}\bar{s_2}}{1 + \bar{s_1}\bar{s_2}}\right)^{|E_\pi|} \tag{3}$$

In this formula, $\bar{s_1} = 1 - s_1$, $\bar{s_2} = 1 - s_2$, $|E_\pi|$ is the edge number of the union of G_1 and G_2 under permutation π, Z is a normalization parameter which is identical among all elements in M. In particular

$$Z = \sum_{\pi \in \Pi} \left(\frac{\bar{s_1}\bar{s_2}}{1 + \bar{s_1}\bar{s_2}}\right)^{|E_\pi|}$$

Note that here we calculate the overall probability transition matrix from a totally different prospective. Recall that in the main result, we enumerate all possible underlying networks, and find all the homomorphisms from both graphs to underlying networks, both of which are time-consuming. In contrast, in this equivalent expression, we alternatively enumerate all the permutations, and calculate the probability that each permutation becomes the true mapping. Although this proposed equivalent process still incurs exponential time complexity, it will facilitate our design of a smart sampling technique that overcomes such huge complexity as well as a more practical approximate algorithm, as we will shortly present in next section.

[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

[4] by 'minimal' we mean the minimal number of edges

5.2 Erdős-Rènyi model as Underlying Network

Now we conduct another case study on obtaining de-anonymizability in the context of Erdős-Rènyi network. An Erdős-Rènyi network is characterized by two parameters n, p, representing the number of nodes, and the probability that any node pair has an edge in between, independently of each other. The de-anonymizability here refers to **the expectation of de-anonymizability of all instances** of graph generated by Erdős-Rènyi model. Since our method focuses on the automorphism and homomorphism of a graph, then the question comes that how 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 structuresin Erdős-Rènyi graph , calculate the expected times of appearance of each of them, and calculate the number of orbits they contract.

Here we only sketch the result due to space limitation. For detailed analysis readers can refer to Section 4 of the technical report [1].

5.2.1 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 5.3. 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 graph G that contains all the nodes in V_s. Notice that, $T(V_s)$ means the number of orbits that V_s 'contracts'.

Here we count up two commonest kinds of motif that have been verified to exist in many complex networks [15]:

(1) Fruits in cherry-like structure: The cherry-like motif is illustrated in [1]. Precisely, a set of k nodes forms a 'k-fruit cherry-like' motif(in short 'k-fruit) if and only if: (1) the degree of each of them is one; (2) they are connected to the same node. Proposition 5.4 shows the effect of cherry-like structures via the expected number of orbits contracted.
(2) Small connected components: A small connected component in a graph is simply a connected component whose size is less than a threshold k. For small connected components, since different values of k will lead to different types of components, here we choose the threshold $k = 7$, of which the detailed illustration of all these 13 types of components are available in Figure 1 in the technical report[1]. Table 1 lists the number of orbits contracted by each type of connected components $T(c_i)$, and Proposition 5.5 shows the effect of small connect components.

Proposition 5.4. *The number of orbits contracted by cherry-like structures (in expectation) is:*

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

Table 1: T($s2_k$) for each kind of motifs

i	$E(c_i)$	$T(c_i)$
1	$n(1-p)^{n-1}$	$2 * E(c_1) - 1$
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	$\binom{n}{6}\frac{6!}{8}p^5(1-p)^{6(n-6)+10}$	$4 * E(c_12) - 2$

Proposition 5.5. *The total number of orbits contracted by small connected components are* $\sum_i T(c_i)$.

Considering the overall effect of both types of motifs, Proposition 5.6 characterizes the de-anonymizability of a Erdős-Rènyi model.

Proposition 5.6. *The expected number of orbits (i.e. de-anonymizability) of a graph generated from Erdős-Rènyi model $G(n, p)$ is $E_{\sigma^*} = E[Orb(G)] = n - T(S) - T(C)$*

Notice that in the above formula, the right hand side is a expression of n and p. Substituting (n, p) in this formula with specified values, we can get an approximation of the de-anonymizability of G.

For ease of understanding, let us now take an Erdős-Rènyi graph with $n = 1000, p = 1/500$ as an example. After calculation we can 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 $G(1000, 1/500)$, more than 3/4 nodes can be (expected to be) de-anonymized.

5.2.2 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 ALGORITHM DESIGN OF DE-ANONYMIZABILITY

In this section we consider the algorithmic aspect of de-anonymizability. Notice that we are not focusing on 'de-anonymization algorithm', which aims to correctly map the nodes between the anonymous graph and the auxiliary graph. Instead, here we put forward algorithms that takes a parameter set $\theta = (G_1, G_2, s_1, s_2)$ as input and then outputs the de-anonymizability of the corresponding de-anonymization problem. For the fully sampled cases, we have already given out an algorithm that simply counts the orbit number

of a graph in Section 4.2. Therefore, here we only focus on the algorithm in partially sampled case.

As noted earlier, the method we have proposed to obtain the de-anonymizability in the partially sampled case, both the original one and the equivalent one, has an exponentially increasing time complexity. In this part, we design a practical algorithm based on the equivalent method of Theorem 5.2. The main idea in this algorithm is to *enumerate a subset of all permutations by sampling*.

Basically, the algorithm samples some of the permutations, and calculates the probability of these permutations. The probability transition matrix M is calculated according to the method stated in Theorem 5.2, but only the sampled permutations are taken into account.

In each sampling, a permutation is randomly sampled or artificially selected (we will discuss the selection strategy later), the union of G_1 and G_2 under this permutation is generated, and the probability of this permutation is calculated. Then, the matching probability matrix M is updated term by term according to Theorem 5.2. When the sampling process ceases, a normalization factor is also multiplied to each term of M, but the normalization factor here is the summation over the sampled permutations only..

It is easy to see the time complexity of this algorithm is $O(K(|E|+n^2))$, where K is the number of samples, $|E|$ is the number of edges in union graph G, and the n^2 term is due to updating each term after each sampling.

However, a practical number of samplings are sure to be much less than the total number of all permutations (i.e. $n!$), which may degrade the performance of accuracy. Therefore, in the sequel we further propose some refinement this algorithm.

We propose a sampling selection strategy to get a good approximation of de-anonymizability within a finite number of samplings. Theorem 5.2 implies that the probability of a permutation monotonically decreases when the number of edges of the union graph increases. Intuitively, a permutation with a larger probability will have a larger impact on the probability transition matrix. Therefore, we try to find some permutations such that the number of edges of the union graph is as small as possible. This is actually consistent with the goal of **graph matching problem**. For this reason, we borrow the idea of some state-of-the-art graph matching algorithms, and take their outputs as the permutations we sampled. Also, we interchange a small number of nodes in these permutations to enlarge the number of samplings.

A remaining consideration is how to obtain the maximum diagonal sum of the matching probability matrix. In fact, finding the maximum diagonal sum of a doubly stochastic matrix is $NP-Complete$, since it can be polynomially reduced to a {0,1} programming problem. Therefore, we use heuristic algorithms to get a sub-optimal solution of maximum diagonal sum, obtaining a lower bound.

To sum up, Algorithm 1 shows the pseudo-code of our designed algorithm.

Algorithm 1: Getting De-anonymizability in Partially Sampled Case

input : De-anonymization problem with $\theta = (G_1, G_2, s_1, s_2)$, the number of samplings
output: De-anonymizability of θ

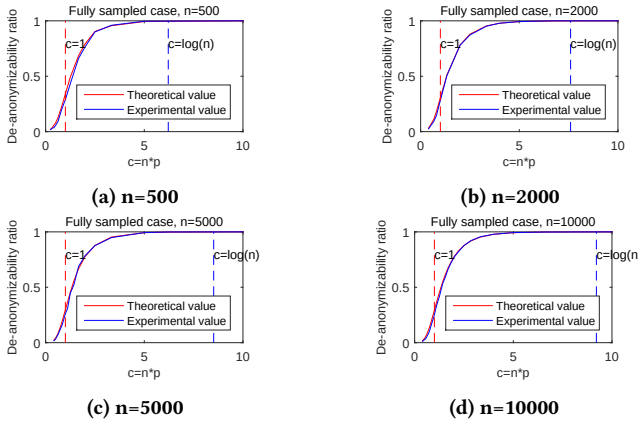
```
M ← zeros (n,n) ;  /* Get a n-by-n zero matrix */
for i = 1 ... k_sample do
    π ← getSample () ;   /* Get a permutation by random shuffling or via selection strategy */
    G ← union( G1,G2, π); |E| ← G.edgeNumber(); prob ← [(1 − s1)(1 − s2)/(1 + (1 − s1)(1 − s2)]^|E|; for k = 1 ... n do
        M [i, π(i)] ← M [i, π(i)]+ prob;
    end
end
M ← M/|M|; return max_diagonal_sum (M)
```

7 EXPERIMENT EVALUATION

7.1 Experiments on Real Dataset

To verify our result in the case study, we conduct experiments on Erdős-Rènyi graph to testify our theoretical results, in the context of both fully sampled cases and partially sampled cases.

Fully sampled case. We choose $N = 500, 2000, 5000, 10000$ as the representative of small-size network and large-size network, respectively. Using the method in Section 6, 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 5 demonstrates the high consistency of our theoretical result with the experimental result. Also, the result accords with a well-known classic conclusion [2] 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 4: Experiments on fully sampled Erdős-Rènyi graph

Partially sampled cases. In partially sampled cases, the model is equivalent to the *correlated Erdős-Rènyi model*. Recall that we use the minimum de-anonymizability of G_1 and G_2. However, we have no way to directly verify our result due to the exponential time

complexity of enumerating underlying networks or enumerating permutations. Therefore, we verify the result by (1) comparing the result with the performance of state-of-the-art de-anonymization algorithms (2) applying the method in Section 5.2.1 to get the de-anonymizability (which is merely an approximation, since the prior probability of G is not uniform among different underlying networks).

We choose $G(2000, 0.5)$ to represent the underlying network. The parameters satisfy $np >> \log(n)$, which is a threshold value related to the symmetry of Erdős-Rènyi graphs. We choose $s_1 = s_2$ for simplicity. We also use three matching algorithms mentioned above as references. The result is shown in Figure 5.

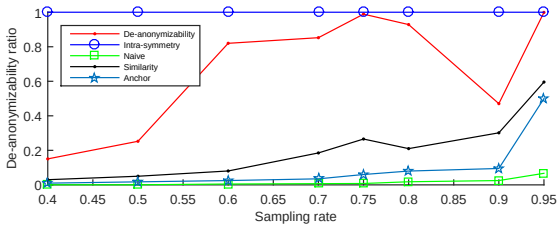


Figure 5: Validation on Partially Sampled Erdős-Rènyi network.

In this figure, *De-anonymizability* is the result applying the method in Section 4.4, and *Inter-symmetry* is expected number of orbits in G_1 and G_2. *Naive*,*Similarity* and *Anchor* represent the performance of three aforementioned algorithms respectively. From the result, we can see that although the sampling algorithm may suffer from oscillation, we indeed obtain an upper bound of the performance of de-anonymization algorithms. Since no similar previous work has been done, there is little reference that we can compare with. However, we show that there remains a large gap between the best state-of-the-art algorithm and the theoretical de-anonymizability for partially-sampled de-anonymization problems.

8 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 few. We proposed a quantitative method to determine the de-anonymizability of a non-asymptotic de-anonymization problems through the lens of symmetry.

We believe that the combination of de-anonymizability and symmetry will be of great value. As a pioneering work, our work only exhibits this idea to a small extent. In fact, the result in this paper can be further generalized. For instance, in the context of *Graph Matching* problem, out method can be used to determine how many nodes can be matched correctly at most without any modification. As a basic problem in the field pattern recognition, chemistry molecular reconstruction, etc., our method can show greater value in various situations.

We do not highlight the algorithmic aspect of this topic in this paper, so few results on the comparison of algorithms have been revealed. However, since we are focusing on the theoretical aspect of de-anonymization problem, we hope to provide a new, theoretical perspective of this problem. We believe that our work is fundamental, and will provide useful guidelines to real network design.

REFERENCES

[1] L. Fu X. Lin X. Wang B. Miao, S. Wang. 2019. Technical report for the paper De-anoymizatibility of Social Network: Through the lens of Symmetry. Retrieved Dec. 10, 2019 from https://github.com/benjie-sjtu/De-anonymizability-TechReport

[2] Béla Bollobás. 1998. *Random Graphs*. Springer New York, New York, NY, 215–252. https://doi.org/10.1007/978-1-4612-0619-4_7

[3] Peter J Cameron et al. 2004. Automorphisms of graphs. *Topics in algebraic graph theory* 102 (2004), 137–155.

[4] 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.

[5] 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.

[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.

Technical Report for the paper De-anonymizability of Social Network: Through the Lens of Symmetry

This document is a technical report, which supports and provides technical details of the proof of key theorems of the paper 'De-anonymizability of Social Network: Through the Lens of Symmetry'(hereafter denoted as the *De-anonymizability* paper).

The symbols, models and denotations of this report are the same as the 'De-anonymizability' paper. To refer to the propositions, definitions, equations or theorems in the 'De-anonymizability' paper, the original index of the 'De-anonymizability' paper is used in this technical report.

1 PROOF OF PROPOSITION 4.1

Proposition 1.1 (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*

$$P(G)\,\mathrm{hom}(G_1, G)\,\mathrm{hom}(G_2, G)((1-s_1)(1-s_2))^{|E|}$$

where $\mathrm{hom}(F, G)$ is the bijective homomorphism counting from F to G. More precisely,

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

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

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

In formula, (0) holds because G_1 and G_2 are independent samplings from G. (1) holds because there are $\mathrm{hom}(G_i, G)$ ways to sample G to get G_i. □

2 PROOF OF PROPOSITION 4.2

Proposition 2.1 (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 $\frac{1}{\mathrm{hom}(G_1,G)}$ to be the true mapping from G_1 to G. Similarly, each h_i has the probability of $\frac{1}{\mathrm{hom}(G_2,G)}$ to be the true mapping from G_2 to G.*

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 $\frac{1}{\mathrm{hom}(G_1,G)}$ to be the true mapping when G is given. Similar for G_2.

3 PROOF OF PROPOSITION 4.7

Proposition 3.1 (Proposition 4.7). *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, $C_G = A_1C_G, D_G = A_2D_G$*

We only prove the result of G_1, i.e. $C_G = A_1C_G$. The proof for G_2 is completely the same. Let $C' = A_1C_G$. Then

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

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 $D_G = A_2 D_G$.

4 PROOF OF THEOREM 5.2

Theorem 4.1. *Given a de-anonymization problem with parameter $\theta = (G_1, G_2, s_1, s_2)$, let Π be all the permutations from V_1 to V_2(i.e. all the permutations on V). Then matching probability matrix M can be calculated by*

$$M_{ij} = Z \sum_{\pi \in \Pi} \mathbb{1}\{\pi(i) = j\} \left(\frac{\bar{s_1}\bar{s_2}}{1 + \bar{s_1}\bar{s_2}}\right)^{|E_\pi|} \tag{1}$$

In this formula, $\bar{s_1} = 1 - s_1$, $\bar{s_2} = 1 - s_2$, $|E_\pi|$ is the edge number of the union of G_1 and G_2 under permutation π, Z is a normalization parameter which is identical among all elements in M. In particular

$$Z = \sum_{\pi \in \Pi} \left(\frac{\bar{s_1}\bar{s_2}}{1 + \bar{s_1}\bar{s_2}}\right)^{|E_\pi|}$$

We denote the matrix calculated by Equation 1 as M', and the matching probability matrix calculated via the original method as M. Then we need to prove that $M' = M$. First, it is easy to see that M' is also a doubly stochastic matrix. Therefore, we only need to prove that, the corresponding elements in M and M' are in proportion. Starting from Equation 1, we have

$$\begin{aligned}
M_{ij} &= \sum_{G \in \mathcal{G}} \sum_{\pi \in \Pi} P(G|\theta) P(\sigma_0 = \pi|\theta, G) \mathbb{1}\{\pi(i) = j\} \\
&= \sum_{\pi \in \Pi} \sum_{G \in \mathcal{G}} P(G|\theta) P(\sigma_0 = \pi|\theta, G) \mathbb{1}\{\pi(i) = j\} \\
&\overset{(0)}{=} \sum_{\pi \in \Pi} \sum_{G \in \mathcal{G}} \sum_{f \in \Pi} P(G|\theta) P(f|\theta, G) P(h|\theta, G)) \mathbb{1}\{\pi(i) = j\} \\
&\propto \sum_{\pi \in \Pi} \sum_{G \in \mathcal{G}} \sum_{f \in \Pi} (\bar{s_1}\bar{s_2})^{|E|} \hom(G_1, G) \hom(G_2, G) \\
&\quad P(f|\theta, G) P(h|\theta, G) \mathbb{1}\{\pi(i) = j\} \\
&\propto \sum_{\pi \in \Pi} \mathbb{1}\{\pi(i) = j\} \sum_{G \in \mathcal{G}} (\bar{s_1}\bar{s_2})^{|E|} \sum_{f \in \Pi} \hom(G_1, G) \hom(G_2, G) \\
&\quad P(f|\theta, G) P(h|\theta, G) \\
&= \sum_{\pi \in \Pi} \mathbb{1}\{\pi(i) = j\} \\
&\quad \sum_{G \in \mathcal{G}} (\bar{s_1}\bar{s_2})^{|E|} \mathbb{1}\{\exists f : f(G_1) \subset G\} \mathbb{1}\{h(G_2) \subset G\}
\end{aligned}$$

In this formula, $h = \pi^{-1} \circ f$ so that we can have $f \circ h^{-1} = \pi$.

Therefore, for each permutation π, we count up all the graphs G whose edge set is a superset of both that of G_1 and G_2. Obviously, the condition holds iff G is the spanning super graph of the union of G_1 and G_2 under permutation π. We denote the union of G_1 and G_2 under permutation π as $G_\pi = (V_\pi, E_\pi)$, and denote $N = \binom{n}{2}$. We have,

$$\begin{aligned}
M_{ij} &\propto \sum_{\pi \in \Pi} \mathbb{1}\{\pi(i) = j\} \\
&\quad \sum_{i=0}^{N - |E_\pi|} \left(\binom{N - |E_\pi|}{i} \bar{s_1}^{|E_\pi| - |E_1| + i} \bar{s_2}^{|E_\pi| - |E_2| + i}\right) \\
&= \sum_{\pi \in \Pi} \mathbb{1}\{\pi(i) = j\} \bar{s_1}^{|E_\pi| - |E_1|} \bar{s_2}^{|E_\pi| - |E_2|} \\
&\quad (1 + (\bar{s_1}\bar{s_2}))^{N - |E_\pi|} \\
&\propto \sum_{\pi \in \Pi} \mathbb{1}\{\pi(i) = j\} \bar{s_1}^{|E_\pi|} \bar{s_2}^{|E_\pi|} (1 + (\bar{s_1}\bar{s_2}))^{-|E_\pi|} \\
&= \sum_{\pi \in \Pi} \mathbb{1}\{\pi(i) = j\} \frac{\bar{s_1}\bar{s_2}}{1 + \bar{s_1}\bar{s_2}}^{|E_\pi|} \propto M'_{ij}
\end{aligned}$$

which completes the proof.

5 A CASE STUDY : DE-ANONYMIZABILITY IN ERDŐS-RÈNYI RANDOM GRAPHS

This part provides the technique details in the case study of de-anonymizability in Erdős-Rènyi graph.

As [2] has mentioned, the symmetric structure in real-world network is more likely to be 'local'. The following graph can serve as an example to show the difference between an artificially constructed graph and a network-generated graph. The left one is the famous (artificially constructed) Petersen Graph, of which the symmetry structure is completely global(that is, we have to search the whole graph to assert the existence of automorphism in the graph). The right one is an instance generated by Erdős-Rènyi model G(n=10, p=0.2). The only symmetric node pair here is (15,18) and (19,20), and detecting both of them only requires local information. On the other hand, in instances of network model, large, complicated symmetric structures are unlikely to appear.

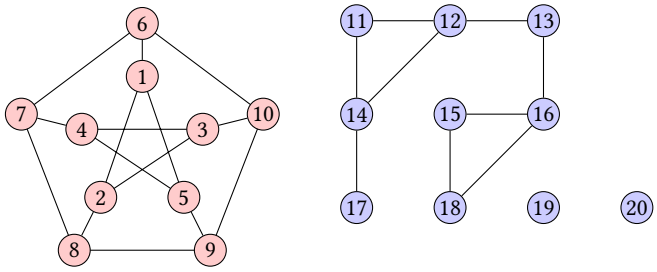


Figure 1: An comparison. The left is the (artificially constructed) Peterson Graph, in which detecting the automorphism requires searching all the nodes. The right is an instance from G(10,0.2). Only two node pairs, (15,18) and (19,20), are symmetric here.

From this phenomenon, we calculate the de-anonymizability via detecting some local automorphism structures in a network. This method is somehow like the *motif detection* [3]. 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.

5.1 Fully Sampled Case

In fully sampled cases we only need to count the orbits of a graph to get its de-anonymizability. As we have said, we only focus on those local symmetric structures of a graph. Slightly abusing the concept, we also use *motif* to refer to those local subgraphs or patterns. In this section, a motif is merely a set of nodes with some specific pattern (especially the pattern in symmetry). In the analysis of Erdős-Rènyi graph, we will clearly specify which kind of patterns we consider. 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 $|Orb(V_s)|$ means the number of orbits in the graph G that contains all the nodes in V_s. Notice that, $T(V_s)$ means the number of orbits that V_s 'contracts'. The following theorem demonstrates an upper bound of the orbit

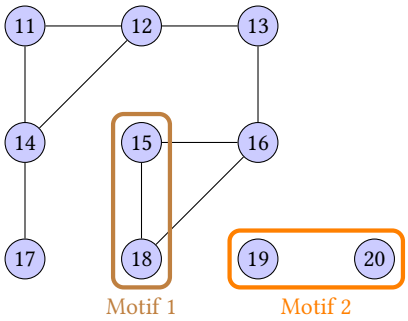


Figure 2: The instance from G(10,0.2). This picture is to show what a motif is.

number in a graph , which can be used to show the approximate de-anonymizability of a graph.

Theorem 5.1. *Given a graph $G = (V, E)$, suppose we have k motifs of G $V_{s1}, V_{s2}, \ldots, V_{sk}$. If any two motifs are vertex-disjoint, then we have $|Orb(G)| \leq |N| - \sum_{i=1}^{k} T(V_{si})$. The equality holds when:*

- *There does not exist a symmetric node pair across two motifs. That is, for any V_{si} and $V_{sj} (i \neq j)$, for any $v_i \in V_{si}$ and $v_j \in V_{sj}$, no automorphism of G will map v_i to v_j.*
- *All the motifs are found. That is, for any node v that does not belong to any of $V_{s1}, V_{s2}, \ldots, V_{sk}$, all the automorphism of G maps v to itself.*

Proof. The inequality is easy to prove by the definition of orbit. For the nodes that does not belong to any motif, it owns at most one orbit. Thus

$$\begin{aligned} |Orb(G)| &\leq |Orb(V - \bigcap_{i=1}^{k} V_{si})| + \sum_{i=1}^{k} |Orb(V_{si})| \\ &\leq |N| - \sum_{i=1}^{k} |V_{si}| + \sum_{i=1}^{k} |Orb(V_{si})| \\ &= |N| - \sum_{i=1}^{k} T(V_{si}) \end{aligned}$$

The equality holds when no additional symmetric node pair exists, which can be expressed by the two given conditions. □

We apply this theorem into the context of Erdős-Rènyi graph. We only consider the situation when $p \leq 0.5$, for a graph and its complement have the same automorphisms. We count up two commonest kinds of motifs in a graph:

(1) The fruits in a cherry-like structure.
(2) The small connected components.

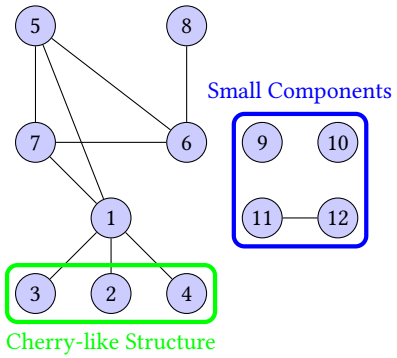


Figure 3: Two kinds of motifs we consider here.

In the sequel we discuss them respectively.

5.1.1 Cherry-like Motif. We first consider cherry-like motif, denoted as S. The cherry-like motif is shown above in Figure 3. Precisely, a set of k nodes forms a 'k-fruit cherry-like' motif(in short 'k-fruit') *iff* (1) the degrees of all of them are equal to one (2) all of them are connected to the same node. For a graph G, we define $S_k = \{V_{s_k}\}$ as the set of the all the 'k-fruit' of the G. We can obtain the expectation of how many times these type of structures appear. Specifically, for k nodes, the probability that all of them are merely connected to the same node is

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

The expectation appearance of 'k-fruit' over the whole graph is

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

And, since each 'k-fruit' obtains only one orbit,

$$\sum_{V_{s_k} \in S_k} T(V_{s_k}) = (k-1)|S_k|$$

We define $S = \bigcup_{k=2}^{n} S_k$ as the set of all cherry-like motifs in graph G. We cannot simply add the result above, since repeated counting appears. For example, a '3-fruit' structure also contains $C_3^2 = 3$ '2-fruit' structures. By the inclusive-exclusive principle, we have

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

by which we can calculate the (expectation) number of 'cherry-like' motifs and the number of orbits they contract given any specific parameters n, p.

5.1.2 Small connected components. According to [1], it has been shown that in Erdős-Rènyi graph $G(n,p)$ with $np > \log(n)$, almost every graph generated by $G(n,p)$ is asymmetric. On the other hand, when $np \leq \log(n)$, most graph generated by $G(n,p)$ is not fully connected, and the existence of small connected components (with contrast to the giant component in Erdős-Rènyi graph) are likely to contain symmetric nodes. Therefore, we consider another type of motif: small connected components (denoted as C). Here we simply define that small connected components in a graph are all the connected components whose sizes are less than a threshold k. Notice that if two components share the same shape, not only the nodes within the motif may belong to the same motif (since most of the small components are symmetric), the corresponding nodes in different components will also belong to the same orbit. To reduce the number of components we have to enumerate, we use the fact that almost all the small components in Erdős-Rènyi graph are trees[1]. Therefore, we only enumerate all the trees with size less than k. As a trade-off, a larger k will detect more automorphisms and improve the accuracy the result, while exponentially increases the number of components that are necessary to enumerate. As an illustration, Figure 4 shows all 13 kinds of trees whose size is less than or equal to 6 [4].

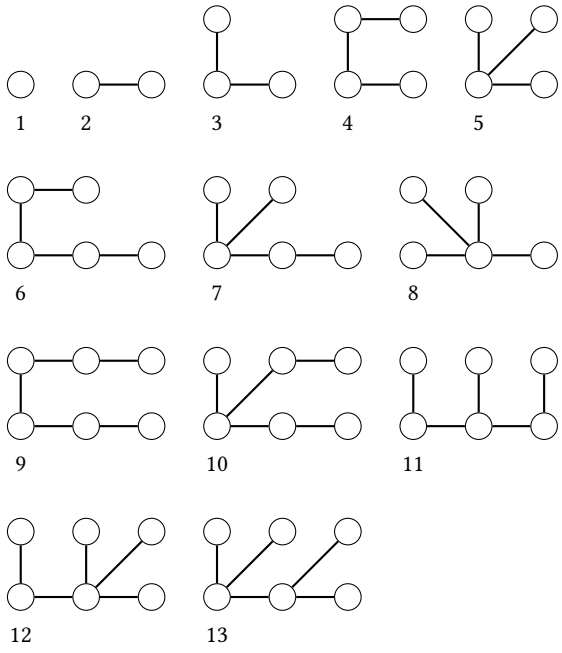


Figure 4: Different types of small components.

For each type, the expectation of times of its appearance is calculated in order to obtain the number of orbits it contracts. As an example, we show in detail how to calculate the expectation of C_1 (i.e. an isolated node). The probability that it is connected to none of the other points is

$$P(deg_i = 0) = (1-p)^{n-1}$$

Then the expectation over the whole graph is equal to

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

which shows the 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. Therefore, we only contract the orbit **across** the components. Take type 3 as an example, the expectation of appearance of type 3 components is:

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

It is easy to observe that, the orbit number, after contraction in C, is 2. However, in each type-3 component (i.e. C_3), the two 'fruit' nodes have already been contracted into one orbit after the analysis Section 5.1.1, and therefore, only $2*E(c_3)-2$ (rather than $3*E(c_3)-2$) orbits are contracted. We list the results of each type in the Table 1. Consequently, $T(C) = \sum_i T(c_i)$ represents the total number of

Table 1: $T(C)$ for each kind of motifs

i	$E(c_i)$	$T(c_i)$
1	$n(1-p)^{n-1}$	$2 * E(c_1) - 1$
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	$\binom{n}{6} \frac{6!}{2} p^5 (1-p)^{6(n-6)+10}$	$6 * E(c_1 0) - 4$
12	$\binom{n}{6} \frac{6!}{3!} p^5 (1-p)^{6(n-6)+10}$	$4 * E(c_1 1) - 4$
13	$\binom{n}{6} \frac{6!}{8} p^5 (1-p)^{6(n-6)+10}$	$4 * E(c_1 2) - 2$

orbits contracted by small connected components. In fact, a larger component size threshold k can detect more small components, but when one motif does not exist, extra orbits are contracted wrongly. Therefore, we dynamically decide k when conducting our experiments. After analyzing those two kinds of motifs, the de-anonymizability of a graph G is then given by

$$E_{\sigma^*} = E[Orb(G)] = n - T(S) - T(C)$$

Notice that in the above formula, the right hand side is a expression of n and p. Substituting (n,p) in this formula with specified values, we can get an approximation of the de-anonymizability of G.

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

$$T(S) = 33.69, T(C) = 181.49$$
$$E_{\sigma^*} = n - T(S) - T(C) = 784.82$$

[1] which means in a graph given by $G(1000, 1/500)$, more than 3/4 nodes can be (expected to be) de-anonymized. We will validate the result in the experiment section. The following theorem is used to

[1]In calculation, we do some modification to our formula in order to prevent extra orbits to be contracted. The detail will be described in the experiment part.

prove the correctness of the method, which shows that, there is almost no symmetric node pairs in the giant component of Erdős-Rènyi graphs.

Theorem 5.2. *For the Erdős-Rènyi model, $p = \Omega(\frac{1}{n})$ and $p \le \frac{1}{2}$, there are $o(1)$ nodes that are symmetric in the giant component.*

Proof. Suppose there are n nodes in the giant component and the edge existence probability is p, where $p = \Omega(\frac{1}{n})$ and $p \le \frac{1}{2}$. The distribution of the node degree satisfies Poisson distribution. We aim to figure out the the probability of two nodes N_a, N_b being symmetric given that both of their degrees are larger than one. Before the derivation, we first denote the nodes that are both N_a and N_b's neighbors to be 1-hop shared neighbors, and denote the nodes that are only N_a's or N_b's neighbors to be 1-hop free neighbors. If N_a and N_b are symmetric, these 1-hop free neighbors can be divided into several pairs, and each of them contains one N_a's neighbor and one N_b's, which are also symmetric. Similarly, we denote those nodes that connect two symmetric $(k-1)$-hop free neighbors as k-hop shared neighbors and nodes that only connect to one of them as k-hop free neighbors. Suppose the numbers of N_a's and N_b's k-hop free neighbors are n_{ak} and n_{bk}, and the numbers of N_a's and N_b's k-hop share neighbors are n'_{ak} and n'_{bk}.

If N_a, N_b are symmetric, $n_{ak} = n_{bk} = n_k$ and $n'_{ak} = n'_{bk} = n'_k$. Moreover, since their k-hop free neighbors can be divided into symmetric pairs, the number of N_a's k-hop free neighbors whose degrees are even is equal to that of N_b's. The probability of a node with even degree is:

$$P_{even} = \sum_{j=1}^{\lfloor \frac{n-1}{2} \rfloor} \frac{\lambda^{2j}}{(2j)!} e^{-\lambda}, \tag{2}$$

where $\lambda = (n-1)p$. Since $n \to \infty$, we can derive that

$$\frac{P_{even}}{P_{odd}} = \frac{\sum_{j=1}^{\lfloor \frac{n-1}{2} \rfloor} \frac{\lambda^{2j}}{(2j)!} e^{-\lambda}}{\sum_{j=1}^{\lfloor \frac{n-2}{2} \rfloor} \frac{\lambda^{2j+1}}{(2j+1)!} e^{-\lambda}} \to \frac{e^{\lambda} + e^{-\lambda}}{e^{\lambda} - e^{-\lambda}} \to 1 \tag{3}$$

Therefore, $P_{even} \to \frac{1}{2}$ and $P_{odd} \to \frac{1}{2}$. Let A_1 be the event that the number of N_a's k-hop free neighbors whose degrees are even is equal to that of N_b's. Based on Stirling's formula, we can get that:

$$\begin{aligned} P_{A_1} &= \sum_{i=0}^{n-1} \left[P_{even}^{i} P_{odd}^{n_k - i} \binom{n_k}{i} \right]^2 \\ &\to \sum_{i=0}^{n-1} 2^{-2n_k} \binom{n_k}{i}^2 \\ &= 2^{-2n_k} \binom{2n_k}{n_k} \\ &= \frac{(2n_k)!}{2^{2n_k} (n_k!)^2} \\ &\to \frac{\sqrt{4\pi n_k} \left(\frac{2n_k}{e}\right)^{2n_k}}{2\pi n_k 2^{2n_k} \left(\frac{n_k}{e}\right)^{2n_k}} \\ &= \frac{1}{\sqrt{\pi n_k}} \end{aligned} \tag{4}$$

Furthermore, when A_1 is satisfied, without loss of generality, we suppose the number of k-hop free neighbors with even degree is larger than or equal to $\frac{n_k}{2}$. Let A_2 be the event that the number of N_a's k-hop free neighbors whose degrees satisfy $d = 4i$, where $i \ge 0$, is equal to that of N_b's. With the same derivation, we can get that $P_{A_2} \ge \sqrt{\frac{2}{\pi n_k}}$. With the same analysis, we can also derive that $P_{A_3} \ge \sqrt{\frac{4}{\pi n_k}}$, $P_{A_4} \ge \sqrt{\frac{8}{\pi n_k}}$ and $P_{A_5} \ge \sqrt{\frac{16}{\pi n_k}}$, where A_3, A_4 and A_5 are events that the number of N_a's k-hop free neighbors whose degrees satisfy $d = 8i$, $d = 16i$ and $d = 32i$, where $i \ge 0$, is equal to that of N_b's respectively.

Suppose N_a, N_b have k-hop free neighbors, where $k \le l$, then we can bound the probability that N_a, N_b are symmetric by:

$$\begin{aligned} P_{(N_a,N_b)\in sym} &\le P_{deg(N_a)=deg(N_b)} \prod_{k=1}^{l} \prod_{j=1}^{5} P_{A_j} \\ &= P_{deg(N_a)=deg(N_b)} \prod_{k=1}^{l} \frac{32}{(\pi n_k)^{\frac{5}{2}}}, \end{aligned} \tag{5}$$

where $P_{deg(N_a)=deg(N_b)} = \sum_{i=2}^{n-1} \left(\frac{\lambda^i}{i!} e^{-\lambda}\right)^2$.

If there exists k_0 satisfies that $n_{k_0} = \Theta(n)$, then we have:

$$P_{(N_a,N_b)\in sym} \le \frac{32}{(\pi n_{k_0})^{\frac{5}{2}}} = o(\frac{1}{n^2}). \tag{6}$$

Then we have $\binom{n}{2} P_{(N_a,N_b)\in sym} = o(1)$.

Otherwise, for any constant $\epsilon > 0$ and two nodes N_1, N_2 in the network, let B be the event: $deg(N_1) = deg(N_2) \ge (\frac{1}{2} + \epsilon)n$. Since $p \le \frac{1}{2}$, then $\lambda = (n-1)p < \frac{n}{2}$ and we have:

$$\begin{aligned} P_B &= \sum_{i=(\frac{1}{2}+\epsilon)n}^{n-1} \left(\frac{\lambda^i}{i!} e^{-\lambda}\right)^2 \le \max_{i \ge (\frac{1}{2}+\epsilon)n} \left(\frac{\lambda^i}{i!} e^{-\lambda}\right) \sum_{i=(\frac{1}{2}+\epsilon)n}^{n-1} \frac{\lambda^i}{i!} e^{-\lambda} \\ &< \left(\frac{\lambda}{\frac{1+\epsilon}{2} n}\right)^{\frac{\epsilon}{2} n} e^{-\lambda} < \left(\frac{1}{1+\epsilon}\right)^{\frac{\epsilon}{2} n}. \end{aligned} \tag{7}$$

Then we have $\binom{n}{2} P_B = o(1)$, which means there are almost no node pairs that have the same degree larger than $(\frac{1}{2} + \epsilon)n$. Therefore, for N_a, N_b or two nodes in k-hop free heighbors, suppose the degrees of them are both d, where $2 < d < (\frac{1}{2} + \epsilon)n$. The probability of the number of nodes that connect only one of them is n_f satisfies $P_f \le \sum_{d=2}^{(\frac{1}{2}+\epsilon)n} P_{deg=d} \binom{d}{d-n_f} p^{d-n_f}$, when $n_f = o(d)$, we can derive that $P_f = O(\frac{1}{n^2})$. Therefore, $n_f = \Theta(d)$ with probability 1 and the total number of free neighbors is $\sum_{k=1}^{l} n_k = n - 1 - \sum_{k=1}^{l} n'_k = \Theta(n)$.

Then we can derive that

$$
\begin{aligned}
P_{(N_a,N_b)\in sym} &< \prod_{k=1}^{l} \frac{32}{(\pi n_k)^{\frac{5}{2}}} \\
&\leq \frac{32}{\pi^{\frac{5}{2}}(\sum_{k=1}^{l} n_k - l)^{\frac{5}{2}}} \prod_{k=1}^{l-1} \frac{1}{\sqrt{\pi}} \\
&= \frac{32}{\pi^{\frac{5}{2}} \left(\pi^{\frac{l-1}{5}} (\sum_{k=1}^{l} n_k - l)\right)^{\frac{5}{2}}} \\
&< \frac{32}{\pi^{\frac{5}{2}} \left(\sum_{k=1}^{l} n_k\right)^{\frac{5}{2}}} \\
&= o(\frac{1}{n^2})
\end{aligned}
\tag{8}
$$

Therefore, $\binom{n}{2} P_{(N_a,N_b)\in sym} = o(1)$.

Above all, there are almost no symmetric node pairs with degree larger than 1 in the network.

□

5.2 Partially Sampled Case

The partially sampled case is relatively difficult to analyze. The underlying network G is generated by model $G(n, p)$, and G_1, G_2 are independent samplings of G, the sampling rate of which is s_1,s_2, respectively. In fact, G_1 and G_2 can be seen generated by $G(n, ps_1)$ and $G(n, ps_2)$ respectively. We can calculate the orbit number of G_1 and G_2, and take the smaller one as an upper bound of de-anonymizability of the problem.

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] Ben D. Macarthur, Rubén J. Sánchez-García, and James W. Anderson. 2008. Symmetry in complex networks. *Discrete Applied Mathematics* 156, 18 (2008), 3525–3531.

[3] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. Network Motifs: Simple Building Blocks of Complex Networks. *Science* 298, 5594 (2002), 824–827. https://doi.org/10.1126/science.298.5594.824 arXiv:https://science.sciencemag.org/content/298/5594/824.full.pdf

[4] Ronald C. Read and Robin J. Wilson. 1998. An atlas of graphs. *Clarendon Press Oxford University Press New York* (1998).