Robust Graph Matching when Nodes are Corrupt

Taha Ameen [1] **Bruce Hajek** [1]

Abstract

Two models are introduced to study the problem of matching two correlated graphs when some of the nodes are corrupt. In the weak model, a random subset of nodes in one or both graphs can interact randomly with their network. For this model, it is shown that no estimator can correctly recover a positive fraction of the corrupt nodes. Necessary conditions for any estimator to correctly identify and match all the uncorrupt nodes are derived, and it is shown that these conditions are also sufficient for the k-core estimator. In the strong model, an adversarially selected subset of nodes in one or both graphs can interact arbitrarily with their network. For this model, detection of corrupt nodes is impossible. Even so, we show that if only one of the networks is compromised, then under appropriate conditions, the maximum overlap estimator can correctly match a positive fraction of nodes albeit without explicitly identifying them.

1. Introduction

Graph matching is the problem of finding the latent correspondence between two edge-correlated networks, i.e recovering (unknown) node identities in a graph by matching to a correlated graph with (known) node identities. It is a ubiquitous problem in machine learning with applications to social networks (Narayanan & Shmatikov, 2009; 2008), biological networks (Singh et al., 2008; Kazemi et al., 2016), computer vision (Schellewald & Schnörr, 2005), natural language processing (Haghighi et al., 2005), and graph neural network based machine translation (Xu et al., 2019). Over a decade of progress has led to a sound understanding of the fundamental limits of graph matching in the case of correlated Erdős-Rényi graphs; an overview is provided in Section 1.1.

Graph matching can be viewed as a noisy version of the graph isomorphism problem, and is motivated by real-world networks often being correlated but non-identical. For instance, the interaction graphs of two social networks (such as Twitter and Flickr) are correlated because users are likely to connect with the same people in both networks. Indeed, it was shown in (Narayanan & Shmatikov, 2009) that the identities of some nodes in the Twitter graph, despite being anonymized, could be recovered simply by matching to the Flickr network. Another example is protein-protein interaction (PPI), where the interactome of an organism is constructed by connecting two interacting proteins with an edge. The interactomes of two closely related species are then correlated through a latent correspondence. Matching these interactomes allows the identification of conserved functional components between the two species (Singh et al., 2008; Bandyopadhyay et al., 2006).

All these networks are more complicated than correlated Erdős-Rényi graphs, and so designing robust algorithms is paramount in practice. In a sense, algorithms for graph matching may themselves be viewed as robust algorithms for graph isomorphism, with the extent of robustness quantified through tolerance to *edge-corruptions*. In the present work, it is argued that robustness towards *node-corruptions* is also an important factor to consider when designing algorithms. For instance, a user's Twitter account may get hacked, causing them to connect and disconnect arbitrarily with other users. Similarly, a protein in a PPI network may interact randomly with other proteins due to a variety of factors. For example, the popular *Yeast two-hybrid* method constructs a PPI network by pairwise examining the interaction between two proteins by fusing them both to a transcription binding domain in the yeast cell (Fionda, 2019). However, if one of the proteins is itself an unknown transcription factor, then false positive interactions may be recorded. Conversely, if it is toxic to the cell, or requires post-translational modifications that do not take place in yeast cells, then false negatives can occur (Koh et al., 2012).

These phenomena are better captured by node corruptions than edge corruptions. Here, the number of corrupted node pairs may even be quadratic in the size of the graph, but there is a spatial clustering of the noise: each corrupted edge has at least one end point in a subset of corrupted nodes.

[1] Department of Electrical and Computer Engineering, and the Coordinated Science Laboratory, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. Correspondence to: Taha Ameen <tahaa3@illinois.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Contributions To our knowledge, this is the first work to consider fundamental limits of graph matching with node-corruptions. Two models are studied:

1. *Weak adversary*: The adversary selects a random set of nodes in each network and resamples all the edges adjacent to the set without observing the graphs. This models random behavior of unknown proteins in a PPI network.
2. *Strong adversary*: The adversary selects an arbitrary set of nodes in each network and rewires all the edges adjacent to the set after observing the graphs. This models malicious behavior of hacked users in social networks.

In the setting of the weak adversary, we show that no estimator correctly matches any positive fraction of corrupted nodes. Conversely, under appropriate conditions, the k-core estimator correctly matches almost all of the uncorrupted nodes and none of the corrupted nodes. Under a further condition that is also necessary, it identifies the corrupted nodes and correctly matches all the uncorrupted nodes. Our simulations suggest that there is a gap between these fundamental limits and the performance of commonly used computationally feasible algorithms: Section 5 provides details.

In the setting of the strong adversary, we show that an analogous detection of corrupted nodes is impossible. Even so, when only one of the networks is corrupted, the maximum overlap estimator outputs a matching that correctly matches a positive fraction of the uncorrupted nodes. An explicit lower bound on the fraction of correctly matched nodes as a function of the fraction of corrupted nodes is also derived.

These results can be contextualized through their practical implications to social and biological networks. A good understanding of these networks is useful to appropriately process and augment data for downstream machine learning tasks. For example, our impossibility result for the strong adversary provides an algorithm to sanitize social network data such that precise de-anonymization of any positive fraction of the users through matching to another network is impossible. Further, our achievability result for the weak adversary provides a framework to detect proteins in PPI networks that register random interactions. This is discussed in Section 5.

1.1. Related work

The problem of finding necessary and sufficient conditions for matching correlated random graphs was considered in (Pedarsani & Grossglauser, 2011). Ever since, a growing line of work has improved these results for exact recovery (Cullina & Kiyavash, 2016; 2017), almost-exact recovery (Cullina et al., 2019; Wu et al., 2022) and partial recovery (Hall & Massoulié, 2023; Ganassali et al., 2021; Ding & Du, 2022). In parallel, other works have investigated computationally feasible algorithms (Barak et al., 2019; Ding et al., 2021; Fan et al., 2022; Mao et al., 2021), culminating in algorithms that run provably well in polynomial time when the graphs are far from isomorphic (Mao et al., 2023a;b; Ding & Li, 2023).

All these works study correlated Erdős-Rényi graphs, for which the fundamental limits of matching are now well understood. Subsequently, an emerging line of work is expanding the scope of the problem. For instance, (Rácz & Sridhar, 2021) and (Gaudio et al., 2022) study the graph matching problem in correlated stochastic block models, while (Rácz & Sridhar, 2023) and (Ding et al., 2023) study graph matching in inhomogeneous random graphs, focusing respectively on fundamental limits and efficient algorithms.

Recently, there is growing interest in studying robust variants of estimation problems in graphs when a positive fraction of nodes interact adversarially with their network. For example, (Acharya et al., 2022) studies the problem of estimating the parameter p of an Erdős-Rényi graph, whereas recently, (Liu & Moitra, 2022) and (Hua et al., 2023) study the community detection problem in this setting. Finally, the model in (Mitzenmacher & Morgan, 2018) allows for a simpler version of node corruptions, where only a sublinear fraction of nodes to be corrupt. All these works provide insight into developing robust algorithms that are better suited for real-world networks.

2. Preliminaries

Notation Let $[n]$ denote the set $\{1, 2, \cdots, n\}$ and let $\binom{[n]}{2}$ denote the set of unordered pairs $\{\{u, v\} : u, v \in [n]$ and $u \neq v\}$. For a graph G on n nodes, assume that its node set $V(G)$ is $[n]$, and so its edge set $E(G)$ is a subset of $\binom{[n]}{2}$. In this work, graphs are undirected and unweighted. For a node pair $\{i, j\}$, denote its *edge status* by $G\{i, j\}$, where $G\{i, j\} = 1$ if $\{i, j\} \in E(G)$ and $G\{i, j\} = 0$ otherwise. The graph G is sampled from the Erdős-Rényi (ER) distribution, denoted $G \sim \mathsf{ER}(n, p)$, if G has n nodes and each edge in G exists with probability p. Let π be a permutation on $[n]$ and denote by G^π the graph obtained by relabeling nodes in G according to π, so that

$$G\{i, j\} = G^\pi\left\{\pi(i), \pi(j)\right\} \; \forall \{i, j\} \in \binom{[n]}{2}.$$

Standard asymptotic notation ($O(\cdot), o(\cdot), \Theta(\cdot), \cdots$) is used throughout, and it is implicit that $n \to \infty$.

In this work, $\mathsf{Bern}(p)$ and $\mathsf{Bin}(n, p)$ denote respectively the Bernoulli and binomial distribution. The hypergeometric distribution is denoted by $\mathsf{HypGeom}(n, k, m)$. A random variable with this distribution counts the number of successes in a sample of k elements drawn without replacement from a population of n individuals, of which m elements are considered successes.

2.1. Correlated graphs and corruption models

In all definitions below, n is a positive integer and p, s are in $[0, 1]$. Further, G_1 and G_2 are graphs with $V(G) = [n]$ and π^* is a permutation on $[n]$.

Definition 2.1 (Correlated Erdős-Rényi graphs)**.** A pair of graphs G_1 and G_2 are said to be correlated ER graphs with parameters p and s, if they are obtained as follows: Two graphs G_1 and G_2' are obtained by independently subsampling each edge of a parent graph $G \sim \mathsf{ER}(n, p)$ with probability s. Independently, a permutation π^* is sampled uniformly at random, and G_2 is obtained as $G_2 = G_2'^{\pi^*}$.

Marginally, G_1 and G_2 each follow the $\mathsf{ER}(n, ps)$ distribution. However, the two graphs are edge-wise correlated according to an underlying permutation π^*. Next, two corruption models are presented, motivated respectively by protein-protein interaction and social network de-anonymization.

Definition 2.2 (Adversaries)**.** For any set $\mathcal{B} \subseteq [n]$, define by $E_{\mathcal{B}}$ the set of all node pairs adjacent to $\mathcal{B}$, i.e.

$$E_{\mathcal{B}} = \left\{ \{i, j\} \in \binom{[n]}{2} : i \in \mathcal{B} \text{ or } j \in \mathcal{B} \right\}.$$

Let G_1 and G_2 be two graphs and let γ and λ be in $[0, 1]$. The weak and strong adversaries are defined as:

- Weak adversary: Let $q \in [0, 1]$. The weak adversary with parameters (γ, λ, q) acts on G_1 and G_2 as follows: It selects sets $\mathcal{B}_1 \subseteq V(G_1)$ and $\mathcal{B}_2 \subseteq V(G_2)$ of nodes such that $|\mathcal{B}_1| = \lambda\gamma n$ and $|\mathcal{B}_2| = (1 - \lambda)\gamma n$, independently and uniformly at random. It then assigns the edge status of each node pair in $E_{\mathcal{B}_1}$ and $E_{\mathcal{B}_2}$ independently from the $\mathsf{Bern}(q)$ distribution.
- Strong adversary: A strong adversary A is a possibly random rule that given $(G_1, G_2, \gamma, \lambda)$ determines sets $\mathcal{B}_1 \subseteq V(G_1)$ and $\mathcal{B}_2 \subseteq V(G_2)$ such that $|\mathcal{B}_1| = \gamma\lambda n$ and $|\mathcal{B}_2| = (1 - \lambda)\gamma n$, and it determines the edge status of all node pairs in $E_{\mathcal{B}_1}$ and $E_{\mathcal{B}_2}$.

The output of the adversary is denoted $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$, where $\widetilde{G}_1$ and $\widetilde{G}_2$ are the corrupted graphs.

In words, the adversary corrupts a total of γn nodes, of which a fraction λ are in G_1 and the rest are in G_2. It then modifies the edge status of each node pair with at least one corrupted end point. Note that if (G_1, G_2) are correlated ER graphs with underlying correspondence π^*, then for any q, the weak adversary defines a joint distribution on $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2, \pi^*)$. Similarly, any strong adversary A defines a distribution on $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2, \pi^*)$.

2.2. Matchings and estimators

Definition 2.3. A *matching* μ is an injective function with domain $\mathsf{dom}(\mu) \subseteq [n]$ and codomain $[n]$.

Note that permutations are matchings with domain equal to $[n]$. An *estimator* $\mathcal{E}$ is a mapping that takes in a pair of corrupted graphs $(\widetilde{G}_1, \widetilde{G}_2)$ and outputs a matching μ. In doing so, it attempts to recover the latent permutation π^* between the uncorrupted graphs G_1 and G_2. Two estimators that have been studied in the absence of any adversary are the maximum overlap estimator $\mathcal{E}_{\mathsf{MO}}$ and the k-core estimator $\mathcal{E}_k$. They are presented next using the following definition.

Definition 2.4 (Intersection Graph)**.** Let H_1 and H_2 be two graphs and let μ be a matching. The intersection graph $H_1 \wedge_\mu H_2$ is a graph with node set $\mathsf{dom}(\mu)$, such that for any two nodes $i, j \in \mathsf{dom}(\mu)$, the pair $\{i, j\}$ is an edge in $H_1 \wedge_\mu H_2$ if and only if $\{i, j\}$ is an edge in H_1 and $\{\mu(i), \mu(j)\}$ is an edge in H_2.

Maximum overlap estimator For two graphs H_1 and H_2, the maximum overlap estimator $\mathcal{E}_{\mathsf{MO}}(H_1, H_2)$ outputs a complete matching, i.e. a permutation $\widehat{\mu}_{\mathsf{MO}}$ that maximizes the number of edges in the corresponding intersection graph:

$$\widehat{\mu}_{\mathsf{MO}} \in \arg\max_{\mu} |E(H_1 \wedge_\mu H_2)|.$$

For correlated ER graphs, the maximum overlap matching is the maximum likelihood estimator for exact recovery in the absence of the adversary, and therefore maximizes the probability of exact recovery.

k-core estimator The k-core of a graph G, denoted $\mathsf{core}_k(G)$ is the largest set of vertices A of G such that the induced subgraph on A has minimum degree at least k. For any two graphs H_1 and H_2 and non-negative integer k, a matching μ is said to be a *k-core matching* of H_1 and H_2 if the minimum degree in $H_1 \wedge_\mu H_2$ is at least k.

The *k-core estimator* $\mathcal{E}_k(H_1, H_2)$ selects a k-core matching $\widehat{\mu}_k$ such that $|\mathsf{dom}(\widehat{\mu}_k)| \geq |\mathsf{dom}(\mu_k)|$, for any other k-core matching μ_k.

2.3. Recovery objectives

For a matching μ and a permutation π^* on $[n]$, denote by $\mathsf{ov}(\mu, \pi^*)$ the overlap between μ and π^*, i.e. the number of nodes on which μ and π^* agree:

$$\mathsf{ov}(\mu, \pi^*) := |\{i \in \mathsf{dom}(\mu) : \mu(i) = \pi^*(i)\}|.$$

Upon observing only the pair of corrupted graphs $(\widetilde{G}_1, \widetilde{G}_2)$, the objective is to find a matching $\widehat{\mu}$ to maximize the overlap between $\widehat{\mu}$ and the latent permutation π^*. Definition 2.5 captures this notion.

Definition 2.5 (α-recovery)**.** Let $\alpha \in (0,1]$. An estimator that outputs a matching $\widehat{\mu}$ is said to achieve

(i) α-recovery, if $\mathbb{P}\left(\frac{\mathsf{ov}(\widehat{\mu},\pi^*)}{n} \geq \alpha\right) = 1 - o(1)$.

(ii) almost α-recovery, if for every $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{\mathsf{ov}(\widehat{\mu},\pi^*)}{n} \geq \alpha - \varepsilon\right) = 1 - o(1).$$

Graph matching is often a precursor to downstream tasks. Consequentially, an estimator is often useful in practice only if it correctly matches all the nodes in its domain. This concept is captured through the notion of *precision*.

Definition 2.6 (Precision)**.** The *precision* ρ of a matching $\widehat{\mu}$ is the fraction of the matching that is correct, i.e.

$$\rho(\widehat{\mu}) := \frac{\mathsf{ov}(\widehat{\mu},\pi^*)}{|\mathsf{dom}(\widehat{\mu})|}.$$

For a sequence of graph-pairs $(\widetilde{G}_1, \widetilde{G}_2)_n$ on n vertices, an estimator $\mathcal{E}(\widetilde{G}_1, \widetilde{G}_2)$ that outputs a matching $\widehat{\pi}$ is *precise* if $\mathbb{P}(\rho(\widehat{\pi}) = 1) = 1 - o(1)$. For any $\varepsilon > 0$, it is said to be ε-imprecise if $\mathbb{P}(\rho(\widehat{\pi}) \leq 1 - \varepsilon) = 1 - o(1)$.

3. Main Results

Impossibility and achievability results are presented separately for the weak and strong adversary. In all the results, n is a positive integer and p, s, γ, λ are real numbers such that $p \in (0,1)$, $s \in (0,1]$, and $\gamma, \lambda \in [0,1]$. For a pair of graphs G_1 and G_2 with underlying correspondence π^*, and the output $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ of an adversary, denote by $\mathcal{B}_2'$ the preimage of $\mathcal{B}_2$ under π^*, i.e. $\mathcal{B}_2' := \{i \in [n] : \pi^*(i) \in \mathcal{B}_2\}$.

3.1. Results on the weak adversary

Our first result is an impossibility result that holds for any estimator.

Theorem 3.1. *Let G_1 and G_2 be two correlated ER graphs with parameters p and s, and let π^* denote the underlying correspondence between them. Let $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ be the output of a weak adversary with parameters (γ, λ, ps). Let $\mathcal{E}(\widetilde{G}_1, \widetilde{G}_2)$ be any estimator that returns a matching μ. Let $\alpha^* = 1 - \gamma + \lambda(1-\lambda)\gamma^2$.*

(i) If $\mathcal{E}$ is precise, then

$$\mathbb{P}(\mathsf{dom}(\mu) \subseteq (\mathcal{B}_1 \cup \mathcal{B}_2')^c) = 1 - o(1).$$

(ii) If $\mathcal{E}$ achieves almost α-recovery, then $\alpha \leq \alpha^*$.

(iii) Let $p = C\log(n)/n$ and $\lambda \in \{0,1\}$. If $\mathcal{E}$ is precise and achieves α^*-recovery, then $C \geq 1/(s^2\alpha^*)$.

Part (i) of Theorem 3.1 states that no precise estimator can correctly recover any of the corrupted nodes with high probability. Part (ii) precludes the possibility of almost 1-recovery (and therefore also 1-recovery) when $\gamma > 0$, in stark contrast to known achievability results in the absence of adversary.

Next, we show that when the average degrees of G_1 and G_2 are logarithmic in the number of nodes n (i.e. $p = C\log(n)/n$ for some positive constant C), the k-core estimator performs optimally for an appropriate choice of k. Specifically, we prove that there is a threshold $\tau \equiv \tau(s, \gamma, \lambda)$ such that if $C > \tau$, then the k-core estimator identifies and matches all the uncorrupted nodes. Further, if $C < \tau$, then the estimator matches all but a vanishing fraction of the uncorrupted nodes and none of the corrupted nodes.

Theorem 3.2. *Let C be a positive constant and suppose that $p = C\log(n)/n$. Let G_1 and G_2 be two correlated ER graphs with parameters p and s, and let π^* denote the underlying correspondence between them. Let $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ denote the output of a weak adversary with parameters (γ, λ, ps), and let $\widehat{\mu}_k$ be the matching output by the k-core estimator $\mathcal{E}_k(\widetilde{G}_1, \widetilde{G}_2)$ with $k = \sqrt{\log n}$. Let $\alpha^* = 1 - \gamma + \lambda(1-\lambda)\gamma^2$.*

(i) $\mathcal{E}_k$ is precise.

(ii) If $C > 1/(s^2\alpha^*)$, then $\mathcal{E}_k(\widetilde{G}_1, \widetilde{G}_2)$ achieves α^*-recovery if $\lambda \in \{0,1\}$, and achieves almost α^*-recovery if $\lambda \in (0,1)$. Further,

$$\mathbb{P}(\mathsf{dom}(\widehat{\mu}_k) = (\mathcal{B}_1 \cup \mathcal{B}_2')^c) = 1 - o(1).$$

(iii) If $C < 1/(s^2\alpha^*)$, then $\mathcal{E}_k(\widetilde{G}_1, \widetilde{G}_2)$ achieves almost α^*-recovery for all $\lambda \in [0,1]$. Further,

$$\mathbb{P}(|(\mathcal{B}_1 \cup \mathcal{B}_2')^c \setminus \mathsf{dom}(\widehat{\mu}_k)| = o(n)) = 1 - o(1). \quad (1)$$

3.2. Results on the strong adversary

First, we show that a strong adversary can modify two graphs so that no estimator can output a precise matching.

Theorem 3.3. *There exists a strong adversary A^* such that given any $(G_1, G_2, \gamma, \lambda)$, its output $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ satisfies the following: If an estimator $\mathcal{E}(\widetilde{G}_1, \widetilde{G}_2)$ is invariant with respect to reordering the node labels of both graphs with the same permutation, and returns a matching μ with $|\mathsf{dom}(\mu)| = \Theta(n)$, then $\mathcal{E}$ is ε^*-imprecise, where ε^* is any real number less than $\max(\lambda, 1-\lambda)\gamma$.*

Note that the estimator being invariant with respect to reordering the labels of both graphs with the same permutation is a natural requirement: it means the estimator uses only the graph structures to match them. Thus, Theorem 3.3 establishes that no such estimator can identify a set $S \subset [n]$ containing a positive fraction of nodes, such that all nodes in S are correctly matched with high probability. Despite this, we show that if the two graphs are correlated ER graphs and only one of the networks is compromised, then imprecise recovery of a positive fraction of nodes is possible under appropriate conditions. We state two achievability results below. Theorem 3.4 deals with the case when average degrees are logarithmic in n, whereas Theorem 3.5 deals with the case when p is constant.

Theorem 3.4. *Let $\lambda \in \{0,1\}$ and $\alpha \in [0,1]$. If*

$$\gamma < \frac{s\left(1-\alpha^2\right)}{4}, \tag{2}$$

then there exists a constant $C' \equiv C'(\alpha,\gamma)$ such that for all $C > C'$ and $p = C\log(n)/n$, and for all strong adversaries $\mathsf{A}(G_1, G_2, \gamma, \lambda)$ acting on two correlated ER graphs G_1 and G_2 with parameters p and s, the maximum overlap estimator achieves α-recovery.

Theorem 3.5. *Suppose p is constant, $\lambda \in \{0,1\}$ and $\alpha \in [0,1]$. If*

$$\gamma < 1 - \sqrt{1 - \frac{s^2 p(1-p)(1-\alpha^2)}{2}}, \tag{3}$$

then for all strong adversaries $\mathsf{A}(G_1, G_2, \gamma, \lambda)$ acting on two correlated ER graphs G_1 and G_2 with parameters p and s, the maximum overlap estimator achieves α-recovery.

4. Proof Outlines

Proofs for all results in Section 3 are outlined, with details deferred to the supplementary material.

4.1. The weak adversary

The performance of the k-core estimator against the weak adversary is analyzed. First, the impossibility result is proved using an indistinguishability argument.

Proof of Theorem 3.1. (i) It suffices to show that no estimator can correctly match any node i in $\mathcal{B}_1 \cup \mathcal{B}_2'$ with high probability. Fix such a node i and onsider the joint distribution of the collection $\left\{\widetilde{G}_1\{i,j\}, \widetilde{G}_2\{\pi^*(i),\pi^*(j)\}\right\}_{j\in[n], j\neq i}$. These $2(n-1)$ random variables are each distributed as $\mathsf{Bern}(ps^2)$ and mutually independent, since either $G_1\{i,j\}$ or $G_2\{\pi^*(i),\pi^*(j)\}$ is resampled because either $i \in \mathcal{B}_1$ or $\pi^*(i) \in \mathcal{B}_2$. Further, this joint distribution is the same for all nodes i in the set $\mathcal{B}_1 \cup \mathcal{B}_2'$, and so the nodes within it are statistically indistinguishable. Consequently, no estimator can match any subset $\overline{M}$ of nodes in $\mathcal{B}_1 \cup \mathcal{B}_2'$ better than random guessing. Lemma A.2 shows that the random guessing estimator is precise if and only if $\mathbb{P}\left(\overline{M} = \phi\right) = 1 - o(1)$, and the desired result follows.

(ii) Lemma A.2 implies that any estimator can at best match correctly the set $(\mathcal{B}_1 \cup \mathcal{B}_2')^c$ and at most a sublinear number of nodes in $(\mathcal{B}_1 \cup \mathcal{B}_2')$. However, $|\mathcal{B}_1 \cup \mathcal{B}_2'|/n$ converges in probability to $\gamma - \lambda(1-\lambda)\gamma^2$. This follows from Lemma A.3, where it is shown that $|\mathcal{B}_1 \cap \mathcal{B}_2'|/n$ converges in probability to $\lambda(1-\lambda)\gamma$. Since no more than a sublinear number of nodes in $\mathcal{B}_1 \cup \mathcal{B}_2'$ are correctly matched, it follows that the fraction of correctly matched nodes, α, is strictly upper bounded by $|(\mathcal{B}_1 \cup \mathcal{B}_2')^c|/n + \varepsilon$ for every $\varepsilon > 0$. We conclude that $\alpha \le 1 - \gamma + \lambda(1-\lambda)\gamma^2$, as desired.

(iii) Assume $\lambda = 1$ so that $\mathcal{B}_2' = \phi$; a similar proof works for $\lambda = 0$. With probability $1 - o(1)$:

$$\alpha^* n \overset{(a)}{\le} \mathsf{ov}(\mu, \pi^*) \le |\mathsf{dom}(\mu)| \overset{(b)}{\le} |\mathcal{B}_1^c| = \alpha^* n,$$

where (a) is true because $\mathcal{E}$ achieves α^* recovery and (b) uses (i) since $\mathcal{E}$ is a precise estimator. The above string of inequalities are thus equalities. Thus, (i) yields that $\mathsf{dom}(\mu) = \mathcal{B}_1^c$ with probability $1 - o(1)$. Since $\mathcal{E}$ achieves α^*-recovery, it follows that μ has correctly matched all the vertices in $\mathcal{B}_1^c$. We show that this is only possible when $C \ge 1/(s^2\alpha^*)$. For a graph G and vertex subset $X \subseteq V(G)$, let $G|_X$ denote the induced subgraph of G on X. Then, with probability $1 - o(1)$:

$$\begin{aligned} H_1 &:= \widetilde{G}_1|_{\mathsf{dom}(\mu)} = \widetilde{G}_1|_{\mathcal{B}_1^c} \overset{(c)}{=} G_1|_{\mathcal{B}_1^c}, \\ H_2 &:= \widetilde{G}_2|_{\pi^*(\mathsf{dom}(\mu))} = \widetilde{G}_2|_{\pi^*(\mathcal{B}_1^c)} \overset{(d)}{=} G_2|_{\pi^*(\mathcal{B}_1^c)}, \end{aligned}$$

where (c) is because no node pair in $\widetilde{G}_1|_{\mathcal{B}_1^c}$ is influenced by the adversary, and (d) is because $G_2 = \widetilde{G}_2$. Thus, H_1 and H_2 are correlated ER graphs on $\alpha^* n$ nodes with parameters p and s, and their underlying correspondence is given by $\pi^* |_{\mathsf{dom}(\mu)}$. Recovering $\pi^*|_{\mathsf{dom}(\mu)}$ is the exact graph matching problem between H_1 and H_2, which is impossible whenever $\alpha^* nps^2 < 1$, i.e. whenever $C < 1/(s^2\alpha^*)$ (Cullina & Kiyavash, 2017). □

Next, a proof sketch for Theorem 3.2 is presented.

Proof of Theorem 3.2. (i) The proof is deferred to Appendix A.1 due to space constraints.

(ii) The union bound yields for any $\delta \ge 0$,

$$\mathbb{P}\big(\mathcal{E}_k \text{ achieves almost } \alpha^*\text{-recovery}\big) \ge 1 - p_1 - p_2 - p_3,$$

where

$$p_1 = \mathbb{P}\big(\widehat{\mu}_k \neq \pi^*|_{\mathsf{core}_k(\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2)}\big), \tag{4}$$

$$p_2 = \mathbb{P}\big(\mathsf{core}_k(\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2) \neq (\mathcal{B}_1 \cup \mathcal{B}_2')^c\big), \tag{5}$$

$$p_3 = \mathbb{P}\big(\,|(\mathcal{B}_1 \cup \mathcal{B}_2')^c| < (1-\delta)\alpha^* n \tag{6}$$

From Theorem A.9 and Lemma A.10 in the proof of (i), it follows that $p_1 = o(1)$. The bulk of the analysis is to show that $p_2 = o(1)$ whenever $C > 1/(s^2\alpha^*)$. This is shown in Lemma A.11 in the supplementary material. Finally, for any $\delta > 0$, it follows from Lemma A.3 that $p_3 = o(1)$ for all $\lambda \in [0, 1]$.

When $\lambda \in \{0, 1\}$, it holds that $p_3 = 0$ even when $\delta = 0$. This is because either $\mathcal{B}_1 = \phi$ or $\mathcal{B}_2' = \phi$, and therefore $|\mathcal{B}_1 \cup \mathcal{B}_2'| = \gamma n = \alpha^* n$ in this setting. However, setting $\delta = 0$ corresponds to achieving α^*-recovery.

(iii) Let M^* denote $\mathsf{core}_k(\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2)$. The union bound yields for any $\delta \geq 0$,

$$\mathbb{P}\big(\mathcal{E}_k \text{ achieves almost } \alpha^*\text{-recovery}\big) \geq 1 - p_1 - p_4,$$

where p_1 is defined in (4), and

$$p_4 = \mathbb{P}\big(|M^*| < (1-\delta)\alpha^* n\big). \tag{7}$$

Lemma A.12 shows that $p_4 = o(1)$ for any $\delta > 0$. From part (i) of this theorem, it follows that $p_1 = o(1)$. Therefore, $\mathcal{E}_k$ achieves almost α^*-recovery. It remains to prove (1), i.e. $\mathcal{E}_k$ recovers all but a vanishing fraction of the uncorrupted nodes. Since $p_1 = o(1)$, it suffices to instead show that for any $\delta' > 0$

$$p_5 := \mathbb{P}\left(|(\mathcal{B}_1 \cup \mathcal{B}_2')^c \setminus M^*| > \delta' n\right) = 1 - o(1).$$

Indeed, denoting $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2'$, it follows that for any $\delta' > 0$, and $\varepsilon = \delta/\alpha^*$ that

$$\begin{aligned} p_5 &\leq \mathbb{P}\left(M^* \not\subseteq \mathcal{B}^c\right) + \mathbb{P}\left(\{|\mathcal{B}^c \setminus M^*| > \delta n\} \cap \{M^* \subseteq \mathcal{B}^c\}\right) \\ &\leq o(1) + \mathbb{P}\left(|\mathcal{B}^c| - |M^*| > \delta n\right) \\ &\leq o(1) + \mathbb{P}\left(|\mathcal{B}^c| > (1+\varepsilon/2)\alpha^* n\right) \\ &\qquad + \mathbb{P}\left(|M^*| < (1-\varepsilon/2)\alpha^* n\right) \\ &\overset{(a)}{=} o(1) + o(1) + o(1), \end{aligned}$$

where (a) uses both Lemma A.3 and Lemma A.12. □

Algorithm 1 Adversary A^*

```
Inputs: $G_1$, $G_2$, $\gamma$, $\lambda$
Set $\widetilde{G}_2 = G_2$
Select $\mathcal{B}_1 \subseteq V(G_1)$ uniformly at random from the set of all subsets of $V(G_1)$ of size $\gamma\lambda n$
Select $\mathcal{B}_2 \subseteq V(G_2)$ arbitrarily from the set of all subsets of $V(G_2)$ of size $\gamma(1-\lambda)n$
Select a permutation $\pi^{\mathcal{B}_1}$ uniformly at random from the set of all permutations on $\mathcal{B}_1$
Extend $\pi^{\mathcal{B}_1}$ to a permutation $\widetilde{\pi}$ on $[n]$ by setting
    $\widetilde{\pi}(i) = \begin{cases} \pi^{\mathcal{B}_1}(i), & i \in \mathcal{B}_1 \\ i, & i \notin \mathcal{B}_1 \end{cases}$    (8)
Set $\widetilde{G}_1 = G_1^{\widetilde{\pi}}$
return $\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2$
```

4.2. The strong adversary

Proof of Theorem 3.3. We may assume $\lambda \geq 1/2$ (else interchange λ and $1-\lambda$ in this proof). Consider the strong adversary A^* in Algorithm 1, which selects a random set $\mathcal{B}_1$ of vertices in G_1 and modifies the edge status of vertex pairs in G_1 so that the output graph is given by $\widetilde{G}_1 = G_1^{\widetilde{\pi}}$ for the permutation $\widetilde{\pi}$ described in (8). Since $\widetilde{\pi}(i) = i$ for all $i \notin \mathcal{B}_1$, it follows that $\widetilde{G}_1$ can be obtained from G_1 by modifying the edge status of only the node pairs in $E_{\mathcal{B}_1}$.

Let $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ denote the output of the adversary A^*. Let $\mathcal{E}$ be any estimator that is invariant with respect to reordering the node labels of both graphs with the same permutation. Let $\mu = \mathcal{E}(G_1, G_2)$ and $\widetilde{\mu} = \mathcal{E}(\widetilde{G}_1, \widetilde{G}_2)$, and denote by M and $\widetilde{M}$ respectively the domains of these matchings. Assume $|M| = \Theta(n)$. Since $\widetilde{G}_2 = G_2$, the above invariance property implies that

$$\widetilde{\mu} = \mathcal{E}(\widetilde{G}_1, \widetilde{G}_2) = \mathcal{E}(G_1^{\widetilde{\pi}}, G_2) = \mu \circ \widetilde{\pi}^{-1},$$

where $\mu \circ \widetilde{\pi}^{-1}$ is the matching on the domain

$$\widetilde{\pi}(M) := \{\widetilde{\pi}(i) \in [n] : i \in M\}$$

and "$\circ$" denotes composition of functions. It follows that $\widetilde{M} = \pi(M)$, and that the precision ρ satisfies

$$\rho(\widetilde{\mu}) = \frac{\mathsf{ov}(\widetilde{\mu}, \pi^*)}{|\widetilde{M}|} = \frac{\mathsf{ov}(\mu \circ \widetilde{\pi}^{-1}, \pi^*)}{|M|} = \frac{\mathsf{ov}(\mu, \pi^* \circ \widetilde{\pi})}{|M|}.$$

Proceed by separately analyzing the overlap on $\mathcal{B}_1$ and $\mathcal{B}_1^c$. Consider the sets

$$\begin{aligned} \mathcal{X} &:= \{i \in M \cap \mathcal{B}_1 : \mu(i) = \pi^*(\widetilde{\pi}(i))\} \\ \mathcal{Y} &:= \{i \in M \cap \mathcal{B}_1^c : \mu(i) = \pi^*(\widetilde{\pi}(i))\}, \end{aligned}$$

so that $\mathsf{ov}(\mu, \pi^* \circ \widetilde{\pi}) = |\mathcal{X}| + |\mathcal{Y}|$. Let ε^* be a real number such that $\varepsilon^* < \gamma\lambda$. Then, for any $\delta > 0$:

$$\begin{aligned} \mathbb{P}\left(\rho(\widetilde{\mu}) > 1 - \varepsilon^*\right) &= \mathbb{P}\left(\frac{|\mathcal{X}|}{|M|} + \frac{|\mathcal{Y}|}{|M|} > 1 - \varepsilon^*\right) \\ &\leq \mathbb{P}\left(\frac{|\mathcal{X}|}{|M|} > \delta\right) + \mathbb{P}\left(\frac{|\mathcal{Y}|}{|M|} > 1 - \varepsilon^* - \delta\right). \end{aligned}$$

The event in the first term is that μ and $\pi^*(\widetilde{\pi}(i))$ agree on more than $\delta|M|$ nodes in $\mathcal{B}_1$.

Note that μ and $\widetilde{\pi}$ are independent. Since $\widetilde{\pi}$ permutes nodes within $\mathcal{B}_1$ uniformly at random, it follows that

$$\mathbb{E}\left[|\mathcal{X}| \,\big|\, M, \mathcal{B}_1\right] = |M \cap \mathcal{B}_1|/|\mathcal{B}_1| \leq 1.$$

By Markov's inequality,

$$\begin{aligned}\mathbb{P}\left(\frac{|\mathcal{X}|}{|M|} > \delta\right) &\le \frac{1}{\delta}\cdot\mathbb{E}\left[\frac{|\mathcal{X}|}{|M|}\right]\\ &= \frac{1}{\delta}\cdot\mathbb{E}\left[\frac{\mathbb{E}\left[|\mathcal{X}|\middle|M,\mathcal{B}_1\right]}{|M|}\right]\\ &\le \frac{1}{\delta}\cdot\mathbb{E}\left[\frac{1}{|M|}\right] = o(1),\end{aligned}$$

Choose δ such that $\varepsilon^* + \delta < \gamma\lambda$. Then, to bound the term with $\mathcal{Y}$, consider that

$$\begin{aligned}&\mathbb{P}\left(\frac{|\mathcal{Y}|}{|M|} > 1-\varepsilon^*-\delta \,\middle|\, M\right)\\ &\qquad \le \mathbb{P}\left(\frac{|M\cap\mathcal{B}_1^c|}{|M|} > 1-\varepsilon^*-\delta \,\middle|\, M\right) \qquad (9)\end{aligned}$$

However, M and $\mathcal{B}_1$ are independent, and the set $M\cap\mathcal{B}_1^c$ is obtained by sampling $|\mathcal{B}_1^c|$ nodes from $[n]$ uniformly without replacement, where each node is considered a success if and only if it is in M. Thus, given M, it follows that $|M\cap\mathcal{B}_1^c| \sim \mathsf{HypGeom}(n, |M|, |\mathcal{B}_1^c|)$. Since $|\mathcal{B}_1^c| = (1-\gamma\lambda)n$:

$$\begin{aligned}\mathbb{E}\left[|M\cap\mathcal{B}_1^c| \;\middle|\, M\right] &= \frac{|M||\mathcal{B}_1^c|}{n} = (1-\gamma\lambda)|M|,\\ \mathrm{Var}(|M\cap\mathcal{B}_1^c| \,|\, M) &= \frac{|M||\mathcal{B}_1^c|}{n}\cdot\frac{n-|M|}{n}\cdot\frac{n-|\mathcal{B}_1^c|}{n-1}\\ &= \gamma\lambda\cdot(1-\gamma\lambda)\cdot|M|\cdot\frac{n-|M|}{n-1}\end{aligned}$$

By Chebyshev's inequality,

$$\begin{aligned}(9) &\le \mathbb{P}\left(\left|\frac{|M\cap\mathcal{B}_{1,1}^c|}{|M|} - (1-\gamma\lambda)\right| > \gamma\lambda-\varepsilon^*-\delta \,\middle|\, M\right)\\ &\le \frac{\gamma\lambda(1-\gamma\lambda)}{(\gamma\lambda-\varepsilon^*-\delta)^2}\cdot\frac{n-|M|}{n-1}\cdot\frac{1}{|M|} = o(1), \qquad (10)\end{aligned}$$

Taking the expectation with respect to M yields that $\mathbb{P}(\rho(\widetilde{\mu}) > \varepsilon^*) = o(1)$. This concludes the proof. □

Next, the proofs for Theorems 3.4 and 3.5 are presented by analyzing the maximum overlap estimator when only one network is compromised. Without loss of generality, assume that $\lambda = 1$, so that $\widetilde{G}_2 = G_2$.

For any matching μ, let $X(\mu)$ denote the number of edges in $G_1 \wedge_\mu G_2$. Similarly, let $\widetilde{X}(\mu)$ denote the number of edges in $\widetilde{G}_1 \wedge_\mu \widetilde{G}_2$. Recall that the maximum overlap matching is defined as $\widehat{\mu}_{\mathsf{MO}} \in \arg\max_\mu \widetilde{X}(\mu)$.

We may assume without loss of generality, that the latent correspondence π^* is the identity permutation id. Recall that a fixed point of a permutation π is an input i such that $\pi(i) = i$. Let $\mathcal{T}^\alpha$ (resp. $\mathcal{T}^{\le\alpha}$) denote the set of all permutations with exactly (resp. at most) αn fixed points. It is shown below that $\widehat{\mu}_{\mathsf{MO}} \notin \mathcal{T}^{\le\alpha}$. It suffices to prove:

$$\mathbb{P}\left(\widetilde{X}(\mathsf{id}) > \widetilde{X}(\pi) \text{ for all } \pi\in\mathcal{T}^{\le\alpha}\right) = 1-o(1). \qquad (11)$$

The following lemma is a useful ingredient for the proof of both Theorem 3.4 and Theorem 3.5.

Lemma 4.1. *For any permutation π, and output $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ of the strong adversary A with $\lambda = 1$, acting on two correlated ER graphs with parameters p and s with underlying correspondence* id*:*

$$\widetilde{X}(\mathsf{id}) - \widetilde{X}(\pi) > X(\mathsf{id}) - X(\pi) - Z,$$

where

$$Z := \max_{\substack{S,T\subseteq[n]\\ |S|,|T|\le\gamma n}} \sum_{\substack{i\in S\cup T\\ j\in[n]}} G_2\{i,j\}. \qquad (12)$$

Proof. For any selection $\mathcal{B}_1$ of γn nodes in G_1, recall that $E_{\mathcal{B}_1} := \left\{\{i,j\}\in\binom{[n]}{2} : i\in\mathcal{B}_1 \text{ or } j\in\mathcal{B}_1\right\}$.

The adversary selects $\mathcal{B}_1$ and sets the edge status of all node pairs in $E_{\mathcal{B}_1}$ to either 0 or 1. For any node pair $e = \{i,j\}$ and graph H, let $H(e)$ be a shorthand for $H\{i,j\}$. For any set $\mathcal{B}_1$ and any permutation π:

$$\begin{aligned}&\left(\widetilde{X}(\pi) - \widetilde{X}(\mathsf{id})\right) - \left(X(\pi) - X(\mathsf{id})\right) &(13)\\ &= \sum_{e\in E_{\mathcal{B}_1}} \left(\widetilde{G}_1(e) - G_1(e)\right)\left(G_2^\pi(e) - G_2(e)\right) &(14)\\ &\overset{(a)}{\le} \sum_{e\in E_{\mathcal{B}_1}} \mathbb{1}\{G_1(e){=}1\}\cdot\mathbb{1}\{G_2(e){=}1\}\cdot\mathbb{1}\{G_2^\pi(e){=}0\}\\ &+ \sum_{e\in E_{\mathcal{B}_1}} \mathbb{1}\{G_1(e){=}0\}\cdot\mathbb{1}\{G_2(e){=}0\}\cdot\mathbb{1}\{G_2^\pi(e){=}1\}\\ &\le \sum_{e\in E_{\mathcal{B}_1}} \left(\mathbb{1}\{G_2(e)=1\} + \mathbb{1}\{G_2^\pi(e)=1\}\right). &(15)\end{aligned}$$

Here, (a) is because each term of the sum in (14) is in the set $\{-1,0,1\}$, and equals 1 if and only if the adversary sets $\widetilde{G}_1(e) = 1 - G_1(e)$ whenever $(G_1(e), G_2(e), G_2^\pi(e))$ is either $(1,1,0)$ or $(0,0,1)$. Note that (15) is maximized when $\mathcal{B}_1$ and π are chosen to maximize $|E(G_2)\cap E_{\mathcal{B}_1}| + |E(G_2^\pi)\cap E_{\mathcal{B}_1}|$. Therefore,

$$\begin{aligned}(13) &\le \max_{\mathcal{B}_1,\pi(\mathcal{B}_1)} \sum_{e\in E_{\mathcal{B}_1}} \left(G_2(e) + G_2^\pi(e)\right)\\ &\overset{(b)}{\le} \max_{\substack{S,T\subseteq[n]\\ |S|,|T|\le\gamma n}} \sum_{\substack{i\in S\cup T\\ j\in[n]}} G_2\{i,j\},\end{aligned}$$

as desired. Here, (b) follows from the fact that $|\mathcal{B}_1| = |\pi(\mathcal{B}_1)| \le \gamma n$. This concludes the proof. □

Proof of Theorem 3.4. Let Z be as in (12). Let Δ_2 denote the maximum node degree in the graph G_2. Since $|S|+|T| \leq 2\gamma n$, it follows that $Z \leq 2\gamma n \Delta_2$. Let E denote the event

$$\mathsf{E} = \bigcup_{\pi \in \mathcal{T}^{\leq \alpha}} \left\{ \widetilde{X}(\mathsf{id}) - \widetilde{X}(\pi) < 0 \right\}. \tag{16}$$

Applying Lemma 4.1 yields for any $\varepsilon > 0$:

$$\mathbb{P}\left(\mathsf{E}\right) \leq \mathbb{P}\left(\bigcup_{\pi \in \mathcal{T}^{\leq \alpha}} \{X(\mathsf{id}) - X(\pi) < 2\gamma n \Delta_2\} \right) \leq \sum_{i=1}^{3} p_i,$$

where

$$p_1 = \mathbb{P}\left(X(\mathsf{id}) \leq (1-\varepsilon)\binom{n}{2} ps^2 \right), \tag{17}$$

$$p_2 = \mathbb{P}\left(\Delta_2 > (1+\varepsilon)nps\right), \tag{18}$$

$$p_3 = \mathbb{P}\Bigg(\bigcup_{\pi \in \mathcal{T}^{\leq \alpha}} \Bigg\{ X(\pi) \geq (1-\varepsilon)\binom{n}{2} ps^2 - 2\gamma n \left(1+\varepsilon\right) nps \Bigg\} \Bigg) \tag{19}$$

Lemmas B.1 and B.2 show that $p_1 = o(1)$ and $p_2 = o(1)$ for any $\varepsilon > 0$ and sufficiently large C. Furthermore, Lemma B.5 shows that $p_3 = o(1)$ for sufficiently small $\varepsilon > 0$ and sufficiently large $C > 0$, whenever $\gamma < s(1-\alpha^2)/4$. This requires a Chernoff argument, using bounds on the moment generating function of $X(\pi)$ obtained by analyzing the orbit decomposition of π. Thus, it follows that $\mathbb{P}\left(\mathsf{E}\right) = o(1)$, which concludes the proof. □

Proof of Theorem 3.5. Let Z be as defined in Equation (12), and let Γ denote the constant $\Gamma = \binom{\gamma n}{2} + \gamma(1-\gamma)n^2$. Since for any choice of S and T:

$$|\{\{i,j\} : i \in S \cup T, j \in [n]\}| \leq 2\Gamma,$$

it follows that $Z \leq 2\Gamma$. Let E denote the error event in (16). Consequently, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\mathsf{E}\right) \leq \mathbb{P}\left(\bigcup_{\pi \in \mathcal{T}^{\leq \alpha}} \{X(\mathsf{id}) - X(\pi) < 2\Gamma\} \right) \leq p_1 + p_4,$$

where

$$p_1 = \mathbb{P}\left(X(\mathsf{id}) \leq (1-\varepsilon)\binom{n}{2} ps^2 \right), \tag{20}$$

$$p_4 = \mathbb{P}\left(\bigcup_{\pi \in \mathcal{T}^{\leq \alpha}} \left\{ X(\pi) \geq (1-\varepsilon)\binom{n}{2} ps^2 - 2\Gamma \right\} \right). \tag{21}$$

Lemma B.1 shows that $p_1 = o(1)$ for any $\varepsilon > 0$. Furthermore, Lemma B.7, shows that $p_4 = o(1)$ whenever the condition (3) is satisfied. This uses similar techniques as the proof of Lemma B.5, although the resulting sufficient conditions for recovery are quite different. Combining, it follows that $\mathbb{P}\left(\mathsf{E}\right) = o(1)$ as desired. □

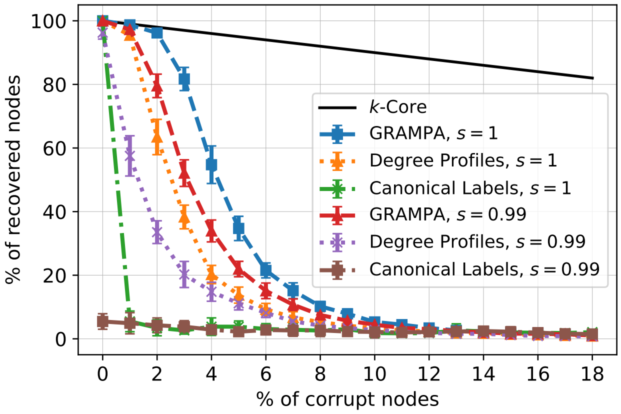

Figure 1. Weak adversary, $(n, p, \lambda) = (10^3, 0.1, 1)$

5. Discussion

Implications Theorem 3.2 provides a sufficient condition for the k-core estimator to match all and only the uncorrupt nodes. This provides a framework to detect corruption: for example, when matching PPI networks, the set of unmatched nodes can be used to identify proteins that register random interactions due to being an unknown transcription factor or being toxic to the cell. Similarly, the proof of Theorem 3.3 provides an algorithm to sanitize social networks in order to prevent de-anonymization through matching. Specifically, social network data can be modified by emulating the adversary A^* in Algorithm 1. It is then impossible for any matching algorithm to recover a positive fraction of vertices with full precision. In practice, one must be wary because theoretical analyses work with generative models that do not fully capture real-world dynamics. Even so, algorithms that perform well on such models and are further robust, are expected to perform better in practical settings.

Feasible Algorithms The maximum overlap and k-core estimators are useful to establish theoretical guarantees, but they do not run in polynomial time. It is an open question to analyze the performance of *computationally feasible* algorithms when an adversary corrupts nodes. Figure 1 compares the asymptotic guarantee of the k-core estimator against simulation results for the following estimators.

1. GRAMPA (Fan et al., 2022) uses the spectrum of the adjacency matrices to match the two graphs. The code is available in (Fan et al., 2020).

2. DEGREE PROFILING (Ding et al., 2021) assigns a signature to each node based on the degrees of its neighbors. It then matches nodes based on signature proximity. The code is available in (Ding et al., 2020).

3. CANONICAL LABELING (Dai et al., 2019) first

Algorithm 2 Adversary A_2, $\lambda = 1/2$.

```
Inputs: $G_1, G_2, \gamma, \lambda$
Initialize $\widetilde{G}_1 = G_1$ and $\widetilde{G}_2 = G_2$
Set $\mathcal{B}_1 = \{1, 2, \cdots, \frac{\gamma n}{2}\}$, $\mathcal{B}_2 = \{\frac{\gamma n}{2}+1, \cdots, \gamma n\}$
Set $\mathcal{G}_1 = \{i \in [n] : i > \gamma n, i \text{ is odd}\}$ and $\mathcal{G}_2 = \{i \in [n] : i > \gamma n, i \text{ is even}\}$
for $k$ in $\{1, 2\}$ do
    for $\{i, j\}$ such that $i \in \mathcal{B}_k$ and $j \in \mathcal{G}_k$ do
        Set $\widetilde{G}_k\{i, j\} = 1$
    end for
end for
return $\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2$
```

matches nodes with outlier degrees, and uses them as seeds to match the remaining nodes.

There is large gap between the performance of these algorithms and the asymptotic guarantees of the k-core estimator. These algorithms are not robust to the *random* noise in the setting without the adversary, since they require $s \to 1$ for good performance. Perhaps unsurprisingly, they are also not robust to the *spatial* noise induced by node corruption.

The strong adversary The maximum overlap estimator is not robust against the strong adversary when $\lambda \notin \{0, 1\}$. For simplicity, assume that $\lambda = 1/2$, and let $p = C \log n/n$ for some positive constant C. Let $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ be the output of the strong adversary A_2 acting on correlated ER graphs with parameters p and s. We show that there is a choice of A_2 (described in Algorithm 2) that can cause the maximum overlap estimator to recover none of the nodes correctly. Following Algorithm 2, the adversary selects disjoint sets $\mathcal{B}_1$ and $\mathcal{B}_2$ and forces nodes to interact such that the overlap is maximized when each node is wrongly matched (see Appendix C). It is an interesting problem to study robustness of other estimators when both graphs are corrupted.

6. Conclusion

This work studied two models for graph matching when nodes interact adversarially with their network: the framework with the strong adversary models the malicious behavior of hacked users in a social network, whereas the framework with the weak adversary models the random behavior of stochastic interactors in PPI networks.

For the weak adversary, our impossibility result states that no positive fraction of the corrupted nodes may be correctly matched. Conversely, under appropriate conditions, the k-core estimator correctly matches almost all of the uncorrupted nodes and none of the corrupted nodes. Under a further condition which is necessary, the k-core estimator also identifies and recovers all the uncorrupted nodes. In contrast, even the simpler problem of detecting corrupted nodes is impossible to solve in the setting of the strong adversary. Even so, the maximum overlap estimator successfully matches a positive fraction of nodes under appropriate conditions.

Acknowledgments

This work was supported by NSF under Grant CCF 19-00636.

Impact statement

This paper presents work whose goal is to advance the study of networks in machine learning. Theoretical analyses of various graphical models conventionally assume underlying structure, and recently there is an interest in robust algorithms for real-world networks. We believe that incorporating *node corruptions* is an important aspect of robustness, with implications to social and biological networks.

References

Acharya, J., Jain, A., Kamath, G., Suresh, A. T., and Zhang, H. Robust estimation for random graphs. In *Conference on Learning Theory*, pp. 130–166. PMLR, 2022.

Bandyopadhyay, S., Sharan, R., and Ideker, T. Systematic identification of functional orthologs based on protein network comparison. *Genome research*, 16(3):428–435, 2006.

Barak, B., Chou, C.-N., Lei, Z., Schramm, T., and Sheng, Y. (Nearly) efficient algorithms for the graph matching problem on correlated random graphs. *Advances in Neural Information Processing Systems*, 32, 2019.

Cullina, D. and Kiyavash, N. Improved achievability and converse bounds for Erdős-Rényi graph matching. *ACM SIGMETRICS performance evaluation review*, 44(1):63–72, 2016.

Cullina, D. and Kiyavash, N. Exact alignment recovery for correlated Erdős-Rényi graphs. *arXiv preprint arXiv:1711.06783*, 2017.

Cullina, D., Kiyavash, N., Mittal, P., and Poor, H. V. Partial recovery of Erdős-Rényi graph alignment via k-core alignment. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–21, 2019.

Dai, O. E., Cullina, D., Kiyavash, N., and Grossglauser, M. Analysis of a canonical labeling algorithm for the alignment of correlated Erdős-Rényi graphs. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(2):1–25, 2019.

Ding, J. and Du, H. Matching recovery threshold for correlated random graphs. *arXiv preprint arXiv:2205.14650*, 2022.

Ding, J. and Li, Z. A polynomial-time iterative algorithm for random graph matching with non-vanishing correlation. *arXiv preprint arXiv:2306.00266*, 2023.

Ding, J., Ma, Z., Wu, Y., and Xu, J. Matlab code for degree profile in graph matching. *Available at: https://github.com/xjmoffside/degree_profile*, 2020.

Ding, J., Ma, Z., Wu, Y., and Xu, J. Efficient random graph matching via degree profiles. *Probability Theory and Related Fields*, 179:29–115, 2021.

Ding, J., Fei, Y., and Wang, Y. Efficiently matching random inhomogeneous graphs via degree profiles. *arXiv preprint arXiv:2310.10441*, 2023.

Fan, Z., Mao, C., Wu, Y., and Xu, J. Matlab code for GRAMPA. *Available at: https://github.com/xjmoffside/grampa*, 2020.

Fan, Z., Mao, C., Wu, Y., and Xu, J. Spectral graph matching and regularized quadratic relaxations II: Erdős-Rényi graphs and universality. *Foundations of Computational Mathematics*, pp. 1–51, 2022.

Fionda, V. Networks in biology. In *Encyclopedia of Bioinformatics and Computational Biology*, pp. 915–921. Academic Press, Oxford, 2019. ISBN 978-0-12-811432-2.

Ganassali, L., Massoulié, L., and Lelarge, M. Impossibility of partial recovery in the graph alignment problem. In *Conference on Learning Theory*, pp. 2080–2102. PMLR, 2021.

Gaudio, J., Rácz, M. Z., and Sridhar, A. Exact community recovery in correlated stochastic block models. In *Conference on Learning Theory*, pp. 2183–2241. PMLR, 2022.

Haghighi, A., Ng, A. Y., and Manning, C. D. Robust textual inference via graph matching. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 387–394, 2005.

Hall, G. and Massoulié, L. Partial recovery in the graph alignment problem. *Operations Research*, 71(1):259–272, 2023.

Hua, Y., Ding, J., d'Orsi, T., and Steurer, D. Reaching Kesten-Stigum threshold in the stochastic block model under node corruptions. In *Conference on Learning Theory*, pp. 4044–4071. PMLR, 2023.

Kazemi, E., Hassani, H., Grossglauser, M., and Pezeshgi Modarres, H. Proper: global protein interaction network alignment through percolation matching. *BMC bioinformatics*, 17(1):1–16, 2016.

Koh, G. C., Porras, P., Aranda, B., Hermjakob, H., and Orchard, S. E. Analyzing protein–protein interaction networks. *Journal of proteome research*, 11(4):2014–2031, 2012.

Liu, A. and Moitra, A. Minimax rates for robust community detection. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 823–831. IEEE, 2022.

Łuczak, T. Size and connectivity of the k-core of a random graph. *Discrete Mathematics*, 91(1):61–68, 1991.

Mao, C., Rudelson, M., and Tikhomirov, K. Random graph matching with improved noise robustness. In *Conference on Learning Theory*, pp. 3296–3329. PMLR, 2021.

Mao, C., Rudelson, M., and Tikhomirov, K. Exact matching of random graphs with constant correlation. *Probability Theory and Related Fields*, 186(1-2):327–389, 2023a.

Mao, C., Wu, Y., Xu, J., and Yu, S. H. Random graph matching at Otter's threshold via counting chandeliers. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 1345–1356, 2023b.

Mitzenmacher, M. and Morgan, T. Reconciling graphs and sets of sets. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 33–47, 2018.

Mitzenmacher, M. and Upfal, E. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge University Press, 2017.

Narayanan, A. and Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125. IEEE, 2008.

Narayanan, A. and Shmatikov, V. De-anonymizing social networks. In *2009 30th IEEE Symposium on Security and Privacy*, pp. 173–187. IEEE, 2009.

Pedarsani, P. and Grossglauser, M. On the privacy of anonymized networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1235–1243, 2011.

Rácz, M. Z. and Sridhar, A. Correlated stochastic block models: Exact graph matching with applications to recovering communities. *Advances in Neural Information Processing Systems*, 34:22259–22273, 2021.

Rácz, M. Z. and Sridhar, A. Matching correlated inhomogeneous random graphs using the k-core estimator. *arXiv preprint arXiv:2302.05407*, 2023.

Schellewald, C. and Schnörr, C. Probabilistic subgraph matching based on convex relaxation. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 171–186. Springer, 2005.

Singh, R., Xu, J., and Berger, B. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.

Wu, Y., Xu, J., and Yu, S. H. Settling the sharp reconstruction thresholds of random graph matching. *IEEE Transactions on Information Theory*, 68(8):5391–5417, 2022.

Xu, K., Wang, L., Yu, M., Feng, Y., Song, Y., Wang, Z., and Yu, D. Cross-lingual knowledge graph alignment via graph matching neural network. *arXiv preprint arXiv:1905.11605*, 2019.

A. Proofs for the weak adversary

This section presents the proofs pertaining to the weak adversary. First, a standard concentration inequality on binomial random variables is presented. This is used heavily in the remainder of the section, often to bound vertex degrees in various graphs of interest.

Lemma A.1. *Let $X \sim \mathsf{Bin}(n,p)$. Then,*

1. *For any $\delta > 0$,*

$$\mathbb{P}\left(X \geq (1+\delta)np\right) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{np}$$

2. *For any $\delta \in (0,1)$,*

$$\mathbb{P}\left(X \leq (1-\delta)np\right) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{np}$$

Proof. The proof follows from the Chernoff bound and can be found, for example, in Theorems 4.4 and 4.5 of (Mitzenmacher & Upfal, 2017). □

Next, the lemmas used in the proof of the impossibility result, Theorem 3.1 are presented. Lemma A.2 bounds the performance of the estimator that matches vertices in $\mathcal{B}_1 \cup \mathcal{B}_2'$ by random guessing, and Lemma A.3 uses a simple concentration argument to bound the size of $\mathcal{B}_1 \cup \mathcal{B}_2'$. Armed with these, we proceed to establish that the k-core estimator is precise in Appendix A.1. It is followed by Appendix A.2, where supporting lemmas for Theorem 3.2(ii) and (iii) are respectively presented.

Lemma A.2. *Let n be a positive integer, and let p, s, γ, λ be in $[0,1]$. Let $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ be the output of the weak adversary with parameters (γ, λ, ps) acting on two correlated ER graphs G_1 and G_2 with parameters p and s, and underlying correspondence π^*. Let $\mathcal{B}_2'$ denote the pre-image of $\mathcal{B}_2$ under π^*. Let $\overline{\mu}$ be a matching output by an estimator $\overline{\mathcal{E}}$ such that its domain $\mathsf{dom}(\overline{\mu}) \subseteq \mathcal{B}_1 \cup \mathcal{B}_2'$ and its codomain is the set $\pi^*(\mathcal{B}_1 \cup \mathcal{B}_2')$. Conditioned on the domain and codomain, suppose that $\overline{\mathcal{E}}$ matches nodes in its domain by random guessing. Let $\delta > 0$.*

(i) If $\overline{\mathcal{E}}$ is precise, then

$$\mathbb{P}\left(\mathsf{dom}(\overline{\mu}) = \phi\right) = 1 - o(1).$$

(ii) If $\overline{\mathcal{E}}$ is δ-imprecise, then for any $\varepsilon > 0$:

$$\mathbb{P}\left(\mathsf{ov}(\overline{\mu}, \pi^*) > \varepsilon n\right) = o(1) \tag{22}$$

Proof. (i) Note that $\mathsf{dom}(\overline{\mu}) \subseteq \mathcal{B}_1 \cup \mathcal{B}_2'$ by definition. Since each element in $\mathsf{dom}(\overline{\mu})$ is mapped randomly to an element in $\pi^*(\mathcal{B}_1 \cup \mathcal{B}_2')$, and since the mapping is injective, it follows that the probability that all nodes in $\mathsf{dom}(\overline{\mu})$ are correctly matched is given by

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i \in \mathsf{dom}(\overline{\mu})} \{\overline{\mu}(i) = \pi^*(i)\}\right) &= \frac{1}{|\mathcal{B}_1 \cup \mathcal{B}_2'|} \times \frac{1}{|\mathcal{B}_1 \cup \mathcal{B}_2'| - 1} \times \cdots \times \frac{1}{|\mathcal{B}_1 \cup \mathcal{B}_2'| - |\mathsf{dom}(\overline{\mu})| + 1} \\ &= \frac{(|\mathcal{B}_1 \cup \mathcal{B}_2'| - |\mathsf{dom}(\overline{\mu})|)!}{|\mathcal{B}_1 \cup \mathcal{B}_2'|!}.\end{aligned}$$

Since $|\mathcal{B}_1 \cup \mathcal{B}_2'| \geq \gamma \max\left(\lambda, 1-\lambda\right) n = \Omega(n)$, it follows that the above probability is $o(1)$ if $|\mathsf{dom}(\overline{\mu})| \geq 1$, and equals 1 if and only if $|\mathsf{dom}(\overline{\mu})| = 0$. Since $\overline{\mathcal{E}}$ achieves precise recovery, the above probability must be $1 - o(1)$, which then implies $\mathbb{P}\left(\mathsf{dom}(\overline{\mu}) = \phi\right) = 1 - o(1)$.

(ii) For any node $m \in \mathsf{dom}(\overline{\mu})$, let X_m denote the indicator event $\mathbb{1}\left\{\overline{\mu}(m) = \pi^*(m)\right\}$. Notice that $\mathsf{ov}(\overline{\mu}, \pi^*) = \sum_{m \in \mathsf{dom}(\overline{\mu})} X_m$, and so it follows that

$$\mathbb{E}\left[\mathsf{ov}(\overline{\mu}, \pi^*)\right] = \sum_{m \in \mathsf{dom}(\overline{\mu})} \mathbb{E}\left[X_m\right] = \sum_{m \in \mathsf{dom}(\overline{\mu})} \mathbb{P}\left(\overline{\mu}(m) = \pi^*(m)\right) = \sum_{m \in \mathsf{dom}(\overline{\mu})} \frac{1}{|\mathsf{dom}(\overline{\mu})|} = 1,$$

$$\mathsf{Var}(\mathsf{ov}(\overline{\mu}, \pi^*)) = \mathbb{E}\left[Y^2\right] - 1 = \left(\sum_{m_1 \in \mathsf{dom}(\overline{\mu})} \sum_{m_2 \in \mathsf{dom}(\overline{\mu})} \mathbb{E}\left[X_{m_1} X_{m_2}\right] \right) - 1 \overset{(a)}{=} 1,$$

where (a) uses the fact that

$$\mathbb{E}\left[X_{m_1} X_{m_2}\right] = \mathbb{P}\left(\{\overline{\mu}(m_1) = \pi^*(m_1)\} \cap \{\overline{\mu}(m_2) = \pi^*(m_2)\}\right) = \begin{cases} \frac{1}{|\mathsf{dom}(\overline{\mu})|} \times \frac{1}{|\mathsf{dom}(\overline{\mu})| - 1}, & \text{if } m_1 \neq m_2 \\ \frac{1}{|\mathsf{dom}(\overline{\mu})|}, & \text{if } m_1 = m_2. \end{cases}$$

Thus, Chebyshev's inequality yields

$$\mathbb{P}\left(\mathsf{ov}(\overline{\mu}, \pi^*) \geq \varepsilon n\right) \leq \mathbb{P}\left(|\mathsf{ov}(\overline{\mu}, \pi^*) - 1| \geq \varepsilon n - 1\right) \leq \frac{1}{(\varepsilon n - 1)^2} = o(1),$$

which concludes the proof. □

Lemma A.3. *Let n be a positive integer, and let p, s, γ, λ be in $[0, 1]$. Let $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ be the output of the weak adversary with parameters (γ, λ, ps) acting on two correlated ER graphs G_1 and G_2 with parameters p and s, and underlying correspondence π^*. Let $\mathcal{B}_2'$ denote the pre-image of $\mathcal{B}_2$ under π^*. Then, for any $\varepsilon > 0$*

$$\mathbb{P}\left(\left| \frac{|\mathcal{B}_1 \cap \mathcal{B}_2'|}{n} - \lambda(1-\lambda)\gamma^2 \right| > \varepsilon \right) = o(1).$$

Proof. Without loss of generality, assume that $\lambda \geq 1/2$, and that $\mathcal{B}_1 = \{1, 2, \cdots, \gamma\lambda n\}$, since the number of elements in $\mathcal{B}_1 \cap \mathcal{B}_2'$ is independent of the elements in $\mathcal{B}_1$. Since π^* is independent of the sets $\mathcal{B}_1$ and $\mathcal{B}_2$, selecting $\mathcal{B}_2$ uniformly at random is equivalent to selecting $\mathcal{B}_2'$ uniformly at random. View $\mathcal{B}_2'$ as being constructed by sampling $\gamma(1-\lambda)n$ nodes from $[n]$ without replacement. A node i sampled this way is labeled a success if $i \in \mathcal{B}_1$ and a failure otherwise. The number of successes after $\gamma(1-\lambda)n$ trials is exactly $|\mathcal{B}_1 \cap \mathcal{B}_2'|$, and is described by the hypergeometric distribution $\mathsf{HypGeom}(n, \gamma(1-\lambda)n, \gamma\lambda n)$. Using standard formulas for the mean and variance of the hypergeometric distribution yields

$$\mathbb{E}\left[|\mathcal{B}_1 \cap \mathcal{B}_2'|\right] = \lambda(1-\lambda)\gamma^2 n,$$

$$\mathsf{Var}\left(|\mathcal{B}_1 \cap \mathcal{B}_2'|\right) = \lambda(1-\lambda)\gamma^2(1-\lambda\gamma)(1-(1-\lambda)\gamma) \times \frac{n^2}{n-1}.$$

Therefore, applying Chebyshev's inequality to $|\mathcal{B}_1 \cap \mathcal{B}_2'|/n$ yields that for any constant $\varepsilon > 0$:

$$\mathbb{P}\left(\left| \frac{|\mathcal{B}_1 \cap \mathcal{B}_2'|}{n} - \lambda(1-\lambda)\gamma \right| \geq \varepsilon \right) \leq \lambda(1-\lambda)\gamma^2(1-\lambda\gamma)(1-(1-\lambda)\gamma) \times \frac{1}{\varepsilon^2(n-1)} = o(1),$$

as desired. □

A.1. Proof of Theorem 3.2(i)

This subsection analyzes the k-core estimator. To show that the k-core estimator is precise, it suffices to establish that with high probability, the output of the k-core estimator $\widehat{\mu}_k$ is such that $\mathsf{dom}(\widehat{\mu}_k)$ is exactly the k-core of the true intersection graph $\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2$, and furthermore that the mapping $\widehat{\mu}_k$ agrees with π^* on its domain. Theorem A.9 and Lemma A.10 together establish this. A similar result is proved in (Cullina et al., 2019) for Erdős-Rényi graphs when no adversary is present. The techniques introduced there were extended to analyze the k-core estimator for graph matching in correlated stochastic block models (Gaudio et al., 2022) and inhomogeneous random graphs (Rácz & Sridhar, 2023). These techniques are adapted to the setting of the weak adversary below.

Definition A.4 (Weak k-core matching)**.** Let H_1 and H_2 be two graphs, and let μ^* and μ be matchings. We say that μ is a weak k-core matching of H_1 and H_2 with respect to μ^* if the average degree in $H_1 \wedge_\mu H_2$ of all the nodes $i \in M$ such that $\mu(i) \neq \mu^*(i)$ is at least k, i.e.

$$f(\mu;\mu^*, H_1, H_2, k) := \sum_{i \in \mathsf{dom}(\mu)\colon \mu(i)\neq\mu^*(i)} \mathsf{deg}_{H_1\wedge_\mu H_2}(i) \geq k \times |\{i \in \mathsf{dom}(\mu)\colon \mu(i) \neq \mu^*(i)\}| \,.$$

When the context is clear, we omit the parameters from the notation and simply use $f(\mu)$.

Definition A.5 (Maximal matching)**.** Let μ^* and μ be matchings. A matching μ is a maximal matching with respect to μ^* (or simply, a μ^*-maximal matching) if for every $i \in \mathsf{dom}(\mu^*)$, either $i \in \mathsf{dom}(\mu)$ or $\mu^*(i) \in \mathsf{range}(\mu)$.

Remark A.6. Every matching can be uniquely extended to a μ^*-maximal matching.

Denote by $\mathcal{M}(\mu^*, d)$ the set of all matchings μ which are μ^*-maximal and such that there are exactly d nodes in M for which the images under μ and μ^* disagree. Let $\mathcal{M}(\mu^*) := \bigcup_{d=0}^n \mathcal{M}(\mu^*, d)$.

Lemma A.7. *Let μ^* be a matching, and suppose that μ is a k-core matching of two graphs H_1 and H_2. Then, there exists a matching $\mu' \in \mathcal{M}(\mu^*)$ such that μ' is a weak k-core matching.*

Proof. Since μ is a k-core matching, it is also a weak k-core matching with respect to μ^*. Let μ' denote the unique extension of μ to a μ^*-maximal matching. Since the extension only involves adding elements from μ^*, it follows that

$$\begin{aligned} k\,|\{i : \mu'(i) \neq \mu^*(i)\}| &= k\,|\{i : \mu(i) \neq \mu^*(i)\}| \\ &\leq \sum_{j \in \{i : \mu(i)\neq\mu^*(i)\}} \mathsf{deg}_{H_1\wedge_\mu H_2}(j) \\ &\leq \sum_{j \in \{i : \mu'(i)\neq\mu^*(i)\}} \mathsf{deg}_{H_1\wedge_{\mu'} H_2}(j), \end{aligned}$$

and so the average degree in $H_1 \wedge_{\mu'} H_2$ of all the nodes $i \in \mathsf{dom}(\mu')$ such that $\mu'(i) \neq \mu^*(i)$ it at least k. Since μ' is a μ^*-maximal matching by construction, the result follows. □

Lemma A.7 establishes that if no weak k-core matchings exist in $\mathcal{M}(\mu^*, d)$ for any $d > 0$, then any k-core matching must agree with μ^*. The main advantage of restricting the search to $\mathcal{M}(\mu^*)$ is that there are much fewer μ^*-maximal matchings than matchings.

Lemma A.8. $|\mathcal{M}(\mu^*, d)| \leq n^{2d}/(d!)$.

Proof. Any μ^*-maximal matching can be identified by the set $\{(i, \mu(i)) : \mu(i) \neq \mu^*(i)\}$. Since the cardinality of this set is d, there are $\binom{n}{d}$ choices of i and at most $\binom{n}{d}$ choices of $\mu(i)$ that preserve the injectivity of μ. Further, there are at most $d!$ ways to match up the d nodes according to μ that preserve the injectivity of μ. It follows that

$$|\mathcal{M}(\mu^*, d)| \leq d! \left(\binom{n}{d}\right)^2 \leq d! \left(\frac{n^d}{d!}\right) = \frac{n^{2d}}{d!}.$$

□

Theorem A.9. *Let $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ be the output of the weak adversary with parameters (γ, λ, ps) acting on two correlated ER graphs G_1 and G_2 with parameters p and s, and underlying correspondence π^*. Let $\widehat{\mu}_k$ denote the matching output by the k-core estimator $\mathcal{E}_k(\widetilde{G}_1, \widetilde{G}_2)$. Then,*

$$\mathbb{P}\left(\mathsf{dom}(\widehat{\mu}_k) = \mathsf{core}_k\left(\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2\right) \text{ and } \widehat{\mu}_k = \pi^* \,|_{\mathsf{dom}(\widehat{\mu}_k)}\right) \geq 2 - \exp\left(n^2\xi\right),$$

where

$$\xi := \max_{1\leq d\leq n} \max_{(M,\mu)\in\mathcal{M}(d)} \mathbb{P}\left(f(\mu) \geq kd\right)^{1/d}.$$

Proof. Let $\mathcal{K}$ denote the set of all k-core matchings of $\widetilde{G}_1$ and $\widetilde{G}_2$, and let $\mathcal{H}$ denote the event

$$\mathcal{H} := \bigcap_{\mu \in \mathcal{K}} \bigcap_{i \in \mathsf{dom}(\mu)} \{\mu(i) = \pi^*(i)\}$$

First, since $\widehat{\mu}_k$ itself is a k-core matching, the event $\mathcal{H}$ implies that $\widehat{\mu}_k(i) = \pi^*(i)$ for all $i \in \mathsf{dom}(\widehat{\mu}_k)$. Let $M^* = \mathsf{core}_k(\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2)$. We now show the event $\mathcal{H}$ implies that $\mathsf{dom}(\widehat{\mu}_k) = M^*$.

First, note that $\pi^*|_{M^*}$ is a k-core matching. Therefore, by the maximality property in the definition of the k-core estimator, we have that $|\mathsf{dom}(\widehat{\mu}_k)| \geq |M^*|$. Therefore, to show that $\mathsf{dom}(\widehat{\mu}_k) = M^*$, it suffices to simply show that $\mathsf{dom}(\widehat{\mu}_k) \subseteq M^*$ whenever $\mathcal{H}$ occurs. Assume to the contrary that $L := \mathsf{dom}(\widehat{\mu}_k) \setminus M^*$ is non-empty. Then, on the event $\mathcal{H}$, the subgraph of $\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2$ that is induced on $M^* \cup L$ also has minimum degree k, contradicting the maximality of M^*. Therefore, $\mathsf{dom}(\widehat{\mu}_k) = M^*$ if $\mathcal{H}$ occurs. It follows that

$$\mathbb{P}\left(\mathsf{dom}(\widehat{\mu}_k) = \mathsf{core}_k\left(\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2\right) \text{ and } \widehat{\mu}_k = \pi^* \mid_{\mathsf{dom}(\widehat{\mu}_k)}\right) \geq \mathbb{P}\left(\mathcal{H}\right).$$

To prove the theorem, it suffices to show that $\mathbb{P}\left(\mathcal{H}^c\right) \leq \exp\left(n^2 \xi\right) - 1$. For any graph G, let $d_{\min}(G)$ denote its minimum degree. Indeed,

$$\begin{aligned}
\mathbb{P}\left(\mathcal{H}^c\right) &= \mathbb{P}\left(\bigcup_{\mu \in \mathcal{K}} \bigcup_{i \in \mathsf{dom}(\mu)} \{\mu(i) \neq \pi^*(i)\}\right) \\
&= \mathbb{P}\left(\bigcup_{\substack{\mu \text{ such that} \\ \exists i \in \mathsf{dom}(\mu) : \mu(i) \neq \pi^*(i)}} d_{\min}(\widetilde{G}_1 \wedge_{\mu} \widetilde{G}_2) \geq k\right) \\
&\overset{\text{(a)}}{\leq} \sum_{d=1}^{n} \mathbb{P}\left(\bigcup_{\substack{\mu \text{ such that} \\ |\{i \in \mathsf{dom}(\mu) : \mu(i) \neq \pi^*(i)\}| = d}} f(\mu) \geq kd\right) \\
&\overset{\text{(b)}}{\leq} \sum_{d=1}^{n} \mathbb{P}\left(\bigcup_{\mu \in \mathcal{M}(\pi^*, d)} f(\mu) \geq kd\right) \\
&\overset{\text{(c)}}{\leq} \sum_{d=1}^{n} |\mathcal{M}(d)| \left(\max_{1 \leq d \leq n} \max_{\mu \in \mathcal{M}(\pi^*, d)} \mathbb{P}\left(f(\mu) \geq kd\right)\right) \\
&\overset{\text{(d)}}{\leq} \sum_{d=1}^{n} \frac{(n^2 \xi)^d}{d!} \\
&\leq \exp\left(n^2 \xi\right) - 1.
\end{aligned}$$

where (a) partitions the set of all matchings based on the number of disagreements of the matching with π^* and uses a union bound, (b) follows from Lemma A.7, (c) follows from a union bound, and (d) is from the definition of ξ and an application of Lemma A.8. □

Next, we show that ξ decays rather quickly under appropriate conditions. The proof below follows (Gaudio et al., 2022) and (Rácz & Sridhar, 2023), although their focus is the stochastic block model and the inhomogeneous random graph model respectively. Below, we adapt the argument to the WCG model.

Lemma A.10. *Let $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ be the output of the weak adversary with parameters (γ, λ, ps) acting on two correlated ER graphs G_1 and G_2 with parameters p and s, and underlying correspondence π^*. Let $\mathcal{M}(\pi^*, d)$ be the set of all π^*-maximal matchings μ such that $|\{i \in \mathsf{dom}(\mu) \colon \mu(i) \neq \pi^*(i)\}| = d$.*

$$\xi = \max_{1 \leq d \leq n} \max_{\mu \in \mathcal{M}(\pi^*, d)} \mathbb{P}\left(f(\mu) \geq kd\right)^{1/d}.$$

Suppose that $p = \frac{C \log n}{n}$ for some $C > 0$. If $k \geq 13$, then $\xi = o(n^{-2})$.

Proof. For any matching $\mu \in \mathcal{M}(\pi^*, d)$, define the following sets:

$$\begin{aligned}
&\mathcal{A}(\mu) := \{(i,j) \in \mathsf{dom}(\mu) \times \mathsf{dom}(\mu) \colon \mu(i) \neq \pi^*(i)\}\,, \\
&\mathcal{T}(\mu) := \{(i,j) \in \mathcal{A}(\mu) \colon \mu(i) = \pi^*(j) \text{ and } \mu(j) = \pi^*(i)\}\,, \\
&\mathcal{N}(\mu) := \mathcal{A}(\mu) \setminus \mathcal{T}(\mu).
\end{aligned}$$

For a graph G, let $G(i,j)$ denote the (i,j)-th entry of its adjacency matrix. First, note that $\mathcal{A}(\mu) \leq d|\mathsf{dom}(\mu)| \leq dn$, and that $|\mathcal{T}(\mu)| \leq d$. This is because μ makes d errors by definition. It follows that

$$\begin{aligned}
f(\mu) &= \sum_{\substack{i \in \mathsf{dom}(\mu) \\ \mu(i) \neq \pi^*(i)}} \mathsf{deg}_{\widetilde{G}_1 \wedge_\mu \widetilde{G}_2}(i) \\
&= \sum_{(i,j) \in \mathcal{A}(\mu)} \widetilde{G}_1(i,j)\widetilde{G}_2(\mu(i),\mu(j)) \\
&= 2 \sum_{\substack{(i,j) \in \mathcal{T}(\mu) \\ i<j}} \widetilde{G}_1(i,j)\widetilde{G}_2(\mu(i),\mu(j)) + \sum_{(i,j) \in \mathcal{N}(\mu)} \widetilde{G}_1(i,j)\widetilde{G}_2(\mu(i),\mu(j)) \\
&=: 2X_{\mathcal{T}} + X_{\mathcal{N}}.
\end{aligned}$$

It is easy to see that $X_{\mathcal{T}}$ and $X_{\mathcal{N}}$ are independent, since they involve disjoint node pairs. Furthermore, for the same reason, the individual terms in $X_{\mathcal{T}}$ are also independent. More importantly,

$$\widetilde{G}_1(i,j)\widetilde{G}_2(\mu(i),\mu(j)) \sim \begin{cases} \mathsf{Bern}(p^2 s^2), & i \in \mathcal{B}_1 \cup \mathcal{B}_2' \text{ or } j \in \mathcal{B}_1 \cup \mathcal{B}_2' \\ \mathsf{Bern}(ps^2), & \text{otherwise} \end{cases}.$$

Therefore, it follows that

$$X_{\mathcal{T}} \preceq \mathsf{Bin}(|\mathcal{T}(\mu)|, ps^2) \preceq \mathsf{Bin}(d, ps^2), \tag{23}$$

where $\preceq$ denotes stochastic domination of the RHS. Next, the $X_{\mathcal{N}}$ is analyzed, which is slightly more complicated because the summands may be correlated. To circumvent this, partition $\mathcal{N}(\mu)$ into $\mathcal{N}_1(\mu)$, $\mathcal{N}_2(\mu)$, and $\mathcal{N}_3(\mu)$ and define

$$X_{\mathcal{N}_j} := \sum_{\substack{(i,j) \in \mathcal{N}_j(\mu) \\ i<j}} \widetilde{G}_1(i,j)\widetilde{G}_2(\mu(i),\mu(j)), \quad j \in \{1,2,3\}.$$

It follows that $X_{\mathcal{N}} \leq 2(X_{\mathcal{N}_1} + X_{\mathcal{N}_2} + X_{\mathcal{N}_3})$. Here, the factor of 2 accounts for the restriction that $i < j$. Furthermore, it is possible to partition $\mathcal{N}$ in such a way that $X_{\mathcal{N}_1}$, $X_{\mathcal{N}_2}$ and $X_{\mathcal{N}_3}$ are mutually independent. To see this, consider two node pairs (i,j) and (a,b) in $\binom{M}{2}$. The random variables $\widetilde{G}_1(i,j)\widetilde{G}_2(\mu(i),\mu(j))$ and $\widetilde{G}_1(a,b)\widetilde{G}_2(\mu(a),\mu(b))$ are dependent if and only if one of the following two conditions hold:

$$\{\mu(i), \mu(j)\} = \{\pi^*(a), \pi^*(b)\} \tag{24}$$

$$\{\mu(a), \mu(b)\} = \{\pi^*(i), \pi^*(j)\}\,. \tag{25}$$

Consider then the dependency graph H on the node set $V(H) := \{\{i,j\} : (i,j) \in \mathcal{N}(\mu)\}$ such that two nodes in this graph $\{i,j\}$ and $\{a,b\}$ have an edge between them if and only if they satisfy (24) or (25). Since each node in H has at most 2 neighbors, it follows that H is 3-colorable. Letting $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$ be the partition corresponding to the 3 colors, it can be seen that $X_{\mathcal{N}_1}$, $X_{\mathcal{N}_2}$ and $X_{\mathcal{N}_3}$ are independent sums of Binomial random variables. Specifically, for each $(i,j) \in \mathcal{N}(\mu)$, we have that $\widetilde{G}_1(i,j)\widetilde{G}_2(\mu(i),\mu(j)) \sim \mathsf{Bern}(p^2s^2)$. It follows that for each $m \in \{1,2,3\}$:

$$X_{\mathcal{N}_m} \preceq \mathsf{Bin}(|\mathcal{N}_m(\mu)|, p^2s^2) \preceq \mathsf{Bin}(dn, p^2s^2),$$

where we have used the fact that $|\mathcal{N}_m(\mu)| \leq |\mathcal{N}(\mu)| \leq |\mathcal{A}(\mu)| \leq dn$. Finally,

$$\begin{aligned}
\mathbb{P}\left(f(\mu) \geq kd\right) &\leq \mathbb{P}\left(2X_{\mathcal{T}} + X_{\mathcal{N}} \geq kd\right) \\
&\leq \mathbb{P}\left(2\left(X_{\mathcal{T}} + X_{\mathcal{N}_1} + X_{\mathcal{N}_2} + X_{\mathcal{N}_3}\right) \geq kd\right) \\
&\leq \sum_{m=1}^{3} \mathbb{P}\left(2X_{\mathcal{T}} + 6X_{\mathcal{N}_m} \geq kd\right).
\end{aligned}$$

The Chernoff bound can then be applied to the binomial distribution. It follows that for any $\theta > 0$:

$$\begin{aligned}
\mathbb{P}\left(f(\mu) \geq kd\right) &\leq \sum_{m=1}^{3} e^{-\theta kd}\mathbb{E}[e^{2\theta X_{\mathcal{T}}}]\mathbb{E}[e^{6\theta X_{\mathcal{N}_m}}] \\
&\leq 3e^{-\theta kd}\left(1 + ps^2(e^{2\theta} - 1)\right)^d \left(1 + p^2s^2e^{6\theta} - 1\right)^{dn} \\
&\overset{(a)}{\leq} 3\exp\left(-d\left(\theta k - e^{2\theta}ps^2 - ne^{6\theta}p^2s^2\right)\right),
\end{aligned}$$

where (a) follows from the fact that $(1+x)^t \leq e^{tx}$. Finally, set $\theta = c' \log n$, and observe that when $c < 1/6$:

$$\begin{aligned}
e^{2\theta}ps^2 &= n^{2c'-1}Cs^2 \log n = o(1), \\
ne^{6\theta}p^2s^2 &= n^{6c-1}C^2s^2\left(\log n\right)^2 = o(1).
\end{aligned}$$

Finally, since $k \geq 13$, it can be ensured that $ck > 2$ by choosing $c = \frac{1}{6} - \varepsilon$ for sufficiently small ε. This yields

$$\xi \leq \mathbb{P}\left(f(\mu) \geq kd\right)^{1/d} \leq \left(3\exp\left(-d(ck\log n - o(1))\right)\right)^{1/d} = o(n^{-2}),$$

as desired. □

A.2. Supporting Lemmas for Theorem 3.2

Lemma A.11. *Let n and k be positive integers, and let p, s, γ, λ be such that $0 \leq p, s, \gamma, \lambda \leq 1$. Let $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ be the output of the weak adversary with parameters (γ, λ, ps) acting on two correlated ER graphs G_1 and G_2 with parameters p and s, and underlying correspondence π^*. Suppose that $p = C\log(n)/n$ for some positive constant C. Let $\alpha^* = 1 - \gamma + \lambda(1-\lambda)\gamma^2$. If $C > 1/\left(s^2\alpha^*\right)$, then the k-core M^* of $\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2$ with $k = \sqrt{\log n}$ satisfies*

$$\mathbb{P}\left(M^* = (\mathcal{B}_1 \cup \mathcal{B}_2')^c\right) = 1 - o(1).$$

Proof. Let the matching $\widetilde{\pi}^*$ with $\mathsf{dom}(\widetilde{\pi}^*) = (\mathcal{B}_1 \cup \mathcal{B}_2')^c$, denote the restriction of the true permutation π^* to the node set $(\mathcal{B}_1 \cup \mathcal{B}_2')^c$. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ denote the events

$$\begin{aligned}
\mathcal{H}_1 &: \bigcap_{i \in \mathcal{B}_1 \cup \mathcal{B}_2'} \left\{\mathsf{deg}_{\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2}(i) < k\right\} \\
\mathcal{H}_2 &: \bigcap_{j \notin \mathcal{B}_1 \cup \mathcal{B}_2'} \left\{\mathsf{deg}_{\widetilde{G}_1 \wedge_{\widetilde{\pi}^*} \widetilde{G}_2}(j) > k\right\}.
\end{aligned}$$

Note that the intersection graph in $\mathcal{H}_1$ is with respect to π^* whereas the intersection graph in $\mathcal{H}_2$ is with respect to $\widetilde{\pi}^*$. Let $\mathcal{H} = \mathcal{H}_1 \cap \mathcal{H}_2$. It suffices to show that $\mathbb{P}\left(\mathcal{H}^c\right) = o(1)$. This implies that with high probability, no node in $\mathcal{B}_1 \cup \mathcal{B}_2'$ has degree greater than k in the intersection graph $\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2$, and therefore cannot belong to its k-core. Furthermore, it also implies that with high probability, the subgraph of $\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2$ induced on the node set $(\mathcal{B}_1 \cup \mathcal{B}_2')^c$ has minimum degree at least k. In turn, this implies that $\mathbb{P}\left(M^* = (\mathcal{B}_1 \cup \mathcal{B}_2')^c\right) = 1 - o(1)$, as desired.

Bounding $\mathbb{P}\left(\mathcal{H}_1^c\right)$ Since the sets $\mathcal{B}_1$ and $\mathcal{B}_2'$ are selected independent of G_1 and G_2, it follows for any $i \in \mathcal{B}_1 \cup \mathcal{B}_2'$ that $\mathsf{deg}_{\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2}(i) \sim \mathsf{Bin}(n-1, p^2s^2)$. This is because for any $i \in \mathcal{B}_1 \cup \mathcal{B}_2'$ and $j \in [n]$ such that $j \neq i$, Therefore, $\widetilde{G}_1(i,j)$

and $\widetilde{G}_2(\pi^*(i), \pi^*(j))$ are independent Bernoulli random variables with mean ps. Therefore, for any $\delta > 0$:

$$
\begin{aligned}
\mathbb{P}\left(\mathcal{H}_1^c\right) &= \mathbb{P}\left(\bigcup_{i \in \mathcal{B}_1 \cup \mathcal{B}_2'} \left\{\mathsf{deg}_{\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2}(i) > (1+\delta) n p^2 s^2)\right\}\right) \\
&\leq \mathbb{P}\left(|\mathcal{B}_1 \cup \mathcal{B}_2'| > 1.01(1-\alpha^*)n\right) + \mathbb{P}\left(\bigcup_{i \in \mathcal{B}_1 \cup \mathcal{B}_2'} \left\{\mathsf{deg}_{\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2}(i) > (1+\delta) n p^2 s^2\right\} \cap \left\{|\mathcal{B}_1 \cup \mathcal{B}_2'| \leq 1.01(1-\alpha^*)n\right\}\right) \\
&\leq \mathbb{P}\left(|\mathcal{B}_1 \cup \mathcal{B}_2'| > 1.01(1-\alpha^*)n\right) + 1.01(1-\alpha^*)n \times \mathbb{P}\left(\mathsf{Bin}(n-1, p^2 s^2) > (1+\delta) n p^2 s^2\right) \\
&\overset{\text{(a)}}{\leq} o(1) + 1.01(1-\alpha^*)n \times \mathbb{P}\left(\mathsf{Bin}(n-1, p^2 s^2) > (1+\delta) n p^2 s^2)\right) \\
&\overset{\text{(b)}}{\leq} o(1) + 2n \times \mathbb{P}\left(\mathsf{Bin}(n, p^2 s^2) > (1+\delta) n p^2 s^2)\right) \\
&\overset{\text{(c)}}{\leq} o(1) + 2n \times \left(\frac{\exp(\delta)}{(1+\delta)^{1+\delta}}\right)^{np^2 s^2},
\end{aligned} \tag{26}
$$

where (a) follows from Lemma A.3 and (b) follows from $\alpha^* \geq 0$ and the stochastic domination $\mathsf{Bin}(n-1, q) \preceq \mathsf{Bin}(n, q)$ for any $q \in [0, 1]$. Finally, (c) follows from Lemma A.1. Setting $\delta = \frac{n}{(\log n)^2}$ and recalling that $p = \frac{C \log n}{n}$, it follows that $(1+\delta) n p^2 s^2 = C^2 s^2 + o(1)$. Therefore, Equation (26) can be written as

$$
\begin{aligned}
\mathbb{P}\left(\mathcal{H}_1^c\right) &\leq \mathbb{P}\left(\bigcup_{i \in \mathcal{B}_1 \cup \mathcal{B}_2'} \left\{\mathsf{deg}_{\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2}(i) > C^2 s^2 + o(1)\right\}\right) \\
&\leq o(1) + 2n \times \left(\frac{\exp\left(\frac{n}{(\log n)^2}\right)}{\left(1 + \frac{n}{(\log n)^2}\right)^{1 + \frac{n}{(\log n)^2}}}\right)^{C^2 s^2 \frac{(\log n)^2}{n}} \\
&= \begin{cases} o(1), & Cs > 1 \\ \omega(1), & Cs < 1 \end{cases}.
\end{aligned}
$$

Since $C > 1/(s^2 \alpha^*)$, it follows that $Cs > \frac{1}{s\alpha^*} \geq 1$. This is because $s\alpha^* \leq 1$. Therefore, for any $\varepsilon > 0$ and any k' such that $k' > C^2 s^2 + \varepsilon$, it is true that

$$
\mathbb{P}\left(\mathcal{H}_1^c\right) \leq \mathbb{P}\left(\bigcup_{i \in \mathcal{B}_1 \cup \mathcal{B}_2'} \left\{\mathsf{deg}_{\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2}(i) > k'\right\}\right) = o(1). \tag{27}
$$

Indeed, our choice of $k = \sqrt{\log n}$ satisfies $k > C^2 s^2 + \varepsilon$ for all sufficiently large n. Therefore, $\mathbb{P}\left(\mathcal{H}_1^c\right) = o(1)$ as desired.

Bounding $\mathbb{P}\left(\mathcal{H}_2^c\right)$ The probability that a node in $\widetilde{G}_1 \wedge_{\widetilde{\pi}^*} \widetilde{G}_2$ has degree lesser than k is computed. For any $i \in (\mathcal{B}_1 \cup \mathcal{B}_2')^c$,

$$
\mathsf{deg}_{\widetilde{G}_1 \wedge_{\widetilde{\pi}^*} \widetilde{G}_2}(i) = \sum_{\substack{j \in (\mathcal{B}_1 \cup \mathcal{B}_2')^c \\ j \neq i}} \widetilde{G}_1(i,j) \widetilde{G}_2\left(\pi^*(i), \pi^*(j)\right).
$$

Since i and j are both in $(\mathcal{B}_1 \cup \mathcal{B}_2')^c$, it follows that $\widetilde{G}_1(i,j)\widetilde{G}_2\left(\pi^*(i), \pi^*(j)\right) = G_1(i,j) G_2\left(\pi^*(i), \pi^*(j)\right) \sim \mathsf{Bern}(ps^2)$. Further, for any $j_1, j_2 \in (\mathcal{B}_1 \cup \mathcal{B}_2')^c$ such that $j_1 \neq j_2 \neq i$, it follows by independence across edges that $G_1\left(i, j_1\right) G_2\left(\pi^*(i), \pi^*(j_1)\right)$ and $G_1\left(i, j_2\right) G_2\left(\pi^*(i), \pi^*(j_2)\right)$ are independent random variables. Therefore, it follows that

$\mathsf{deg}_{G_1 \wedge_{\widetilde{\pi}^*} G_2}(i) \sim \mathsf{Bin}((1-\gamma)\,n-1, ps^2)$. We have from a union bound that for any $\delta' \in (0,1)$:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{H}_2^c\right) &= \mathbb{P}\left(\bigcup_{j \in (\mathcal{B}_1 \cup \mathcal{B}_2')^c} \left\{ \mathsf{deg}_{\widetilde{G}_1 \wedge_{\widetilde{\pi}^*} \widetilde{G}_2}(i) < (1-\delta')\,(\alpha^* n - 1)\, ps^2 \right\} \right) \\
&\leq \mathbb{P}\left(|(\mathcal{B}_1 \cup \mathcal{B}_2')| > 1.01(1-\alpha^*)n\right) \\
&\qquad + \mathbb{P}\left(\bigcup_{j \in (\mathcal{B}_1 \cup \mathcal{B}_2')^c} \left\{ \mathsf{deg}_{\widetilde{G}_1 \wedge_{\widetilde{\pi}^*} \widetilde{G}_2}(i) < (1-\delta')\,(\alpha^* n - 1)\, ps^2 \right\} \cap |(\mathcal{B}_1 \cup \mathcal{B}_2')| \leq 1.01(1-\alpha^*)n \right) \\
&\overset{\text{(d)}}{\leq} o(1) + 2n \times \mathbb{P}\left(\mathsf{Bin}\left(\alpha^* n - 1, ps^2\right) < (1-\delta')\,(\alpha^* n - 1)\, ps^2\right) \\
&\overset{\text{(e)}}{\leq} o(1) + 2n \times \left(\frac{e^{-\delta'}}{(1-\delta')^{1-\delta'}} \right)^{ps^2(\alpha^* n - 1)} \\
&\overset{\text{(f)}}{=} o(1) + 2n \times \left(\frac{e^{-\delta'}}{(1-\delta')^{1-\delta'}} \right)^{\left(Cs^2\alpha^* - \frac{Cs^2}{n}\right)\log n} \\
&\overset{\text{(g)}}{=} \begin{cases} o(1), & Cs^2\alpha^* > 1 \\ \omega(1), & \text{otherwise} \end{cases}.
\end{aligned} \tag{28}$$

Here, in (d), Lemma A.3 is used along with the fact that $\alpha^* \geq 0$. Further, in (e), Lemma A.1 is used, and in (f) we have set $p = C\log(n)/n$.

To see why (g) is true, consider the function

$$g(a,x,n) = nx^{a\log n}.$$

Notice that if $a\log(x) < -1$, then $\lim_{n\to\infty} g(a,x,n) = 0$. Equivalently, this sufficient condition requires $x < \exp\left(-1/a\right)$. Setting $a = Cs^2\alpha^* - \frac{Cs^2}{n}$, this condition reduces to

$$x < \exp\left(-\frac{1}{Cs^2\alpha^* - \frac{Cs^2}{n}} \right). \tag{29}$$

When $Cs^2\alpha^* > 1$, there is a sufficiently small constant $\varepsilon > 0$ such that for all n sufficiently large, it is true that $Cs^2\alpha^* - \frac{Cs^2}{n} > 1 + \varepsilon$. Therefore, (29) holds if $x < \exp\left(-\frac{1}{1+\varepsilon}\right)$. For our purpose, x is a function of δ':

$$x(\delta') = \frac{e^{-\delta'}}{(1-\delta')^{1-\delta'}}.$$

Clearly, $x(0) = 1$ and the right-hand side above is continuous and decreasing on $(0,1)$. Therefore, there exists $\delta'' > 0$ such that for all sufficiently small ε:

$$x(\delta'') = \exp\left(-\frac{1}{1+\varepsilon/2}\right) < \exp\left(-\frac{1}{1+\varepsilon}\right),$$

which satisfies (29) and hence implies (g). Selecting $\delta' = \delta''$, it follows from (28) that for any k' such that $k' < (1-\delta')Cs^2\alpha^*\log(n) - o(1)$:

$$\mathbb{P}\left(\mathcal{H}_2^c\right) \leq \mathbb{P}\left(\bigcup_{j \in (\mathcal{B}_1 \cup \mathcal{B}_2')^c} \left\{ \mathsf{deg}_{\widetilde{G}_1 \wedge_{\widetilde{\pi}^*} \widetilde{G}_2} < k' \right\} \right) = o(1). \tag{30}$$

Indeed, our choice of $k = \sqrt{\log n}$ satisfies this condition for sufficiently large n. Combining Equation (27) and Equation (30) via a union bound, it follows that $\mathbb{P}\left(\mathcal{H}^c\right) = o(1)$ as desired. This concludes the proof. □

Lemma A.12. *Let n and k be positive integers, and let p, s, γ, λ be real numbers such that $0 \leq p, s, \gamma, \lambda \leq 1$. Let $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ be the output of the weak adversary with parameters (γ, λ, ps) acting on two correlated ER graphs G_1 and G_2 with parameters p and s, and underlying correspondence π^*. Let $\mathcal{B}_2'$ denote the pre-image of $\mathcal{B}_2$ under π^*. Suppose that $p = \frac{C\log(n)}{n}$ for some positive constant C. Let $\alpha^* = 1 - \gamma + \lambda(1-\lambda)\gamma^2$. If $C < \frac{1}{s^2\alpha^*}$, then the k-core M^* of $\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2$ with $k = \sqrt{\log n}$ satisfies*

$$\mathbb{P}\left(|(\mathcal{B}_1 \cup \mathcal{B}_2')^c \setminus M^*| = o(n)\right) = 1 - o(1).$$

Proof. From Theorem 3.2 (i) and Theorem 3.1 (i), it follows that $\mathbb{P}\left(M^* \subseteq (\mathcal{B}_1 \cup \mathcal{B}_2')^c\right) = 1 - o(1)$. It remains to show that all but a vanishing fraction of the nodes in $(\mathcal{B}_1 \cup \mathcal{B}_2')^c$ are also in M^*. Let $\widetilde{\pi}^*$ with $\mathsf{dom}(\widetilde{\pi}^*) = (\mathcal{B}_1 \cup \mathcal{B}_2')^c$ denote the restriction of the permutation π^* to the node set $(\mathcal{B}_1 \cup \mathcal{B}_2')^c$. Let $\mathsf{core}_k(G)$ denote the k-core of a graph G. First, it is shown that none of the nodes in $\mathcal{B}_1 \cup \mathcal{B}_2'$ belong to the k-core with high probability. Since for any $i \in \mathcal{B}_1 \cup \mathcal{B}_2'$, it holds that $\mathsf{deg}_{\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2}(i) \sim \mathsf{Bin}(n-1, p^2s^2) \preceq \mathsf{Bin}(n, p^2s^2)$, and since $|\mathcal{B}_1 \cup \mathcal{B}_2'| \leq |\mathcal{B}_1| + |\mathcal{B}_2'| \leq \gamma n$, it follows by the union bound that for any $\delta > 0$

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i \in \mathcal{B}_1 \cup \mathcal{B}_2'} \left\{\mathsf{deg}_{\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2}(i) \geq (1+\delta)np^2s^2\right\}\right) &\leq \gamma n \times \mathbb{P}\left(\mathsf{Bin}(n, p^2s^2) > (1+\delta)np^2s^2\right) \\ &\leq \gamma n \times \left(\frac{\exp(\delta)}{(1+\delta)^{1+\delta}}\right)^{(n-1)p^2s^2}.\end{aligned}$$

Setting $p = C\log(n)/n$ yields

$$\mathbb{P}\left(\bigcup_{i \in \mathcal{B}_1 \cup \mathcal{B}_2'} \left\{\mathsf{deg}_{\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2}(i) \geq (1+\delta)C^2s^2\frac{\log(n)^2}{n}\right\}\right) \leq \gamma n \times \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{C^2s^2\frac{\log(n)^2}{n}}.$$

Setting $\delta = \frac{n}{C^2s^2\log(n)^{3/2}} - 1$ yields

$$\mathbb{P}\left(\bigcup_{i \in \mathcal{B}_1 \cup \mathcal{B}_2'} \left\{\mathsf{deg}_{\widetilde{G}_1 \wedge_{\pi^*} \widetilde{G}_2}(i) \geq \sqrt{\log(n)}\right\}\right) \leq \gamma n \left(\frac{\exp\left(\frac{n}{C^2s^2\log(n)^{3/2}} - 1\right)}{\left(\frac{n}{C^2s^2\log(n)^{3/2}}\right)^{\frac{n}{C^2s^2\log(n)^{3/2}}}}\right)^{C^2s^2\frac{\log(n)^2}{n}} = o(1), \qquad (31)$$

whenever $Cs > 0$. Therefore, setting $k = \sqrt{\log(n)}$ yields $\mathbb{P}(\mathcal{E}) = o(1)$, where $\mathcal{E}$ denotes the event that there exists a node in $\mathcal{B}_1 \cup \mathcal{B}_2'$ with degree larger than k. Therefore, on the event $\mathcal{E}^c$, it follows that no node in $\mathcal{B}_1 \cup \mathcal{B}_2'$ belongs to the k-core. It remains to show that all but a vanishing fraction of nodes in $(\mathcal{B}_1 \cup \mathcal{B}_2')^c$ form the k-core. Consider then the trimmed graph $H := \widetilde{G}_1 \wedge_{\widetilde{\pi}^*} \widetilde{G}_2$ on the node set $(\mathcal{B}_1 \cup \mathcal{B}_2')^c$. Thus, on the event $\mathcal{E}^c$, it is true that

$$\mathsf{core}_k(H) \subseteq M^* \subseteq (\mathcal{B}_1 \cup \mathcal{B}_2')^c. \qquad (32)$$

On the other hand, since $\mathcal{B}_1$ and $\mathcal{B}_2'$ are selected independent of the graphs G_1 and G_2, it follows that the graph H itself is an Erdős-Rényi graph on $|(\mathcal{B}_1 \cup \mathcal{B}_2')^c|$ nodes, where each edge is present with probability $Cs\frac{\log n}{n}$. The k-core of an Erdős-Rényi graph is a well studied problem. We invoke Theorem 2 of (Łuczak, 1991), restated below for convenience.

Proposition A.13. *Let $c(n) = (n-1)p$ denote the average degree in an Erdős-Rényi graph G on n nodes with edge probability p. For every $\varepsilon > 0$, there is a constant d, such that for $c = c(n) > d$ and $k = k(n) \leq c - c^{0.5+\varepsilon}$, it is true that*

$$\mathbb{P}\left(|\mathsf{core}_k(G)| \geq n\left(1 - \exp\left(-c^\varepsilon\right)\right)\right) = 1 - o(1).$$

First, some intuition is presented. Observe that the average degree c in H is given by $(|(\mathcal{B}_1 \cup \mathcal{B}_2')^c| - 1) \times \frac{Cs\log(n)}{n} = \Theta(\log(n))$ with high probability. Further, $k = \sqrt{\log(n)} \leq c - c^{0.5+\varepsilon}$ whenever $\varepsilon < 0.5$. Therefore, the graph H with

$k = \sqrt{\log n}$ satisfies the conditions of Proposition A.13 with high probability, and so the k-core contains almost all the nodes of H with high probability. Let $\alpha^* = 1 - \gamma + \lambda(1-\lambda)\gamma^2$. Formally, for any $\delta > 0$:

$$\begin{aligned}\mathbb{P}\left(|(\mathcal{B}_1 \cup \mathcal{B}_2')^c \setminus M^*| > \delta n\right) &\leq \mathbb{P}\left(\mathcal{E}\right) + \mathbb{P}\left(|(\mathcal{B}_1 \cup \mathcal{B}_2')^c| < (1-\delta)(1-\alpha^*)n\right) \\ &\quad + \mathbb{P}\left(|\mathsf{core}_{\sqrt{\log n}}(H)| < (1-\delta)(1-\alpha^*)n\left(1 - \exp\left(-c^{0.1}\right)\right)\right) \qquad (33) \\ &\leq o(1) + o(1) + o(1), \qquad (34)\end{aligned}$$

where we have showed in (31) that the first term is $o(1)$. The second term is $o(1)$ due to Lemma A.3, and the third term is $o(1)$ due to Proposition A.13. It is emphasized that (33) holds because $\mathsf{core}_k(H) \subseteq M^* \subseteq (\mathcal{B}_1 \cup \mathcal{B}_2')^c$ under the event $\mathcal{E}^c$. This concludes the proof. □

B. Proofs for the strong adversary

This section analyzes the maximum overlap estimator in the context of the strong adversary. First, a simple concentration inequality is shown for the number of edges in $G_1 \wedge_{\pi^*} G_2$. Then, the moment generating function of the random variable $X(\pi)$ is bounded using techniques introduced in (Cullina & Kiyavash, 2017) and refined in (Wu et al., 2022), among others.

Lemma B.1. *Let n be a positive integer and s be a real number such that $s \in (0,1]$. Let $C > 0$ be constant and let $p \geq C\log(n)/n$. Let G_1 and G_2 be two correlated ER graphs with parameters p and s, and underlying correspondence π^*. Let $X(\mathsf{id})$ denote the number of edges in the intersection graph $G_1 \wedge_{\pi^*} G_2$. Then, for any $\varepsilon \in (0,1)$:*

$$\mathbb{P}\left(X(\mathsf{id}) \leq (1-\varepsilon)\binom{n}{2}ps^2\right) = o(1).$$

Proof. First, notice that $G_1 \wedge_{\pi^*} G_2$ is also an Erdős-Rényi graph on n nodes, where each edge is present with probability ps^2. It follows that the number of edges in the graph $X(\mathsf{id}) \sim \mathsf{Bin}\left(\binom{n}{2}, ps^2\right)$. Thus,

$$\begin{aligned}\mathbb{P}\left(X(\mathsf{id}) \leq (1-\varepsilon)\binom{n}{2}ps^2\right) &\overset{\text{(a)}}{\leq} \left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right)^{\binom{n}{2}ps^2} \\ &\overset{\text{(b)}}{\leq} \left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right)^{\frac{Cs^2}{2}(n-1)\log(n)} \\ &= o(1),\end{aligned}$$

where (a) follows from Lemma A.1 and (b) is true because $e^{-\varepsilon} < (1-\varepsilon)^{1-\varepsilon}$ for all $\varepsilon \in (0,1)$. □

Lemma B.2. *Let $s \in (0,1]$ and $\varepsilon > 0$ be fixed, and $C > 0$ be constant. Let G_1 and G_2 denote two correlated ER graphs with parameters p and s, where $p = C\log(n)/n$. Let Δ_2 denote the maximum node degree in G_2. If $C > 3/\left(s\varepsilon^2\right)$, then*

$$\mathbb{P}\left(\Delta_2 > (1+\varepsilon)nps\right) = o(1).$$

Proof. Let $v_i \in [n]$ denote a node and let $\mathsf{deg}(v_i)$ denote its degree in G_2. By the union bound and the fact that $p = C\log(n)/n$,

$$\begin{aligned}\mathbb{P}\left(\Delta_2 > (1+\varepsilon)nps\right) &\leq \sum_{i=1}^{n}\mathbb{P}\left(\mathsf{deg}(v_i) > (1+\varepsilon)\cdot Cs\log(n)\right) \\ &= n \times \mathbb{P}\left(\mathsf{Bin}(n-1, C\log(n)/n) > (1+\varepsilon)\cdot Cs\log(n)\right) \\ &\leq n \times \mathbb{P}\left(\mathsf{Bin}(n, C\log(n)/n) > (1+\varepsilon)\cdot Cs\log(n)\right) \\ &\overset{\text{(a)}}{\leq} n \times \exp\left(-\frac{\varepsilon^2 Cs\log(n)}{3}\right) \\ &= n^{1-\frac{\varepsilon^2 Cs}{3}} \\ &\overset{\text{(b)}}{=} o(1),\end{aligned}$$

where (a) uses Theorem 4.4 of (Mitzenmacher & Upfal, 2017), and (b) is because $C > \frac{3}{s\varepsilon^2}$. □

Lemma B.3. *Let π be a permutation on $[n]$, and let G_1 and G_2 denote two correlated ER graphs with parameters p and s, and underlying correspondence π^*. Let $X(\pi)$ denote the number of edges in $G_1 \wedge_\pi G_2$. The moment generating function of $X(\pi)$ is given by*

$$\mathbb{E}\left[e^{tX(\pi)}\right] = \prod_{k=1}^{\binom{n}{2}} L_k^{N_k}$$

where N_k is the number of k-orbits in the edge decomposition of π and

$$L_k := \mathit{Tr}(L^k), \tag{35}$$

where the 2×2 matrix L is given as

$$\begin{bmatrix} 1-ps & \frac{ps}{1-ps}\left(1-p+p(1-s)^2+ps(1-s)e^t\right) \\ 1-ps & ps\left(1-s+se^t\right) \end{bmatrix}$$

Moreover, $L_k \leq L_2^{k/2}$.

Proof. For a permutation π, Let $\mathcal{O}^\pi = \bigcup_{k=1}^{\binom{n}{2}} \mathcal{O}_k^\pi$ denote the edge orbit decomposition of π, where $\mathcal{O}_k^\pi$ is the set of all k length edge orbits of π. By independence across edge orbits:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(tX(\pi)\right)\right] &= \mathbb{E}\left[\exp\left(t \sum_{\{i,j\}\in\binom{[n]}{2}} G_1\left\{i,j\right\} G_2\left\{\pi(i),\pi(j)\right\}\right)\right] \\
&= \mathbb{E}\left[\exp\left(t \sum_{O\in\mathcal{O}^\pi} \sum_{(i,j)\in O} G_1\left\{i,j\right\} G_2\left\{\pi(i),\pi(j)\right\}\right)\right] \\
&= \prod_{O\in\mathcal{O}^\pi} \mathbb{E}\left[\exp\left(tX_O(\pi)\right)\right],
\end{aligned}$$

where $X_O(\pi) = \sum_{i,j\in O} G_1\left\{i,j\right\} G_2\left\{\pi(i),\pi(j)\right\}$. Let $(a_i, b_i)_{i=1}^{k+1}$ with $(a_{k+1}, b_{k+1}) = (a_1, b_1)$ denote mutually independent random variables so that $a_i, b_i \sim \mathsf{Bern}(ps)$ and $\mathbb{P}\left(b_i = 1 \mid a_i = 1\right) = s$. Let $L_k := \mathbb{E}\left[e^{tX_O}\right]$ for any edge orbit O such that $|O| = k$. Then,

$$\begin{aligned}
L_k &= \mathbb{E}\left[\exp\left(t\sum_{j=1}^k a_j b_{j+1}\right)\right] = \mathbb{E}\left[\prod_{j=1}^k \exp\left(t\cdot a_j b_{j+1}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\prod_{j=1}^k \exp\left(t\cdot a_j b_{j+1}\right) \mid b_1,\cdots,b_k\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\prod_{j=1}^k \exp\left(t\cdot a_j b_{j+1}\right) \mid b_j, b_{j+1}\right]\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\prod_{j=1}^k \mathbb{E}\left[\exp\left(t\cdot a_j b_{j+1}\right) \mid b_j, b_{j+1}\right]\right] \\
&= \sum_{\substack{(\theta_1,\cdots,\theta_k)\in\{0,1\}^k \\ \theta_{k+1}=\theta_1}} \mathbb{P}\left((b_1,\cdots,b_k) = (\theta_1,\cdots,\theta_k)\right) \prod_{j=1}^k \mathbb{E}\left[\exp\left(t\cdot a_j b_{j+1} \mid b_j = \theta_j, b_{j+1} = \theta_{j+1}\right)\right]
\end{aligned} \tag{36}$$

where (a) follows from conditional independence of edges in G_1 given G_2. On the other hand, Equation (36) is exactly equal to

$$(36) = \mathrm{Tr}(L^k),$$

where the 2×2 matrix L is given by

$$L(\ell, m) = \mathbb{E}\left[\exp\left(t \cdot a_1 b_2\right) \mid b_1 = \ell, b_2 = m\right] \times \mathbb{P}\left(b_2 = m\right), \quad \ell, m \in \{0, 1\} \tag{37}$$

Computing the conditional expectations gives as desired,

$$\begin{bmatrix} 1-ps & \frac{ps}{1-ps}\left(1-p+p(1-s)^2+ps(1-s)e^t\right) \\ 1-ps & ps\left(1-s+se^t\right) \end{bmatrix}.$$

Finally, we show that $L_k \leq L_2^{k/2}$. Since the eigenvalues of L are given by $\frac{T \pm \sqrt{T^2-4D}}{2}$, where T and D are the trace and determinant of L respectively, it follows that

$$L_k = \mathrm{Tr}(L^k) = \left(\frac{T+\sqrt{T^2-4D}}{2}\right)^k + \left(\frac{T-\sqrt{T^2-4D}}{2}\right)^k,$$

and it follows that $L_k \leq L_2^{k/2}$. □

Corollary B.4. *Evaluating the trace of L and L^2,*

$$\begin{aligned} L_1 &= 1-ps+ps\left(1-s+se^t\right) \\ L_2 &= (1-ps)^2 + 2ps\left(1-p+p(1-s)^2+ps(1-s)e^t\right) + \left(ps\left(1-s+se^t\right)\right)^2. \end{aligned}$$

Proof. This follows from a direct computation using Equation (35). □

Lemma B.5. *Let n be a positive integer and s, γ, α be real numbers such that $s \in (0, 1]$, and $\gamma, \alpha \in [0, 1]$. Let $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ be the output of a strong adversary A with $\lambda = 1$, acting on two correlated ER graphs G_1 and G_2 with parameters p and s, and underlying correspondence π^*. Let p_3 be the event defined in (19). If $\gamma < s(1-\alpha^2)/4$, then $p_3 = o(1)$.*

Proof. By definition of p_3 and the union bound,

$$p_3 \leq \sum_{k=0}^{\alpha n} \mathbb{P}\left(\bigcup_{\pi \in \mathcal{T}^{k/n}} \left\{X(\pi) \geq (1-\varepsilon)\binom{n}{2}ps^2 - 2\gamma n\left(1+\varepsilon\right)nps\right\}\right).$$

The Chernoff bound then yields for any $t > 0$:

$$\mathbb{P}\left(X(\pi) \geq (1-\varepsilon)\binom{n}{2}ps^2 - \gamma n\left(1+\varepsilon\right)nps\right) \leq \exp\left(-t\left[(1-\varepsilon)\binom{n}{2}ps^2 - 2\gamma n\left(1+\varepsilon\right)nps\right]\right) \cdot \mathbb{E}\left[e^{tX(\pi)}\right]. \tag{38}$$

For each π, let n_k^π (resp. N_k^π) denote the number of k-orbits in the node (resp. edge) decomposition of π. The MGF of $X(\pi)$ is handled using Lemma B.3:

$$\begin{aligned} \mathbb{E}\left[e^{tX(\pi)}\right] &= \prod_{\ell=1}^{\binom{n}{2}} L_\ell^{N_\ell^\pi} \leq L_1^{\binom{n_1^\pi}{2}+n_2^\pi} \cdot \prod_{\ell=2}^{\binom{n}{2}} L_2^{\ell/2} \\ &= L_1^{\binom{n_1^\pi}{2}+n_2^\pi} \cdot L_2^{\frac{1}{2}\left(\binom{n}{2}-\binom{n_1^\pi}{2}-n_2^\pi\right)} \end{aligned} \tag{39}$$

Substituting Equation (39) in Equation (38) yields

$$\mathbb{P}\left(X(\pi) \geq (1-\varepsilon)\binom{n}{2}ps^2 - 2\left(1+\varepsilon\right)\gamma n^2 ps\right) \leq \exp\left(\zeta(\pi)\right),$$

where

$$\zeta(\pi) = -t\left((1-\varepsilon)\binom{n}{2}ps^2 - 2\left(1+\varepsilon\right)\gamma n^2 ps\right) + \frac{1}{2}\binom{n}{2}\log(L_2) + \frac{1}{2}\left(\binom{n_1^\pi}{2} + n_2^\pi\right)\log\left(\frac{L_1^2}{L_2}\right) \tag{40}$$

and

$$
\begin{aligned}
L_1 &= 1 - ps + ps\left(1 - s + se^t\right) \\
L_2 &= (1-ps)^2 + 2ps\left(1 - p + p(1-s)^2 + ps(1-s)e^t\right) + (ps\left(1 - s + se^t\right))^2.
\end{aligned}
$$

It is easy to verify that for all $p, s \in [0,1]$ and $t > 0$, the quantity $\frac{L_1^2}{L_2} \geq 1$, and so an upper bound on $\zeta(\pi)$ can be obtained by using the bound $n_2^\pi \leq n - n_1^\pi$. Let β_π denote the fraction of fixed points in the permutation π. By definition, $n_1^\pi = \beta_\pi n$. Using $p = \frac{C\log(n)}{n}$ yields

$$
\begin{aligned}
L_1 &= 1 + \frac{C \cdot (e^t - 1)s^2 \log(n)}{n}, \\
L_2 &= 1 + \frac{C^2\left(e^t - 1\right)s^2\left(2 + (e^t-1)s^2\right)(\log n)^2}{n^2}.
\end{aligned}
$$

Substituting these in Equation (40) yields

$$
\zeta(\pi) \leq T_1(\pi) + T_2(\pi) + T_3(\pi) + T_4(\pi) + T_5(\pi) + T_6(\pi),
$$

where

$$
\begin{aligned}
T_1(\pi) &= -t\left(\frac{(1-\varepsilon)s^2}{2} - 2(1+\varepsilon)\gamma s\right) \times \frac{C\log n}{n} \times n^2 = \Theta(n\log n) \\
T_2(\pi) &= \frac{\beta_\pi^2 n^2}{2}\log(L_1) = \frac{\beta_\pi^2 n^2}{2}\log\left(1 + \frac{C\cdot(e^t-1)s^2\log(n)}{n}\right) \\
&\leq \frac{\beta_\pi^2 n^2}{2} \times \frac{C(e^t-1)s^2\log n}{n} = O(n\log n) \\
T_3(\pi) &= \frac{(1-\beta_\pi^2)n^2}{4}\log(L_2) = \frac{(1-\beta_\pi^2)n^2}{4}\log\left(1 + \frac{C^2\left(e^t-1\right)s^2\left(2+(e^t-1)s^2\right)(\log n)^2}{n^2}\right) \\
&\leq \frac{(1-\beta_\pi^2)n^2}{4} \times \frac{C^2\left(e^t-1\right)s^2\left(2+(e^t-1)s^2\right)(\log n)^2}{n^2} = O((\log n)^2) \\
T_4(\pi) &= \frac{t(1-\varepsilon)ps^2}{2} \times n = \frac{C\cdot t(1-\varepsilon)s^2}{2}\log n = \Theta(\log n) \\
T_5(\pi) &= \frac{(2-3\beta_\pi)n}{2}\log(L_1) = \frac{(2-3\beta_\pi)n}{2}\log\left(1 + \frac{C(e^t-1)s^2\log(n)}{n}\right) \\
&\leq \frac{(2-3\beta_\pi)n}{2} \times \frac{C(e^t-1)s^2\log(n)}{n} = O(\log n) \\
T_6(\pi) &= -\frac{3(1-\beta_\pi)n}{4}\log\left(L_2\right) = -\frac{3(1-\beta_\pi)n}{4}\log\left(1 + \frac{C^2\left(e^t-1\right)s^2\left(2+(e^t-1)s^2\right)(\log n)^2}{n^2}\right) \\
&\leq -\frac{3(1-\beta_\pi)n}{4} \times \frac{C^2\left(e^t-1\right)s^2\left(2+(e^t-1)s^2\right)(\log n)^2}{n^2} = O\left(\frac{(\log n)^2}{n}\right).
\end{aligned}
$$

Since the dominant terms are $T_1(\pi)$ and $T_2(\pi)$, it follows that for sufficiently large n and any $t > 0$:

$$
\sum_{i=1}^{6} T_i(\pi) \leq (1+\varepsilon)\left(\frac{-t(1-\varepsilon)s^2}{2} + 2t(1+\varepsilon)\gamma s + \frac{\beta_\pi^2 s^2}{2}\left(e^t - 1\right)\right) \times Cn\log n. \tag{41}
$$

Next, the condition in the hypothesis of the theorem is invoked, i.e. $\gamma < s(1-\alpha^2)/4$. Since $\beta_\pi \leq \alpha < 1$ for all $\pi \in \mathcal{T}^{\leq\alpha}$, it follows also that $\gamma < s(1-\beta_\pi^2)/4$. Let

$$
t^* = \log\left(\frac{1}{\beta_\pi^2}\left(1 - \frac{4\gamma}{s}\right)\right). \tag{42}
$$

Note that $t^* > 0$. Also note that

$$-\frac{t^* s^2}{2} + 2t^*\gamma s + \frac{\beta_\pi^2}{2}s^2\left(e^{t^*} - 1\right) = -\frac{s}{2}\left(s\beta_\pi^2 + 4\gamma - s + (s-4\gamma)\log(s-4\gamma) - (s-4\gamma)\log\left(s\beta_\pi^2\right)\right), \tag{43}$$

and that the RHS of Equation (43) is strictly negative whenever the condition $\gamma < s(1-\beta_\pi^2)/4$ is satisfied (see Lemma B.6). Therefore, there exists a sufficient small $\varepsilon > 0$ and a sufficiently large $C > 0$ such that

$$\sum_{i=1}^{6} T_i(\pi) \le (1+\varepsilon)\left(\frac{-t^*(1-\varepsilon)s^2}{2} + 2t^*(1+\varepsilon)\gamma s + \frac{\beta_\pi^2 s^2}{2}\left(e^{t^*} - 1\right)\right) \times Cn\log n < -\frac{\beta_\pi + 1}{2} n\log(n).$$

This yields that for sufficiently large n:

$$\zeta(\pi) \le -\frac{\beta_\pi + 1}{2} n \log n.$$

Finally, this yields

$$\begin{aligned}
p_3 &\le \sum_{k=0}^{\alpha n} \sum_{\pi \in \mathcal{T}^{k/n}} \mathbb{P}\left(X(\pi) \ge (1-\varepsilon)\binom{n}{2}ps^2 - 2(1+\varepsilon)\gamma n^2 ps\right)\\
&= \sum_{k=0}^{\alpha n} n^k \exp\left(-\frac{1}{2}\left(\frac{k}{n}+1\right)n\log(n)\right)\\
&= \sum_{k=0}^{\alpha n} \exp\left(-\frac{1}{2}(n-k)\log n\right)\\
&= \sum_{k=(1-\alpha)n}^{n} \exp\left(-\frac{1}{2}k\log n\right)\\
&\le \frac{\exp\left(-\frac{1-\alpha}{2}n\log n\right)}{1-\exp\left(-\frac{1}{2}\log n\right)}\\
&= o(1)
\end{aligned}$$

This concludes the proof. □

Lemma B.6. *Suppose β_π, γ are constants such that $\beta_\pi > 0$ and $\gamma < s(1-\beta_\pi^2)/4$. Then,*

$$s\beta_\pi^2 + 4\gamma - s + (s-4\gamma)\log(s-4\gamma) - (s-4\gamma)\log\left(s\beta_\pi^2\right) > 0$$

Proof. Let $x = s\beta_\pi^2$ and $y = s - 4\gamma$. Further, let $z = y/x$. Then,

$$s\beta_\pi^2 + 4\gamma - s + (s-4\gamma)\log(s-4\gamma) - (s-4\gamma)\log\left(s\beta_\pi^2\right) = x(1 - z + z\log(z)).$$

First, note that $x > 0$. Furthermore, $z > 1$ since $\gamma < \frac{s(1-\beta_\pi^2)}{4}$ implies $x < y$. Finally, note that the function $z \mapsto 1 - z + z\log z$ is convex and minimized at $z = 1$, where it equals 0. It follows that the desired quantity is positive whenever $z > 1$. □

Lemma B.7. *Let n be a positive integer and p, s, γ, α be real numbers such that $p \in (0,1)$, $s \in (0,1]$, and $\gamma, \alpha \in [0,1]$. Let $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ be the output of a strong adversary A with $\lambda = 1$, acting on two correlated ER graphs G_1 and G_2 with parameters p and s, and underlying correspondence π^*. Let $\varepsilon \in (0,1)$ and let p_4 be the event defined in (21). If*

$$\gamma < 1 - \sqrt{1 - \frac{s^2 p(1-p)(1-\alpha^2)}{2}},$$

then $p_4 = o(1)$.

Proof. By definition of p_4 and the union bound,

$$\begin{aligned} p_4 &= \mathbb{P}\left(\bigcup_{k=0}^{\alpha n} \bigcup_{\pi \in \mathcal{T}^{k/n}} \left\{X(\pi) \geq (1-\varepsilon)\binom{n}{2}ps^2 - 2\left[\binom{\gamma n}{2} - \gamma(1-\gamma)n^2\right]\right\}\right) \\ &\leq \sum_{k=0}^{\alpha n} \mathbb{P}\left(\bigcup_{\pi \in \mathcal{T}^{k/n}} \left\{X(\pi) \geq (1-\varepsilon)\binom{n}{2}ps^2 - 2\left[\binom{\gamma n}{2} - \gamma(1-\gamma)n^2\right]\right\}\right) \end{aligned}$$

The Chernoff bound is used below to bound these terms. It follows for any $t > 0$:

$$\begin{aligned} &\mathbb{P}\left(X(\pi) \geq (1-\varepsilon)\binom{n}{2}ps^2 - 2\left[\binom{\gamma n}{2} - \gamma(1-\gamma)n^2\right]\right) \\ &\qquad \leq \exp\left(-t\left((1-\varepsilon)\binom{n}{2}ps^2 - 2\left[\binom{\gamma n}{2} - \gamma(1-\gamma)n^2\right]\right)\right) \cdot \mathbb{E}\left[e^{tX(\pi)}\right] \qquad (44) \end{aligned}$$

For each π, let n_k^π (resp. N_k^π) denote the number of k-orbits in the node (resp. edge) decomposition of π. The MGF of $X(\pi)$ is handled using Lemma B.3:

$$\mathbb{E}\left[e^{tX(\pi)}\right] = \prod_{\ell=1}^{\binom{n}{2}} L_\ell^{N_\ell^\pi} \leq L_1^{\binom{n_1^\pi}{2}+n_2^\pi} \cdot \prod_{\ell=2}^{\binom{n}{2}} L_2^{\ell/2} = L_1^{\binom{n_1^\pi}{2}+n_2^\pi} \cdot L_2^{\frac{1}{2}\left(\binom{n}{2}-\binom{n_1^\pi}{2}-n_2^\pi\right)} \qquad (45)$$

Substituting Equation (45) in Equation (44) yields

$$\mathbb{P}\left(X(\pi) \geq (1-\varepsilon)\binom{n}{2}ps^2 - 2\left[\binom{\gamma n}{2} - \gamma(1-\gamma)n^2\right]\right) \leq \exp\left(\zeta(\pi)\right),$$

where

$$\zeta(\pi) = -t\left((1-\varepsilon)\binom{n}{2}ps^2 - 2\left[\binom{\gamma n}{2} - \gamma(1-\gamma)n^2\right]\right) + \frac{1}{2}\binom{n}{2}\log(L_2) + \frac{1}{2}\left(\binom{n_1^\pi}{2} + n_2^\pi\right)\log\left(\frac{L_1^2}{L_2}\right)$$

Let β_π denote the fraction of fixed points in the permutation π. By definition, $n_1^\pi = \beta_\pi n$. Letting $\widetilde{p} = p(1-\varepsilon)$ and using the fact that $n_2^\pi \leq n - n_1^\pi$, it follows that

$$\begin{aligned} \zeta(\pi) &\leq \left(-\frac{t}{2}\left(\widetilde{p}s^2 + 2\gamma^2 - 4\gamma\right) + \frac{1}{4}\log(L_2) + \frac{\beta_\pi^2}{4}\log\left(\frac{L_1^2}{L_2}\right)\right)n^2 \qquad (46) \\ &\quad + \left(-\frac{t}{2}\left(2\gamma - \widetilde{p}s^2\right) - \frac{1}{4}\log(L_2) + \frac{2-3\beta_\pi}{4}\log\left(\frac{L_1^2}{L_2}\right)\right)n \qquad (47) \end{aligned}$$

Next, it is shown that if (3) is satisfied, then there exists $\delta > 0$ such that coefficient of the n^2 term in (46) is less than or equal to $-\delta$. To that end, notice that this coefficient is strictly negative when

$$\widetilde{p}s^2 + 2\gamma^2 - 4\gamma > \inf_{t>0}\left\{\frac{1}{t}\left(\frac{1}{2}\log(L_2) + \frac{\beta_\pi^2}{2}\log\left(\frac{L_1^2}{L_2}\right)\right)\right\} = \inf_{t>0}\left\{\log\left(L_1^{\frac{\beta_\pi^2}{t}} \cdot L_2^{\frac{1-\beta_\pi^2}{2t}}\right)\right\} \qquad (48)$$

Therefore, (48) holds if

$$\begin{aligned} \exp\left(\widetilde{p}s^2 + 2\gamma^2 - 4\gamma\right) &> \exp\left(\inf_{t>0}\log\left(L_1^{\frac{\beta_\pi^2}{t}} \cdot L_2^{\frac{1-\beta_\pi^2}{2t}}\right)\right) \qquad (49) \\ &= \inf_{t>0} L_1^{\frac{\beta_\pi^2}{t}} \cdot L_2^{\frac{1-\beta_\pi^2}{2t}} \qquad (50) \end{aligned}$$

since $L_1, L_2 \geq 1$ and therefore the objective function being infimized in Equation (49) is always non-negative. Observing that the objective function for the infimization in (50) is monotonically increasing in t, and substituting for L_1 and L_2

from Corollary B.4, it follows that

$$\begin{aligned}(50) &= \lim_{t\to 0}\left\{ \left(1-ps+ps(1-s+se^t)\right)^{\frac{\beta_\pi^2}{t}} \left((1-ps)^2+2ps\left(1-p+p(1-s)^2+ps(1-s)e^t\right)+\left(ps(1-s+se^t)\right)^2\right)^{\frac{1-\beta_\pi^2}{2t}} \right\}\\ &= \exp\left(ps^2\left(p+\beta_\pi^2-p\beta_\pi^2\right)\right),\end{aligned}$$

i.e. the condition (48) holds if $\exp\left(\widetilde{p}s^2+2\gamma^2-4\gamma\right) > \exp\left(ps^2\left(p+\beta_\pi^2-p\beta_\pi^2\right)\right)$. Rewriting this as a condition on γ yields that Equation (48) holds whenever

$$\gamma < 1-\sqrt{1-\frac{ps^2\left((1-p)(1-\beta_\pi^2)-\varepsilon\right)}{2}}. \tag{51}$$

Indeed, since $\beta_\pi \le \alpha$ for all $\pi \in \mathcal{T}^{\le\alpha}$, it follows that whenever (3) holds, ε can be chosen sufficiently small so that (51) is satisfied for all $\pi \in \mathcal{T}^{\le\alpha}$. It follows that there exists $\delta > 0$ such that the coefficient of the n^2 term in (46) is less than or equal to $-\delta$. Therefore, for sufficiently large n:

$$\zeta(\pi) \le -\frac{\delta}{2}n^2 \le -\frac{1}{2}\left(\beta_\pi+1\right)n\log(n).$$

Combining all the above,

$$\begin{aligned}p_4 &\le \sum_{k=0}^{\alpha n}\sum_{\pi\in\mathcal{T}^{k/n}} \mathbb{P}\left(X(\pi) \ge \binom{n}{2}\widetilde{p}s^2 - 2\left[\binom{\gamma n}{2} - \gamma(1-\gamma)n^2\right]\right)\\
&\le \sum_{k=0}^{\alpha n} n^k \exp\left(-\frac{1}{2}\left(\frac{k}{n}+1\right)n\log(n)\right)\\
&= \sum_{k=0}^{\alpha n}\exp\left(-\frac{1}{2}(n-k)\log n\right)\\
&= \sum_{k=(1-\alpha)n}^{n}\exp\left(-\frac{1}{2}k\log n\right)\\
&\le \frac{\exp\left(-\frac{1-\alpha}{2}n\log n\right)}{1-\exp\left(-\frac{1}{2}\log n\right)}\\
&= o(1)\end{aligned}$$

as desired. This concludes the proof. □

C. The strong adversary when both networks are compromised

In this section, the maximum overlap estimator is analyzed for the case when the adversary corrupts the same number of nodes in both graphs, i.e. $\lambda = 1/2$. The action of the adversary is described by Algorithm 2. A similar adversary can cause the maximum overlap estimator to fail horribly when $\lambda \in (0,1)$ but $\lambda \ne 1/2$. However, the setting of $\lambda = 1/2$ conveys the main ideas without complicating notation and will be the focus of the analysis.

Theorem C.1. *Let $\gamma > 0$ and let A be the strong adversary described in Algorithm 2. Let $(\mathcal{B}_1, \mathcal{B}_2, \widetilde{G}_1, \widetilde{G}_2)$ be the output of the A with $\lambda = 1/2$, acting on two correlated ER graphs G_1 and G_2 with parameters p and s, and underlying correspondence id. Let $\widehat{\pi}_{\mathsf{MO}}$ denote the matching output by the maximum overlap estimator $\mathcal{E}_{\mathsf{MO}}(\widetilde{G}_1, \widetilde{G}_2)$. Then, for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\mathsf{ov}(\widehat{\pi}_{\mathsf{MO}}, \mathsf{id}) \ge \varepsilon n\right) = o(1).$$

Proof. Assume without loss of generality that $\gamma \le 1/2$, since the same argument works by appropriately interchanging the role of γ and $1-\gamma$. For any permutation π, let $X(\pi)$ denote the number of edges in the intersection graph $G_1 \wedge_\pi G_2$.

Similarly, let $\widetilde{X}(\pi)$ denote the number of edges in the intersection graph $\widetilde{G}_1 \wedge_\pi \widetilde{G}_2$. For $j \in \{1,2\}$, let $E(\widetilde{G}_j)$ denote the set of edges present in $\widetilde{G}_j$, and furthermore, let $\widetilde{E}(\widetilde{G}_j) \subseteq E(\widetilde{G}_j)$ denote the edges added by the adversary in Algorithm 2.

Let $\widetilde{\pi}$ be a matching with $\mathsf{dom}(\widetilde{\pi}) = [n]$ such that

$$\widetilde{\pi}(i) \in \begin{cases} \mathcal{B}_2, & i \in \mathcal{B}_1 \\ \mathcal{B}_1, & i \in \mathcal{B}_2 \\ \mathcal{G}_2, & i \in \mathcal{G}_1 \\ \mathcal{G}_1, & i \in \mathcal{G}_2 \end{cases},$$

where $\mathcal{G}_1$, $\mathcal{G}_2$ are as defined in Algorithm 2. Since $|\mathcal{B}_1| = |\mathcal{B}_2|$ and $|\mathcal{G}_1| = |\mathcal{G}_2|$, such a matching exists. Note that $\mathsf{ov}(\widetilde{\pi}, \mathsf{id}) = 0$. First, a lower bound on $\widetilde{X}(\widetilde{\pi})$ is computed.

$$\begin{aligned} \widetilde{X}(\widetilde{\pi}) &= \left| E\left(\widetilde{G}_1 \wedge_{\widetilde{\pi}} \widetilde{G}_2\right)\right| \\ &\geq \left| \widetilde{E}\left(\widetilde{G}_1 \wedge_{\widetilde{\pi}} \widetilde{G}_2\right)\right| \\ &\overset{\text{(a)}}{=} |\mathcal{B}_1| \times |\mathcal{G}_1| \\ &= \frac{\gamma n}{2} \times \frac{(1-\gamma)n}{2} \\ &= \frac{\gamma(1-\gamma)}{4} \times n^2, \end{aligned}$$

where (a) follows from the fact that each node in $\mathcal{B}_1$ connects to every node in $\mathcal{G}_1$ in the graph $\widetilde{G}_1 \wedge_{\widetilde{\pi}} \widetilde{G}_2$.

Thus, in order to complete the proof, it suffices to show that

$$\mathbb{P}\left(\bigcup_{\varepsilon \in (0,1]} \bigcup_{\substack{\pi \\ \mathsf{ov}(\pi, \mathsf{id}) = \varepsilon n}} \widetilde{X}(\pi) > \frac{\gamma(1-\gamma)}{4} \times n^2 \right) = o(1). \tag{52}$$

A union bound argument is presented below to show (52). Before proceeding, let us collect some observations about the degrees of nodes in $\widetilde{G}_1$ and $\widetilde{G}_2$.

Remark C.2. Let $\mathcal{B}_1, \mathcal{B}_2, \mathcal{G}_1, \mathcal{G}_2$ be as defined in Algorithm 2. Since the adversary only adds edges and does not remove any edges,

$$\mathsf{deg}_{\widetilde{G}_1}(i) \leq \begin{cases} \mathsf{deg}_{G_1}(i) + \frac{(1-\gamma)n}{2}, & i \in \mathcal{B}_1 \\ \mathsf{deg}_{G_1}(i) + \frac{\gamma n}{2}, & i \in \mathcal{G}_1 \\ \mathsf{deg}_{G_1}(i), & i \in \mathcal{B}_2 \cup \mathcal{G}_2 \end{cases}$$

and

$$\mathsf{deg}_{\widetilde{G}_2}(j) \leq \begin{cases} \mathsf{deg}_{G_2}(j) + \frac{(1-\gamma)n}{2}, & j \in \mathcal{B}_2 \\ \mathsf{deg}_{G_2}(j) + \frac{\gamma n}{2}, & j \in \mathcal{G}_2 \\ \mathsf{deg}_{G_2}(j), & j \in \mathcal{B}_1 \cup \mathcal{G}_1 \end{cases} \tag{53}$$

Continuing, let $([n], \pi')$ be a matching such that $\mathsf{ov}(\pi', \mathsf{id}) = \varepsilon n$. Then,

$$\widetilde{X}(\pi') = \frac{1}{2} \sum_{i \in [n]} \mathsf{deg}_{\widetilde{G}_1 \wedge_{\pi'} \widetilde{G}_2}(i) \leq \frac{1}{2} \sum_{i \in [n]} \min\left(\mathsf{deg}_{\widetilde{G}_1}(i), \mathsf{deg}_{\widetilde{G}_2}(\pi'(i))\right) = \frac{1}{2}\left(Z_1 + Z_2\right), \tag{54}$$

where

$$Z_1 := \sum_{\substack{i\in[n]\\ \pi'(i)=i}} \min\left(\mathsf{deg}_{\widetilde{G}_1}(i), \mathsf{deg}_{\widetilde{G}_2}(i)\right),$$

$$Z_2 := \sum_{\substack{i\in[n]\\ \pi'(i)\neq i}} \min\left(\mathsf{deg}_{\widetilde{G}_1}(i), \mathsf{deg}_{\widetilde{G}_2}(\pi'(i))\right)$$

First, the term Z_1 is analyzed. Clearly, for any $i \in [n]$, Remark C.2 yields

$$\min\left(\mathsf{deg}_{\widetilde{G}_1}(i), \mathsf{deg}_{\widetilde{G}_2}(i)\right) \leq \min\left(\mathsf{deg}_{G_1}(i), \mathsf{deg}_{G_2}(i)\right),$$

and so it follows that

$$Z_1 \leq \varepsilon n \times \max_{i\in[n]}\left(\min\left(\mathsf{deg}_{G_1}(i), \mathsf{deg}_{G_2}(i)\right)\right) \leq n \times \max_{i\in[n]} \mathsf{deg}_{G_1}(i). \tag{55}$$

Next, the term Z_2 is analyzed. Notice that it can be upper bounded as

$$Z_2 \leq Z_{2,1} + Z_{2,2} + Z_{2,3} + Z_{2,4}, \tag{56}$$

where each term is defined and then bounded using Remark C.2 as

$$Z_{2,1} := \sum_{\substack{i\in[n]\\ i\in\mathcal{B}_2\cup\mathcal{G}_2}} \min\left(\mathsf{deg}_{\widetilde{G}_1}(i), \mathsf{deg}_{\widetilde{G}_2}(\pi'(i))\right) \leq |\mathcal{B}_2 \cup \mathcal{G}_2| \max_{i\in[n]}\left(\mathsf{deg}_{G_1}(i)\right) \leq n \max_{i\in[n]}\left(\mathsf{deg}_{G_1}(i)\right), \tag{57}$$

$$Z_{2,2} := \sum_{\substack{i\in[n]\\ \pi'(i)\in\mathcal{B}_1\cup\mathcal{G}_1}} \min\left(\mathsf{deg}_{\widetilde{G}_1}(i), \mathsf{deg}_{\widetilde{G}_2}(\pi'(i))\right) \leq |\mathcal{B}_1 \cup \mathcal{G}_1| \max_{\pi'(i)\in[n]}\left(\mathsf{deg}_{G_2}(\pi'(i))\right) \leq n \max_{i\in[n]}\left(\mathsf{deg}_{G_2}(i)\right), \tag{58}$$

$$Z_{2,3} := \sum_{\substack{i\in[n]\\ i\in\mathcal{B}_1\\ \pi'(i)\in\mathcal{B}_2\cup\mathcal{G}_2}} \min\left(\mathsf{deg}_{\widetilde{G}_1}(i), \mathsf{deg}_{\widetilde{G}_2}(\pi'(i))\right) \leq |\{i : i\in\mathcal{B}_1, \pi'(i)\in\mathcal{B}_2\cup\mathcal{G}_2\}| \left[\frac{(1-\gamma)n}{2} + \max_{i\in[n]}\left(\mathsf{deg}_{G_1}(i)\right)\right], \tag{59}$$

$$Z_{2,4} := \sum_{\substack{i\in[n]\\ i\in\mathcal{G}_1\\ \pi'(i)\in\mathcal{B}_2\cup\mathcal{G}_2}} \min\left(\mathsf{deg}_{\widetilde{G}_1}(i), \mathsf{deg}_{\widetilde{G}_2}(\pi'(i))\right) \leq |\{i : i\in\mathcal{G}_1, \pi'(i)\in\mathcal{B}_2\cup\mathcal{G}_2\}| \left[\frac{\gamma n}{2} + \max_{i\in[n]}\left(\mathsf{deg}_{G_1}(i)\right)\right]. \tag{60}$$

Let $\varepsilon_1(\pi')$ denote the fraction of nodes i in $\mathcal{B}_1$ such that $\pi'(i) \in \mathcal{B}_2 \cup \mathcal{G}_2$. Similarly, let $\delta_1(\pi')$ denote the fraction of nodes j in $\mathcal{G}_1$ such that $\pi'(j) \in \mathcal{B}_2 \cup \mathcal{G}_2$. Therefore, (59) and (60) can be written as

$$Z_{2,3} \leq \varepsilon_1(\pi')|\mathcal{B}_1| \left(\frac{(1-\gamma)n}{2} + \max_{i\in[n]}\left(\mathsf{deg}_{G_1}(i)\right)\right) \leq \varepsilon_1(\pi')\frac{\gamma(1-\gamma)}{4}n^2 + n \max_{i\in[n]}\left(\mathsf{deg}_{G_1}(i)\right), \tag{61}$$

$$Z_{2,4} \leq \delta_1(\pi')|\mathcal{G}_1| \left(\frac{\gamma n}{2} + \max_{i\in[n]}\left(\mathsf{deg}_{G_1}(i)\right)\right) \leq \delta_1(\pi')\frac{\gamma(1-\gamma)}{4}n^2 + n \max_{i\in[n]}\left(\mathsf{deg}_{G_2}(i)\right). \tag{62}$$

Therefore, combining (56)-(62) yields

$$Z_2 \leq 2n \max_{i\in[n]}\left(\mathsf{deg}_{G_1}(i)\right) + 2n \max_{i\in[n]}\left(\mathsf{deg}_{G_2}(i)\right) + \left(\varepsilon_1(\pi') + \delta_1(\pi')\right)\frac{\gamma(1-\gamma)}{4} \times n^2 \tag{63}$$

Substituting (55) and (63) in (54),

$$\widetilde{X}(\pi') \leq \frac{1}{2}\left(Z_1 + Z_2\right) \leq \frac{1}{2}\left(3n \max_{i\in[n]} \mathsf{deg}_{G_1}(i) + 2n \max_{i\in[n]}\left(\mathsf{deg}_{G_2}(i)\right) + \left(\varepsilon_1(\pi') + \delta_1(\pi')\right)\frac{\gamma(1-\gamma)}{4} \times n^2\right). \tag{64}$$

The following claim is made next.

Claim C.3. *Let $\varepsilon > 0$ and π' be any permutation such that $\mathsf{ov}(\pi', \mathsf{id}) \geq \varepsilon n$. Let ε_1 and δ_1 be as defined above. Then, $\varepsilon_1(\pi') + \delta_1(\pi') < 2$.*

Proof. The proof of the claim relies on a simple observation: If node i is not a fixed point of π', then neither is the node $\pi'(i)$. Specifically, for each node $i \in \mathcal{B}_1 \cup \mathcal{G}_1$ such that $\pi'(i) \in \mathcal{B}_2 \cup \mathcal{G}_2$, there exists a unique node $k = \pi'(i)$ in $\mathcal{B}_2 \cup \mathcal{G}_2$ such that $\pi'(k) \neq k$. This is true simply because π' is a permutation and therefore bijective.

Assume to the contrary that $\varepsilon_1(\pi') + \delta_1(\pi') = 2$. Since $\varepsilon_1(\pi')$ and $\delta_1(\pi')$ are fractions, our assumption also implies that $\varepsilon(\pi') = 1$ and $\delta_1(\pi') = 1$. Since the number of wrongly matched nodes in $\mathcal{B}_1$ (resp. $\mathcal{G}_1$) is at least as large as the number of nodes in $\mathcal{B}_1$ (resp. $\mathcal{G}_1$) that are mapped to $\mathcal{B}_2 \cup \mathcal{G}_2$, it follows that

$$\begin{aligned}|\{i \in [n] : \pi'(i) \neq i\}| &\overset{\text{(a)}}{\geq} 2\left(\varepsilon_1(\pi')|\mathcal{B}_1| + \delta_1(\pi')|\mathcal{G}_1|\right) \\ &= 2\left(|\mathcal{B}_1| + |\mathcal{G}_1|\right) \\ &= n,\end{aligned}$$

which contradicts the fact that $\mathsf{ov}(\pi', \mathsf{id}) \geq \varepsilon n$. Here, the factor of 2 in (a) is due to the aforementioned observation. It is concluded that $\varepsilon_1(\pi') + \delta_1(\pi') < 2$. □

Claim C.3 confirms the existence of a small positive constant ε' such that $\varepsilon_1(\pi') + \delta_1(\pi') \leq 2 - \varepsilon'$. Combining this with (64) and using the fact that $2n \leq 3n$ for all n yields

$$\widetilde{X}(\pi') \leq \frac{3n}{2}\left(\max_{i\in[n]} \mathsf{deg}_{G_1}(i) + \max_{i\in[n]} \mathsf{deg}_{G_2}(i)\right) + (1 - \varepsilon'/2) \times \left(\frac{\gamma(1-\gamma)}{4} \times n^2\right). \tag{65}$$

Therefore,

$$\begin{aligned}\mathbb{P}\left(\widetilde{X}(\pi') > \frac{\gamma(1-\gamma)}{4} \times n^2\right) &\leq \mathbb{P}\left(\frac{3n}{2}\left(\max_{i\in[n]} \mathsf{deg}_{G_1}(i) + \max_{i\in[n]} \mathsf{deg}_{G_2}(i)\right) > \frac{\varepsilon'}{2} \times \frac{\gamma(1-\gamma)}{4} \times n^2\right) \\ &= \mathbb{P}\left(\max_{i\in[n]} \mathsf{deg}_{G_1}(i) + \max_{i\in[n]} \mathsf{deg}_{G_2}(i) > \frac{\gamma(1-\gamma)}{12} \times \varepsilon' n\right) \\ &\leq \mathbb{P}\left(\max_{i\in[n]} \mathsf{deg}_{G_1}(i) > \frac{\gamma(1-\gamma)}{24} \times \varepsilon' n\right) + \mathbb{P}\left(\max_{i\in[n]} \mathsf{deg}_{G_2}(i) > \frac{\gamma(1-\gamma)}{24} \times \varepsilon' n\right) \\ &\overset{\text{(a)}}{\leq} 2n \times \mathbb{P}\left(\mathsf{Bin}(n, ps^2) > \frac{\gamma(1-\gamma)}{24} \times \varepsilon' n\right), \\ &\overset{\text{(b)}}{\leq} 2n \times \mathbb{P}\left(\mathsf{Bin}(n, ps^2) > n\right)\end{aligned}$$

where (a) uses a union bound, the stochastic dominance of $\mathsf{Bin}(n, ps^2)$ over $\mathsf{Bin}(n-1, ps^2)$, and (b) uses the fact that $\varepsilon'\gamma(1-\gamma)/24 < 1$. We use the tail bound on the Binomial distribution (Mitzenmacher & Upfal, 2017)

$$\mathbb{P}\left(\mathsf{Bin}(n, ps^2) > (1+\delta)nps^2\right) \leq \exp\left(-nps^2\delta^2/3\right),$$

with $\delta = \frac{n}{Cs^2\log(n)} - 1$ to get

$$\mathbb{P}\left(\widetilde{X}(\pi') > \frac{\gamma(1-\gamma)}{4} \times n^2\right) \leq 2n \times \mathbb{P}\left(\mathsf{Bin}(n, ps^2) > n\right) \leq 2n \times \exp\left(-\frac{n^2}{3Cs^2\log(n)}\right).$$

It remains to perform a union bound over all π' such that $\mathsf{ov}(\pi', \mathsf{id}) > \varepsilon n$. However, the number of such permutations is trivially upper bounded by $n! < \exp\left(n^{1.1}\log(n)\right)$ for all sufficiently large n. Therefore,

$$\mathbb{P}\left(\bigcup_{\varepsilon\in(0,1]} \bigcup_{\substack{\pi \\ \mathsf{ov}(\pi,\mathsf{id})=\varepsilon n}} \widetilde{X}(\pi) > \frac{\gamma(1-\gamma)}{4} \times n^2\right) \leq 2n\exp\left(n^{1.1}\log(n)\right) \times \exp\left(-\frac{n^2}{3Cs^2\log(n)}\right) = o(1),$$

as desired. This completes the proof. □