Perfect Recovery for Random Geometric Graph Matching with Shallow Graph Neural Networks

Suqi Liu
Harvard University

Morgane Austern
Harvard University

Abstract

We study the graph matching problem in the presence of vertex feature information using shallow graph neural networks. Specifically, given two graphs that are independent perturbations of a single random geometric graph with sparse binary features, the task is to recover an unknown one-to-one mapping between the vertices of the two graphs. We show under certain conditions on the sparsity and noise level of the feature vectors, a carefully designed two-layer graph neural network can, with high probability, recover the correct mapping between the vertices with the help of the graph structure. Additionally, we prove that our condition on the noise parameter is tight up to logarithmic factors. Finally, we compare the performance of the graph neural network to directly solving an assignment problem using the noisy vertex features and demonstrate that when the noise level is at least constant, this direct matching fails to achieve perfect recovery, whereas the graph neural network can tolerate noise levels growing as fast as a power of the size of the graph. Our theoretical findings are further supported by numerical studies as well as real-world data experiments.

1 INTRODUCTION

Graph neural networks (GNNs) (Kipf and Welling, 2016) have seen broad application in many important domains involving graph-structured data since their inception, including social networks (Hamilton et al., 2017), computational biology (Fan et al., 2019), chemistry (Gilmer et al., 2017), and knowledge graphs (Schlichtkrull et al., 2018). A standard graph neural network model, also known as a message passing neural network, consists of traditional multilayer perceptrons (MLPs) injected with a message passing step that aggregates the hidden representation from neighboring vertices in the graph. Although the message passing idea seems rather simple, GNNs have achieved wide success in various tasks, including node classification, link prediction, and learning graph properties (see Zhou et al. (2020) for a recent survey and references therein).

Despite their wide popularity and dominant performance in graph-based tasks, the theoretical understanding of GNNs is only emerging in recent years (Jegelka, 2022). Most of the works focus on traditional learning theory aspects such as representation power (Loukas, 2020) and generalization (Scarselli et al., 2018). A few of them touch on more classic graph-theoretic problems such as graph isomorphism (Morris et al., 2019) and graph properties (Garg et al., 2020). Recently there has been a growing literature studying GNNs using graphons (Ruiz et al., 2021, 2023; Chung et al., 2024). However, there is still a divide between the theory of GNNs and how they are used in the real world. The benefits and limitations of GNNs in many scenarios to which they are commonly applied remain enigmatic. Our goal is to expand the theoretical understanding by investigating a common use case of GNNs.

To this end, we focus on the problem of aligning two geometric graphs together with noisy observations of their vertex features. This simplified setup resembles several practical situations. For example, suppose that we have access to two different social networks that are actually built around the same group of people, together with inaccurate features of each node, while the node correspondence is not known to us (think of Instagram and Facebook). The task is to recover the underlying node correspondence from the networks and node features. We remark that this seemingly simplistic graph alignment task has numerous instan-

Proceedings of the 28^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

tiations, such as cross-lingual knowledge graph alignment (Wang et al., 2018) and molecular network comparison (Sharan and Ideker, 2006). Traditional methods include structure-based and iteration-based approaches (see (Zeng et al., 2021, Section 3) for a comprehensive survey). More recently, entity alignment based on representation learning, including the application of GNNs, has become mainstream (Zeng et al., 2021).

The purpose of this work is not to add new methods to this already abundant literature, but instead to examine the performance of GNNs for the alignment problem through the lens of probability theory. Specifically, suppose that we observe two incomplete copies, G and G', of the same random geometric graph with sparse binary features. Additionally, for each graph, we also observe their vertex features perturbed by independent Gaussian noise. The goal is to match each vertex of G with a vertex of G'. In this work, we characterize how GNNs can facilitate this task leveraging both the noisy features and the graph structure.

1.1 Contributions

Our main contribution is to analyze the performance of GNNs for graph alignment tasks on a random geometric graph model and theoretically prove the benefit of message passing in the presence of vertex features. Specifically, the contribution is threefold:

1. We propose a random geometric graph model that generalizes many existing models and closely resembles real-life settings, allowing for the formal study of the graph alignment problem with the presence of vertex features.
2. We prove that in certain parameter regimes of the model, perfect recovery is possible with a specially-designed two-layer graph neural network. We also show that the dependence on the noise parameter is tight up to logarithmic factors.
3. Meanwhile, we demonstrate that directly aligning the vertices with noisy features fails to recover the vertex correspondence perfectly under certain conditions, while it remains possible with the help of the graph neural network.

1.2 Limitations

Although the current work is theoretical in nature and aims to explain the effectiveness of GNNs with probability tools, we discuss the limitations from both theoretical and application perspectives:

1. The random geometric graph model studied in this paper is still rather idealistic, even though we attempt to incorporate intuitions from real-world applications as much as possible.
2. Only perfect recovery is considered in the current work for clarity of the theoretical message, while in practice, other notions of alignment performance could also be relevant.
3. In the same way classical neural network literature has been able to gain valuable insights, we learn some interesting phenomena of GNNs from studying this simple setting. Notably we can highlight the role of message passing and nonlinearity, which other works fail to do.

1.3 Related work

Network alignment problem is usually approached through graph representation learning methods in modern machine learning literature (Yan et al., 2021). Earlier works focus on learning the low-dimensional vector representation of the entity by simple similarity metrics (Bordes et al., 2013) or probabilistic models (Grover and Leskovec, 2016). Later on, many variants of GNNs have been successfully applied, achieving state-of-the-art performance. Listing all relevant papers would be impossible, so instead, we direct interested readers to the survey article (Zeng et al., 2021). However, most of the advances are made from purely application-based considerations and lack theoretical justification.

On the other hand, random graph matching has attracted considerable theoretical investigation in recent years (Ding et al., 2021; Mao et al., 2021; Wu et al., 2022; Rácz and Sridhar, 2023; Fan et al., 2023a,b). However, the majority of the literature focuses on the correlated Erdős–Rényi model where two samples are independently created from the same Erdős–Rényi graph through edge-resampling. This method of introducing noise is similar to our subsampling process of the geometric graph. GNNs have recently been introduced to this line of research by Yu et al. (2023), which studies seeded graph matching with a carefully-designed architecture. The study of matching random graphs with latent geometric structure has emerged only in the past few years. The correlated stochastic block model introduced by Racz and Sridhar (2021) can be viewed as an intersection graph in one dimension. A later work (Rácz and Sridhar, 2023) studies the so-called k-core estimator for a wide range of inhomogeneous random graphs including random geometric graphs. A recent paper by Wang et al. (2022) considers matching random geometric graphs, but they assume that all pairwise dot products or Euclidean distances are observed under the Gaussian setting. One significant difference that sets our model apart from previous literature is that we consider a novel setting of sparse binary features with noisy observation. Nevertheless, we would like to emphasize that the main goal of our work is not to solve the graph matching problem optimally but rather to show that the effectiveness of GNNs in practice is indeed supported by theoretical analysis.

Due to the existence of vertex features, our work is also related to matching two random point clouds which has a long history in probability and combinatorics (Ajtai et al., 1984). More recently, Kunisky and Niles-Weed (2022) studied matching recovery for Gaussian perturbed Gaussian vectors in various regimes by solving an assignment problem. Notably, our feature generating process is similar to that investigated by the authors (Kunisky and Niles-Weed, 2022). Our work builds upon the same approach of solving an assignment problem but shows theoretically that with the help of the graph structure, modern machine learning methods are capable of achieving much better recovery guarantees. Again, we do not attempt to provide better recovery bounds on matching random points, but rather to supply a theoretical understanding of how graphs aid in the alignment problem.

The last literature we should mention pertains to learning guarantees for GNNs trained on random graphs generated according to a graphon. Stability and transferability of certain untrained GNNs have been established in (Ruiz et al., 2021; Maskey et al., 2023; Ruiz et al., 2023; Keriven et al., 2020). Generalization capabilities of GNNs (Maskey et al., 2022; Esser et al., 2021) and the capacity of GNNs to distinguish different graphons (Magner et al., 2020) have also been studied. In a different direction, Kawamoto et al. (2018); Lu (2021) considered the performance of GNNs for community detection respectively through heuristic mean-field approximations and formally for two community stochastic block models (SBMs) when the GNN is trained via coordinate descent. Baranwal et al. (2021) studied node classification for contextual stochastic block models (CSBMs) and showed that an oracle GNN can significantly boost the performance of linear classifiers. Wang and Wang (2024) investigated both a spectral algorithm and graph convolutional networks (GCNs) for node classification in CSBMs. Duranthon and Zdeborova (2024) derived a belief-propagation-based algorithm for the same task and showed a considerable gap between the accuracy reached by the proposed algorithm and the existing GNN architectures. Finally, Chung et al. (2024) have obtained guarantees for linear GNNs for the edge prediction task. Our goal is significantly different from all of these aforementioned works as we aim to establish the performance of GNNs for graph alignment tasks.

1.4 Notations

We use bold uppercase letters to denote matrices and the corresponding lowercase letters to denote their rows. For example, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is an n-by-d real matrix with row vectors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. For a vector $\boldsymbol{x}$, $\boldsymbol{x}(k)$ stands for its kth entry. The p-norm of a vector is denoted by $\|\cdot\|_p$ and it defaults to 2-norm, or the Euclidean norm, when p is not specified. The inner product between two d-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is written as $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \coloneqq \sum_{k=1}^{d} \boldsymbol{x}(k)\boldsymbol{y}(k)$. The notation $|\cdot|$ is generally overloaded. When it is applied to a number a, $|a|$ is the absolute value of a. And when applied to a set S, $|S|$ denotes its cardinality. We use $[n] \coloneqq \{1, \ldots, n\}$ to denote the set of all natural numbers up to n. For a set S, S^c stands for its complement. Gaussian (normal) distribution with mean μ and variance σ^2 is written as $\mathcal{N}(\mu, \sigma^2)$ and correspondingly the d-dimensional Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is written as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. $\Phi(x), x \in \mathbb{R}$ is the distribution function of $\mathcal{N}(0,1)$, and $\Phi^c(x) \coloneqq 1 - \Phi(x)$ represents the upper tail. We also make use of the standard big O notation: For positive functions $f(x)$ and $g(x)$, $f(x) \lesssim g(x)$ or $f(x) = O(g(x))$ if $f(x) \le Cg(x)$ for a constant $C > 0$ independent of x; $f(x) \gtrsim g(x)$ or $f(x) = \Omega(g(x))$ if $g(x) \lesssim f(x)$. We write $f(x) \asymp g(x)$ if $f(x) = O(g(x))$ and $g(x) = O(f(x))$. We denote $f(x) \ll g(x)$, $g(x) \gg f(x)$, or $f(x) = o(g(x))$ if $\lim_{x \to \infty} f(x)/g(x) = 0$.

2 PROBLEM DEFINITION

Our model is a generalization of the random intersection graph within the family of random geometric graphs. We first introduce several notations necessary for defining the model. A graph G on n vertices is denoted by a pair $G = (\boldsymbol{X}, E)$ where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is the list of vertex features and E is the set of undirected edges. That is, vertices i and j are connected by an undirected edge if and only if $\{i, j\} \in E$.

Definition 1 (Random intersection graph)**.** The random intersection graph $G_0 = (\boldsymbol{X}, E_0)$ is defined as follows. A d-dimensional binary feature vector $\boldsymbol{x}_i \in \{0,1\}^d$ is associated with each vertex i. Given a sparsity parameter $s \le d$, the entries of $\boldsymbol{x}_i$ follow an independent Bernoulli distribution with parameter s/d, i.e., $\mathbb{P}(\boldsymbol{x}_i(j) = 1) = s/d$ and $\mathbb{P}(\boldsymbol{x}_i(j) = 0) = 1 - s/d$ independently. For each pair of vertices $i \ne j \in [n]$, $\{i, j\} \in E_0$ if and only if

$$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \ge t$$

for a fixed threshold $t \ge 1$. We denote the graph by $\mathrm{RIG}(n, d, s, t)$.

Traditionally, a random intersection graph is defined by n random subsets of total d elements, and sets are connected if their intersection is nonempty (Singer, 1996). It is clear that this corresponds to the case $t = 1$ in our more general definition of RIG. Written in the form of Definition 1, it is also clear that RIG is a special case of random inner (dot) product graphs (Young and Scheinerman, 2007) when features are Bernoulli random variables. In contrast to the Erdős–Rényi model, edges in RIG are regulated by the underlying geometric space, which is in line with other random geometric graphs (Penrose, 2003).

Suppose that for a graph G_0, we do not have direct access to it. Instead, we are given two noisy and incomplete copies of it, where we observe perturbed feature vectors and retain only a subset of the edges. The procedure is designed to mimic the data patterns in real-world networks.

Definition 2 (Noisy and incomplete RIG)**.** Given a graph $G_0 = (\boldsymbol{X}, E_0) \sim \mathrm{RIG}(n, d, s, t)$, we assume that the observed graph $G = (\boldsymbol{Y}, E)$ is created by the following process.

1. For each vertex $i \in [n]$, the observed feature vector is given by
$$\boldsymbol{y}_i = \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i$$
where $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d)$ is independent noise for some noise parameter $\sigma \geq 0$.
2. For each pair of vertices $i \neq j \in [n]$, an edge $\{i, j\} \in E$ between them is observed with probability q only if $\{i, j\} \in E_0$. In other words,
$$\mathbb{P}(\{i, j\} \in E \mid \{i, j\} \in E_0) = q$$
and $\{i, j\} \notin E$ otherwise.

The graph is denoted by $\mathrm{NIRIG}(\boldsymbol{X}, \sigma, q)$.

We create a pair of correlated NIRIGs as follows. First, we generate a truth graph $G_0 = (\boldsymbol{X}, E_0) \sim \mathrm{RIG}(n, d, s, t)$. Then, we obtain two independent samples of $\mathrm{NIRIG}(\boldsymbol{X}, \sigma, q)$, G and G', from the same G_0, along with a vertex permutation between them. In other words, given $\boldsymbol{X}$,
$$G = (\boldsymbol{Y}, E) \sim \mathrm{NIRIG}(\boldsymbol{X}, \sigma, q)$$
and
$$G' = (\boldsymbol{Y}', E') \sim \mathrm{NIRIG}(\pi^*(\boldsymbol{X}), \sigma, q)$$
independently for an unknown permutation π^*. Our goal is to recover the true permutation π^* from the observations $G = (\boldsymbol{Y}, E)$ and $G' = (\boldsymbol{Y}', E')$.

3 MAIN RESULTS

The vanilla graph neural network (also called a graph convolutional network or message passing neural network) iteratively applies the following propagation rule (from layer l to $l+1$) to each vertex i in $G = (\boldsymbol{X}, E)$:
$$\boldsymbol{x}_i^{l+1} = \eta\left(\frac{1}{|N_i|} \sum_{j \in N_i} \boldsymbol{W}^l \boldsymbol{x}_j^l\right)$$
where η is a nonlinear activation function, $N_i \coloneqq \{j \in [n] : \{i, j\} \in E\}$ is the neighborhood of i, and $\boldsymbol{W}^l$ is the trainable weight matrix.

We employ a specially-designed two-layer graph neural network to find the matching between $G = (\boldsymbol{Y}, E)$ and $G' = (\boldsymbol{Y}', E')$. Define the neighborhood of vertex i in G and G' respectively as
$$N_i \coloneqq \{j \in [n] : \{i, j\} \in E\}$$
and
$$N_i' \coloneqq \{j \in [n] : \{i, j\} \in E'\}.$$

We apply a message passing layer to the observations $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ followed by a thresholding function
$$\eta(u) = \mathbb{1}\left\{u \geq \frac{t}{2s}\right\}.$$

Hence for each vertex $i \in [n]$ the hidden units of the two-layer graph neural network from the input graphs G and G' are respectively
$$\boldsymbol{z}_i = \eta\left(\frac{1}{|N_i|} \sum_{j \in N_i} \boldsymbol{y}_j\right) \quad \text{and} \quad \boldsymbol{z}_i' = \eta\left(\frac{1}{|N_i'|} \sum_{j \in N_i'} \boldsymbol{y}_j'\right).$$

These values are finally used to find the matching $\pi^\star$. This is done by solving an assignment problem between $\boldsymbol{Z}$ and $\boldsymbol{Z}'$:
$$\begin{aligned}\hat{\pi} &= \operatorname*{arg\,min}_{\pi \in \mathcal{P}(n)} \left(\ell(\pi) \coloneqq \sum_{i=1}^n \|\boldsymbol{z}_i - \boldsymbol{z}_{\pi(i)}'\|^2\right) \\ &= \operatorname*{arg\,max}_{\pi \in \mathcal{P}(n)} \sum_{i=1}^n \langle \boldsymbol{z}_i, \boldsymbol{z}_{\pi(i)}' \rangle\end{aligned} \tag{1}$$
where $\mathcal{P}(n)$ is the permutation group on $[n]$. This problem can be solved in polynomial time with combinatorial optimization algorithms such as the Hungarian algorithm (Kuhn, 1955).

The RIG is parametrized by the sparsity parameter s of the underlying features. However, it can be directly translated to properties of the underlying graph G_0. When n, d, and t are fixed, the edge density of the graph is indeed determined by s in a nontrivial way.

To make this dependence explicit, we define another parameter $m \in [0, n]$ such that

$$s = \sqrt{td\left(\frac{m}{n}\right)^{1/t}}. \tag{2}$$

It will be demonstrated by a later lemma (Lemma 3 in the appendix) that m roughly characterizes the average degree in $\mathrm{RIG}(n, d, s, t)$ when t is constant. For simplicity, we keep $t \geq 1$ fixed in our discussion. This notably implies, using (2), that $s \lesssim \sqrt{d}$.

Our main finding is that when the parameters n, d, m, σ, and q satisfy certain conditions, the solution found by (1) recovers the true permutation π^* with high probability.

Theorem 1 (Perfect recovery)**.** *Let the matching problem be defined in Section 2, and we solve it using* (1)*. With probability approaching* 1 *as* $n \to \infty$ *we recover both the true vertex features and the matching if*

$$\min\left\{s, \frac{qm}{s}, \frac{qm}{\sigma^2 s^2}\right\} \gg \log n + \log d.$$

Intuitively, there is a bias–variance trade-off in aggregating feature information from neighbors in a geometric graph: The noise cancels out by averaging, thus reducing the variance in estimation, while borrowing features from connected vertices increases the bias. Therefore, the parameters must satisfy certain conditions for the GNN to achieve perfect recovery. The proofs make use of the concentration of measure to derive various upper and lower bounds on important quantities such as the support of feature vectors and the neighborhood size. Finer and more detailed results can be found in Theorems 4 and 5 in the appendix along with their proofs.

Remark 1 (Reparameterization)**.** The parameter regime is specified by a mixture of s and m for the simplicity of the presentation. However, as defined in (2), s is determined by m and vice versa. Since we are more interested in graph properties, we choose the more natural parameter m. Replacing s with the definition in (2), we immediately have

$$\min\left\{\sqrt{d\left(\frac{m}{n}\right)^{1/t}}, \sqrt{\frac{q^2m^2}{d}\left(\frac{n}{m}\right)^{1/t}}, \frac{qm}{\sigma^2 d}\left(\frac{n}{m}\right)^{1/t}\right\} \\ \gg \log n + \log d.$$

Since q also affects the edge density of the graph but in a less sophisticated way (only through subsampling), for the interest of discussion, we also fix it to be a constant. Now we are ready to present an interplay between n, d, m, and σ in several graph density regimes. Notably for different graph densities m, we identify the permissible range for d, within which exact matching recovery is feasible, contingent on the noise level σ^2 being adequately small. We similarly derive the maximum size the noise σ^2 can take for which perfect recovery is still possible. We list the results in Table 1 when $t = 1$ and $t = 2$ for $m = a \log^{2+\epsilon} n, \epsilon > 0$ (sparse) and $m = bn^\alpha$ with $0 < \alpha < 1$ (intermediate) and $\alpha = 1$ (dense), where $a, b > 0$ are absolute constants.

Remark 2 (Phase diagram)**.** Note that only the last term in the bound depends on the noise parameter σ. All previous terms are functions of n, d, and m. If we consider only the intermediate regime when $m \asymp n^\alpha$ and $d \asymp n^\beta$, we can visualize the theorem as a diagram in the space of m and d as in Figure 1. The entire colored region is specified by the first two terms (and $m \leq n$) without σ. The term that involves σ "cuts through" the region and moves down as σ increases as a power of n. The dark blue region is where perfect recovery is possible for the GNN.

Remark 3 (Correlated random graph matching)**.** Information-theoretic results on correlated random graph matching have been investigated in a line of research (Cullina and Kiyavash, 2016; Wu et al., 2022; Ding and Du, 2023) and the sharp threshold for exact recovery was established for the Erdős–Rényi model. These results were later extended to inhomogeneous random graphs including random geometric graphs by Rácz and Sridhar (2023). Their result suggests that using the incomplete graphs alone, one can find the exact matching when $mq^2 \gg \log n$, compared to $\min\{\frac{mq}{s}, \frac{mq}{\sigma^2 s^2}\} \gg \log n$ as required in the theorem. Therefore, when $q \ll \min\{\frac{1}{s}, \frac{1}{\sigma^2 s^2}\}$, i.e., the graph is very sparse, GNN is able to recover the matching while the k-core estimator cannot. One additional note here is that the algorithms used to prove the information-theoretic results are usually not computationally efficient. For instance, the MAP estimator in (Cullina and Kiyavash, 2016) would require inspecting all permutations.

Remark 4 (Trainability of the GNN)**.** We focus on understanding the message passing in this work, and the GNN used for the alignment does not contain weight matrices. Nevertheless, the threshold in the activation function η can, in fact, be trained. Since we assume the RIG parameters are known to us, the threshold is chosen "optimally" in the algorithm. However, when dealing with real-world data where the parameters may not be known even if the generative model assumptions hold, the threshold may be learned from the data. We further investigate this point empirically in our real-world data experiments.

We have the following negative result that establishes the conditions under which the GNN cannot recover the matching perfectly.

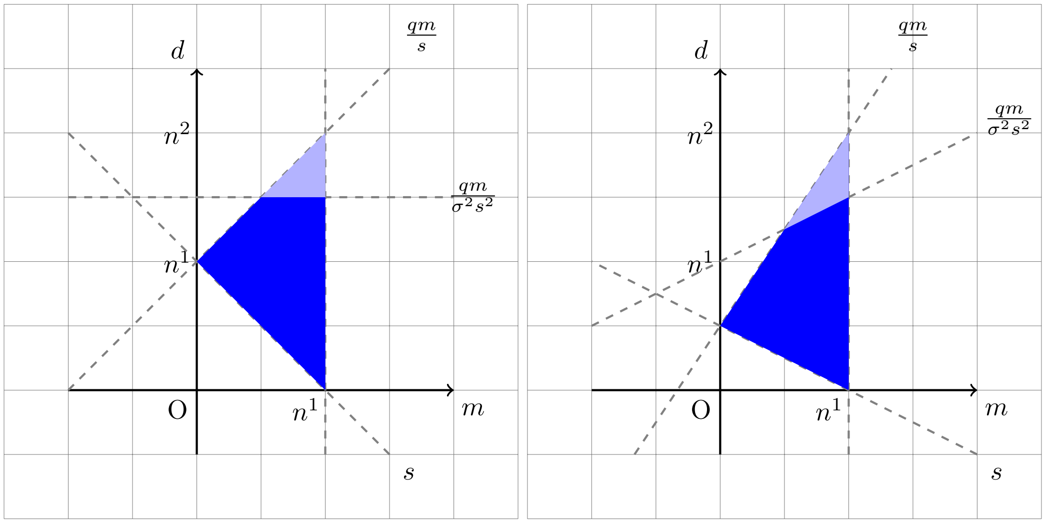


Figure 1: Phase diagram of perfect recovery for $t = 1$ (left) and $t = 2$ (right). Here $\sigma^2 \asymp 1/\sqrt{n}$.

Table 1: Recovery conditions for different thresholds t and sparsity levels m.

m	$(\log n)^{2+\epsilon}$	n^α
d for $t = 1$	$(\log n)^\epsilon n \gg d \gg \frac{n}{(\log n)^\epsilon}$	$\frac{n^{1+\alpha}}{(\log n)^2} \gg d \gg (\log n)^2 n^{1-\alpha}$
σ^2 for $t = 1$	$\sigma^2 \ll \frac{1}{(\log n)^{1-\epsilon}}$	$\sigma^2 \ll \frac{n^\alpha}{(\log n)^3}$
d for $t = 2$	$\sqrt{n}(\log n)^{1+\frac{3}{2}\epsilon} \gg d \gg \sqrt{n}(\log n)^{1-\frac{\epsilon}{2}}$	$\frac{n^{\frac{1+3\alpha}{2}}}{(\log n)^2} \gg d \gg (\log n)^2(\sqrt{n})^{1-\alpha}$
σ^2 for $t = 2$	$\sigma^2 \ll \frac{1}{(\log n)^{1-\epsilon}}$	$\sigma^2 \ll \frac{n^\alpha}{(\log n)^3}$

Theorem 2 (Impossibility of perfect recovery)**.** *Assume that* $\min\{s, \frac{qm}{s}\} \geq c$ *for a constant* $c > 0$*. Then there exists a constant* $\delta > 0$ *such that* $\mathbb{P}(\hat{\pi} \neq \pi^*) \geq \delta'$ *for all* $\delta' \in [0, \delta]$ *if*

$$\frac{qm}{\sigma^2 s^2} \leq C_\delta$$

for some constant $C_\delta > 0$*.*

To prove the impossibility result, we carefully characterize the event in which two vertices are misaligned with each other in the two observed graphs. Hence perfect recovery is not possible when this event occurs. The proof can be found in the appendix.

Remark 5. We note that the conditions $\min\{s, \frac{qm}{s}\} \geq c$ assumed here are also necessary for perfect recovery. These regularity conditions ensure that the vertices are unique and that the signal is sufficiently strong in the graph. The primary distinction between the possibility and impossibility bounds lies in the term $\frac{qm}{\sigma^2 s^2}$ which is also the only term that involves the noise parameter σ. For perfect recovery, it must be $\gg \log n + \log d$, whereas for impossibility, it must be bounded. This implies that the bound concerning the noise parameter σ is tight up to logarithmic factors. Notably combined with the perfect recovery results, in the dense regime $m \asymp n^\alpha$, we show that the algorithm can tolerate noise level σ that grows as a function of n.

Instead of using the graph neural network, one can obtain a matching by directly solving an alignment problem from the noisy vertex features $\boldsymbol{Y}$ and $\boldsymbol{Y}'$:

$$\begin{aligned} \tilde{\pi} &= \arg\min_{\pi \in \mathcal{P}(n)} \tilde{\ell}(\pi) := \sum_{i=1}^{n} \|\boldsymbol{y}_i - \boldsymbol{y}'_{\pi(i)}\|^2 \\ &= \arg\min_{\pi \in \mathcal{P}(n)} \sum_{i=1}^{n} \langle \boldsymbol{y}_i, \boldsymbol{y}'_{\pi(i)} \rangle. \end{aligned} \tag{3}$$

We call this the linear method. It is worth mentioning that this method has been widely used for alignment problems and also attracted significant theoretical analysis under different settings. However, as we will show in the next theorem, when the noise is large or the dimension is high, directly solving the assignment problem will not achieve perfect recovery with at least constant probability.

Theorem 3 (Impossibility of perfect recovery with vertex features). *If $\sigma^2 \geq 2s\big(1+\frac{K}{n}\big)^{-2}$ for any $K > 4$ we have that*

$$\mathbb{P}(\tilde{\pi} = \pi^*) \leq e^{-\frac{K-4}{4}}.$$

Hence if $\sigma^2 \gg s$, the probability of perfect recovery converges to 0. If instead $\frac{1}{4}\sigma^2 \leq s \leq \frac{d}{2}$ and $\sigma^2 \gg \frac{s}{\sqrt{d \log n}}$ then

$$\lim_{n\to\infty} \mathbb{P}(\tilde{\pi} = \pi^*) = 0.$$

To prove the first part of the theorem when the noise parameter is large, we make use of several information-theoretic inequalities. The proof of the second part involves analyzing the misalignment event. Finer and more detailed results are postponed to Propositions 9 and 10, along with their proofs in the appendix.

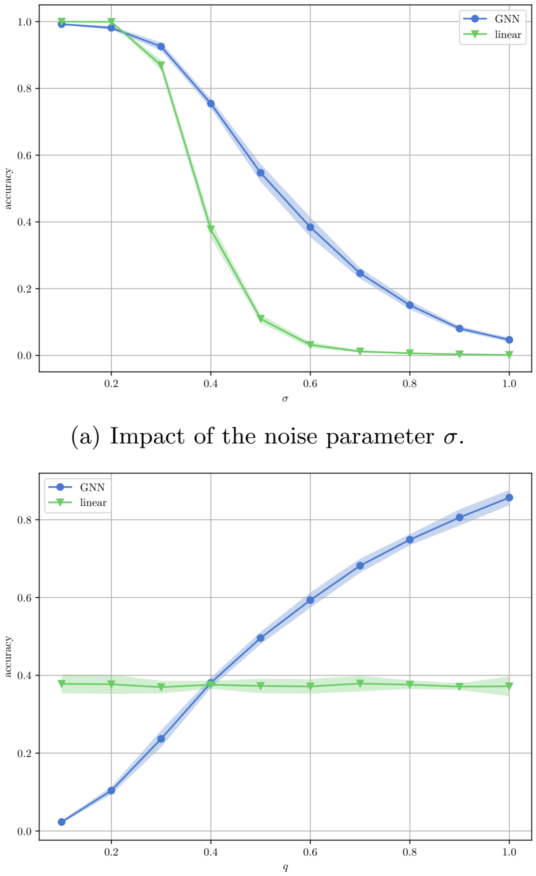
Remark 6. The impossibility result is split into two regimes: the signal-to-noise ratio being small and large. We are more interested in how these bounds compare to those achieved by the graph neural network. To make a direct comparison, we may translate the parameter s to m, although it does not have a clear physical meaning for aligning features. In one scenario, the message is clear: When σ is at least a constant, in all parameter regimes where the GNN has perfect recovery, the linear method would fail.

4 EXPERIMENTS

We validate our findings through experiments on artificially generated graphs as specified by our definitions, as well as two real-world network datasets. With a slight deviation from the theoretical results, we present the matching accuracy in the experimental results for computational concerns and practical considerations. This corresponds to the notion of partial recovery, the theory of which we leave for future work. All experiments are averaged over 5 independent runs, and we report the two standard deviations plotted as the shaded areas along the curve. (Note that due to the small variance in a few experiments, some shaded areas are very narrow and may not be clearly visible.) The results were produced on a shared research computing cluster. Required resources are 4 computing units with 20G RAM. Experiments for each set of plots typically finish within 20 minutes. Code for generating the experimental results is available at `https://github.com/6457/matchgnn/`.

4.1 Synthetic graphs

We first generate graphs according to the model introduced in Section 2. In Figure 2, we show the alignment accuracy, which is defined as the proportion of correctly matched pairs. We first fix the sparsification parameter q and vary the variance parameter σ, and then fix σ and vary q. As suggested by the theory, the linear method fails to recover the alignment when the feature noise is large, while the GNN is less prone to being affected by noise. The accuracy of the GNN also improves as q grows. Note that for the parameters chosen, the original graph is already very sparse. The average degree is about 56 compared to $n = 4000$ nodes. This explains why the error of the GNN is high when q is small.



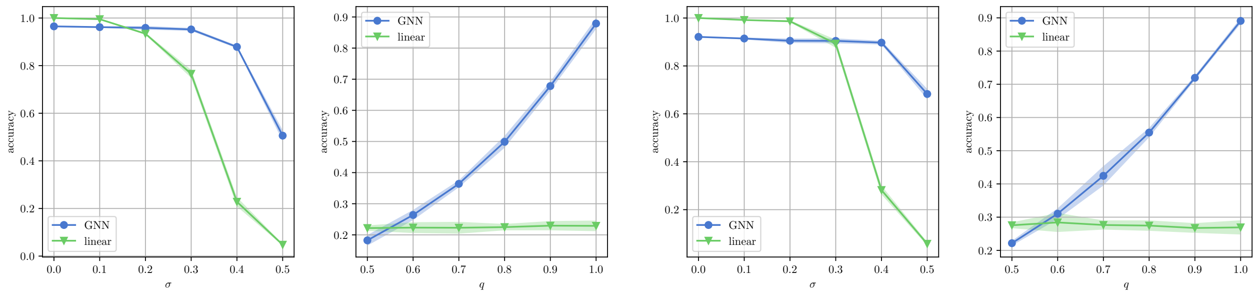
(a) Impact of the noise parameter σ.

(b) Impact of the sparsification parameter q.

Figure 2: Comparison of the GNN and the linear method. Parameters in the experiments are $n = 4000$, $d = 200$, $s = 10$, $t = 3$. For (a), $q = 0.8$ is fixed and σ ranges from 0.1 to 1 in 0.1 increments. For (b), $\sigma = 0.4$ and q changes from 0.1 to 1 in 0.1 increments.

4.2 Real-world datasets

We perform the alignment task on two public benchmark datasets, Cora and CiteSeer (Sen et al., 2008), which have been widely used to evaluate the performance of GNNs (Yang et al., 2016). Both datasets contain research papers with their bag-of-words rep-



(a) Cora

(b) CiteSeer

Figure 3: Comparison of the GNN and the linear method on real-world datasets. In the plots on the left of each group, $q = 1$ is fixed (using all edges from the datasets) and σ varies from 0 to 0.5 in 0.1 increments. In the plots on the right of each group, $\sigma = 0.4$ is fixed and q varies from 0.5 to 1 in 0.1 increments.

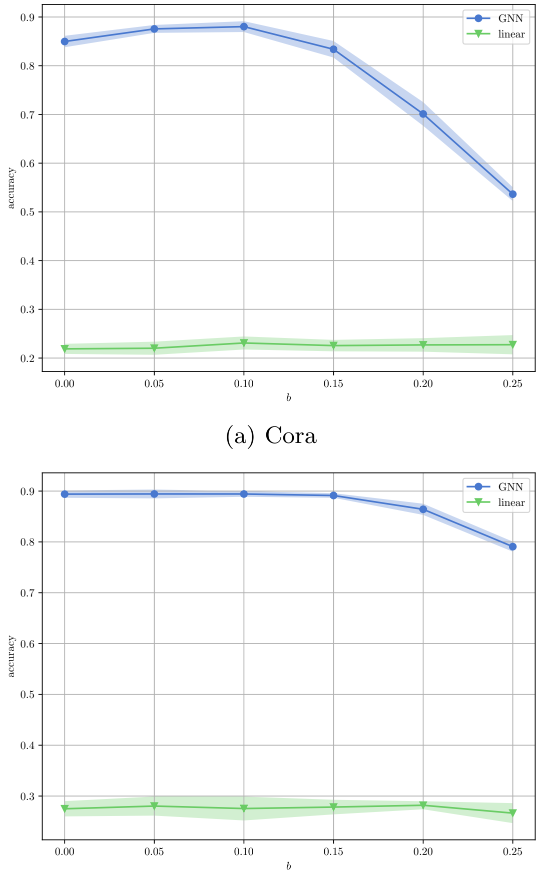
resentation. The dataset statistics may be found in Table 2.

Table 2: Summary of datasets.

Dataset	# Vertices	# Edges	# Features
Cora	2,708	5,429	1,433
CiteSeer	3,327	4,732	3,703

The evaluation task follows a similar approach to Yan et al. (2021). We treat the original network as our "ground truth" graph, and then create two copies of the graph by following the generating process of NIRIG as in Definition 2: Each word feature is perturbed with Gaussian noise and each edge is sampled with a fixed probability. The goal is to recover the article correspondence between them. The experiments are similar to those with the synthetic data. One additional tuning parameter here is the threshold in the activation function, as it is not known to us, even if the real-world networks are indeed geometric graphs. (We delve into this in the next set of experiments.) We found that the accuracy is not very sensitive to it in a large range within which the threshold is small. So we fix the threshold $b = 0.1$ in this part. We also replace the hard thresholding with a soft one for practical purposes. The results are presented in Figure 3.

Next, we evaluate the recovery accuracy of the GNN with different thresholds on our two real-world datasets (Figure 4). Interestingly, the two datasets exhibit very different response curves to the threshold. In Cora, the accuracy first grows when the threshold becomes larger and then drops. The U-shape curve suggests that there may exist an optimal choice of the threshold. Meanwhile, in CiteSeer, the accuracy remains the same in a large range of thresholds and only starts to decrease much later. This may be due to the fact that Cora consists of only machine learning publications, hence the number of overlapping words is comparably larger than in CiteSeer, which has more diverse fields in computer science, resulting in more accurate words from the neighborhood.



(a) Cora

(b) CiteSeer

Figure 4: Impact of the threshold parameter on real-world datasets. We fix $q = 1$ and $\sigma = 0.4$.

5 OPEN PROBLEMS

In this paper, we presented various possibility and impossibility results for perfect recovery under the noisy sparse feature setting. Similar inquiries could be made regarding other feature distributions, such as Gaussian features, which are the central object of study by Kunisky and Niles-Weed (2022), or spherical distributions arising from high-dimensional random geometric graphs (Devroye et al., 2011; Bubeck et al., 2016; Brennan et al., 2020; Liu and Rácz, 2023a). However, it would be less clear what role message passing plays under these settings. Nevertheless, averaging over the neighbors should still be effective thanks to the symmetry of the distributions.

Perfect recovery is the primary objective of this work. It would also be interesting and potentially useful in practice to explore partial recovery for both vertex features and the unknown alignment. There has been considerable research on partial recovery in either aligning Gaussian features (Kunisky and Niles-Weed, 2022) or random graph matching (Cullina et al., 2019). However, these questions still remain open for the current model. The techniques developed in the proofs of our perfect recovery results should bring us closer to finding the answers.

Acknowledgements

We thank Yuling Yan for many insightful discussions and helpful suggestions on an early draft of the paper, as well as the reviewers for their useful comments.

References

Miklós Ajtai, János Komlós, and Gábor Tusnády. On optimal matchings. *Combinatorica*, 4:259–264, 1984.

Aseem Baranwal, Kimon Fountoulakis, and Aukosh Jagannath. Graph convolution for semi-supervised classification: Improved linear separability and out-of-distribution generalization. *arXiv preprint arXiv:2102.06966*, 2021.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, 26, 2013.

Matthew Brennan, Guy Bresler, and Dheeraj Nagaraj. Phase transitions for detecting latent geometry in random graphs. *Probability Theory and Related Fields*, 178(3-4):1215–1289, 2020.

Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49(3):503–532, 2016.

Alan Chung, Amin Saberi, and Morgane Austern. Statistical guarantees for link prediction using graph neural networks. *arXiv preprint arXiv:2402.02692*, 2024.

Daniel Cullina and Negar Kiyavash. Improved achievability and converse bounds for erdos-rényi graph matching. *ACM SIGMETRICS Performance Evaluation Review*, 44(1):63–72, 2016.

Daniel Cullina, Negar Kiyavash, Prateek Mittal, and H Vincent Poor. Partial recovery of Erdös-Rényi graph alignment via k-core alignment. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–21, 2019.

Luc Devroye, András György, Gábor Lugosi, and Frederic Udina. High-Dimensional Random Geometric Graphs and their Clique Number. *Electronic Journal of Probability*, 16:2481 – 2508, 2011.

Jian Ding and Hang Du. Matching recovery threshold for correlated random graphs. *The Annals of Statistics*, 51(4):1718–1743, 2023.

Jian Ding, Zongming Ma, Yihong Wu, and Jiaming Xu. Efficient random graph matching via degree profiles. *Probability Theory and Related Fields*, 179: 29–115, 2021.

O Duranthon and Lenka Zdeborova. Optimal inference in contextual stochastic block models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Pascal Esser, Leena Chennuru Vankadara, and Debarghya Ghoshdastidar. Learning theory can (sometimes) explain generalisation in graph neural networks. *Advances in Neural Information Processing Systems*, 34:27043–27056, 2021.

Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. In *The World Wide Web Conference*, pages 417–426, 2019.

Zhou Fan, Cheng Mao, Yihong Wu, and Jiaming Xu. Spectral graph matching and regularized quadratic relaxations i algorithm and gaussian analysis. *Foundations of Computational Mathematics*, 23(5):1511–1565, 2023a.

Zhou Fan, Cheng Mao, Yihong Wu, and Jiaming Xu. Spectral graph matching and regularized quadratic relaxations ii: Erdős-rényi graphs and universality. *Foundations of Computational Mathematics*, 23(5): 1567–1617, 2023b.

Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *International Conference on Machine Learning*, pages 3419–3430. PMLR, 2020.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR, 2017.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, 2016.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Stefanie Jegelka. *Theory of graph neural networks: representation and learning*, volume 7, pages 5450–5476. EMS Press, 2022.

Jean-Pierre Kahane. *Some random series of functions*, volume 5. Cambridge University Press, 1985.

Tatsuro Kawamoto, Masashi Tsubaki, and Tomoyuki Obuchi. Mean-field theory of graph neural networks in graph partitioning. *Advances in Neural Information Processing Systems*, 31, 2018.

Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. Convergence and stability of graph convolutional networks on large random graphs. *Advances in Neural Information Processing Systems*, 33:21512–21523, 2020.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2016.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Dmitriy Kunisky and Jonathan Niles-Weed. Strong recovery of geometric planted matchings. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 834–876. SIAM, 2022.

Suqi Liu and Miklós Z Rácz. Phase transition in noisy high-dimensional random geometric graphs. *Electronic Journal of Statistics*, 17(2):3512–3574, 2023a.

Suqi Liu and Miklós Z Rácz. A probabilistic view of latent space graphs and phase transitions. *Bernoulli*, 29(3):2417–2441, 2023b.

Andreas Loukas. What graph neural networks cannot learn: depth vs width. In *International Conference on Learning Representations*, 2020.

Wei Lu. Learning guarantees for graph convolutional networks on the stochastic block model. In *International Conference on Learning Representations*, 2021.

Abram Magner, Mayank Baranwal, and Alfred O Hero. The power of graph convolutional networks to distinguish random graph models. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2664–2669. IEEE, 2020.

Cheng Mao, Mark Rudelson, and Konstantin Tikhomirov. Random graph matching with improved noise robustness. In *Conference on Learning Theory*, pages 3296–3329. PMLR, 2021.

Sohir Maskey, Ron Levie, Yunseok Lee, and Gitta Kutyniok. Generalization analysis of message passing neural networks on large random graphs. *Advances in Neural Information Processing Systems*, 35:4805–4817, 2022.

Sohir Maskey, Ron Levie, and Gitta Kutyniok. Transferability of graph neural networks: an extended graphon approach. *Applied and Computational Harmonic Analysis*, 63:48–83, 2023.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and Leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4602–4609, 2019.

Galileo Namata, Ben London, Lise Getoor, and Bert Huang. Query-driven active surveying for collective classification. In *International Workshop on Mining and Learning with Graphs*, Edinburgh, Scotland, 2012.

Raymond EAC Paley and Antoni Zygmund. On some series of functions,(3). In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pages 190–205. Cambridge University Press, 1932.

Mathew Penrose. *Random Geometric Graphs*. Oxford University Press, 05 2003.

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

Miklós Z Rácz and Anirudh Sridhar. Matching correlated inhomogeneous random graphs using the k-core estimator. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 2499–2504. IEEE, 2023.

Luana Ruiz, Zhiyang Wang, and Alejandro Ribeiro. Graphon and graph neural network stability. In

ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5255–5259. IEEE, 2021.

Luana Ruiz, Luiz FO Chamon, and Alejandro Ribeiro. Transferability properties of graph neural networks. *IEEE Transactions on Signal Processing*, 2023.

Franco Scarselli, Ah Chung Tsoi, and Markus Hagenbuchner. The vapnik–chervonenkis dimension of graph and recursive neural networks. *Neural Networks*, 108:248–259, 2018.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer, 2018.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–93, 2008.

Roded Sharan and Trey Ideker. Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24(4):427–433, 2006.

Karen B Singer. *Random intersection graphs*. PhD thesis, Johns Hopkins University, 1996.

Ilya S Tyurin. An improvement of upper estimates of the constants in the lyapunov theorem. *Russian Mathematical Surveys*, 65(3):201–202, 2010.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Haixiao Wang and Zhichao Wang. Optimal exact recovery in semi-supervised learning: a study of spectral methods and graph convolutional networks. In *Proceedings of the 41st International Conference on Machine Learning*, pages 51614–51649, 2024.

Haoyu Wang, Yihong Wu, Jiaming Xu, and Israel Yolou. Random graph matching in geometric models: the case of complete graphs. In *Conference on Learning Theory*, pages 3441–3488. PMLR, 2022.

Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 349–357, 2018.

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

Yuchen Yan, Si Zhang, and Hanghang Tong. Bright: A bridging algorithm for network alignment. In *Proceedings of the Web Conference 2021*, pages 3907–3917, 2021.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International Conference on Machine Learning*, pages 40–48. PMLR, 2016.

Stephen J Young and Edward R Scheinerman. Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 138–149. Springer, 2007.

Liren Yu, Jiaming Xu, and Xiaojun Lin. Seedgnn: graph neural network for supervised seeded graph matching. In *International Conference on Machine Learning*, pages 40390–40411. PMLR, 2023.

Kaisheng Zeng, Chengjiang Li, Lei Hou, Juanzi Li, and Ling Feng. A comprehensive survey of entity alignment for knowledge graphs. *AI Open*, 2:1–13, 2021.

Anru R Zhang and Yuchen Zhou. On the non-asymptotic and sharp lower tail bounds of random variables. *Stat*, 9(1):e314, 2020.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes]

   (b) Complete proofs of all theoretical results. [Yes]

   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A PRELIMINARIES

In this section we present some preliminary results on the tails of different random variables. These will be used later in our proofs.

Proposition 1 (Chernoff bound)**.** *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables and $N = \sum_{i=1}^n X_i$ be their sum. Denote $\mathbb{E}[N] = \mu$ the expected value of the sum. Then for any $\delta \geq 0$,*

$$\mathbb{P}(N \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2+\delta}\right)$$

and for any $0 < \delta < 1$,

$$\mathbb{P}(N \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2}\right).$$

The following lower bound for the tail of a nonnegative random variable is due to Paley and Zygmund (1932) (see also (Kahane, 1985, Inequality II, p. 8)).

Proposition 2 (Paley–Zygmund)**.** *Let Z be a nonnegative random variable with finite variance. For $0 \leq \theta \leq 1$,*

$$\mathbb{P}(Z \geq \theta \mathbb{E}[Z]) \geq (1-\theta)^2 \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}.$$

We exploit this proposition to deduce the following tail lower bound for the dot product of Gaussian vectors.

Proposition 3 (Lower bound for Gaussian inner product)**.** *Let $\boldsymbol{x}, \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ be two independent standard Gaussian random vectors. We have that for any $u \leq \sqrt{d}/16$, the following holds*

$$\mathbb{P}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle \geq u\sqrt{d}) \geq (1 - e^{-4u^2})^2 e^{-22u^2}.$$

Proof. Denote $Z = \exp(t\langle \boldsymbol{x}, \boldsymbol{y} \rangle)$ for $0 < t < 1/2$. By the proof of (Liu and Rácz, 2023b, Proposition 2.14), we have

$$\mathbb{E}[Z] = \mathbb{E}[e^{t\langle \boldsymbol{g}_1, \boldsymbol{g}_2 \rangle}] = \prod_{i=1}^d \mathbb{E}[e^{t\boldsymbol{g}_{1,i}\boldsymbol{g}_{2,i}}] = (1-t^2)^{-d/2}.$$

Hence this implies that for every $t < \frac{1}{4}$ the following holds

$$\mathbb{E}[Z^2] = \mathbb{E}[e^{2t\langle \boldsymbol{g}_1, \boldsymbol{g}_2 \rangle}] = (1-4t^2)^{-d/2}.$$

By Paley–Zygmund (Proposition 2), for all $\theta \in (0,1)$ we have

$$\mathbb{P}\left(\langle \boldsymbol{x}, \boldsymbol{y} \rangle \geq -\frac{d}{2t}\log(1-t^2) + \frac{1}{t}\log\theta\right) \geq (1-\theta)^2 \left(\frac{1-t^2}{\sqrt{1-4t^2}}\right)^{-d}.$$

By taking $t = 4u/\sqrt{d}$ and $\theta = \exp(-4u^2)$ for $u \leq \sqrt{d}/16$, we have that

$$-\frac{d}{2t}\log(1-t^2) + \frac{1}{t}\log\theta = -\frac{d^{3/2}}{8u}\log\left(1 - \frac{16u^2}{d}\right) - u\sqrt{d} \geq u\sqrt{d}$$

where we used $\log(1-x) \leq -x$, and

$$\left(\frac{1-t^2}{\sqrt{1-4t^2}}\right)^{-d} = \left(1 + 2t^2\left(\frac{1+t^2/2}{1-4t^2}\right)\right)^{-d/2} \geq \left(1 + \frac{44u^2}{d}\right)^{-d/2} \geq e^{-22u^2}$$

where we used $(1+x/d)^d \leq e^x$. □

Proposition 4 (Lower bound for lazy random walk tail)**.** *Let $S_n = \sum_{i=1}^n X_i$ be a lazy random walk such that $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = p$ and $\mathbb{P}(X_i = 0) = 1 - 2p$. Assume $pn \geq 1$. Then for $0 \leq u \leq pn/\beta$, there exist constants $c_\beta, C_\beta > 0$ that depend only on β such that*

$$\mathbb{P}(S_n \geq u) \geq c_\beta \exp\left(-C_\beta \frac{u^2}{pn}\right).$$

Proof. Let $Y_i \in \{-1, 1\}$ for $i \in [n]$ be i.i.d. Rademacher random variables, i.e., $\mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i = -1) = 1/2$. Let $Z_i \sim \mathrm{Bern}(2p), i \in [n]$ be i.i.d. Bernoulli random variables. Then, we can write $X_i = Y_i Z_i$. Let $N_n = \sum_{i=1}^n Z_i$ be the sum of Z_i's. Then given N_n, S_n is a sum of N_n i.i.d. Rademacher random variables. By (Zhang and Zhou, 2020, Corollary 4), for any $\beta > 1$, there exists constants $c_\beta, C_\beta > 0$, such that for all $0 \leq u \leq N_n/\beta$,

$$\mathbb{P}(S_n \geq u \mid N_n) \geq c_\beta \exp\left(-C_\beta \frac{u^2}{N_n}\right).$$

Chernoff bound (Proposition 1) gives

$$\mathbb{P}\left(N_n \leq \frac{pn}{2}\right) \leq \exp\left(-\frac{pn}{8}\right).$$

Therefore, we conclude that

$$\begin{aligned}\mathbb{P}(S_n \geq u) &\geq \mathbb{P}\left(S_n \geq u, N_n \geq \frac{pn}{2}\right) = \mathbb{P}\left(S_n \geq u \,\middle|\, N_n \geq \frac{pn}{2}\right)\left(1 - \mathbb{P}\left(N_n \leq \frac{pn}{2}\right)\right) \\ &\geq c_\beta \exp\left(-C_\beta \frac{8u^2}{pn}\right)\left(1 - \exp\left(-\frac{pn}{8}\right)\right).\end{aligned}$$

The claim directly follows. □

B AUXILIARY LEMMAS

A necessary condition for perfect recovery is that the feature vector for each vertex is unique. The following proposition tells us that this happens with high probability in RIG when s is sufficiently large.

Proposition 5 (No two vectors are the same)**.** *Let $\boldsymbol{x}_i$'s be defined in Definition 1. If $s \geq (1 + c)\log n$ for some $c > 0$, then no two vectors in $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are the same with probability at least $1 - n^{-2c}$.*

Proof. As the entries $\boldsymbol{x}_i(k), \boldsymbol{x}_j(k) \overset{i.i.d}{\sim} \mathrm{Bern}(\frac{s}{d})$, we obtain that

$$\mathbb{P}(\boldsymbol{x}_i(k) \neq \boldsymbol{x}_j(k)) = \frac{2s}{d}.$$

Hence, by independence of the entries, we have

$$\mathbb{P}(\boldsymbol{x}_i = \boldsymbol{x}_j) = \mathbb{P}(\boldsymbol{x}_i(k) = \boldsymbol{x}_j(k) \ \forall k \leq d) = \left(1 - \frac{2s}{d}\right)^d.$$

Hence by a union bound argument we obtain that

$$\mathbb{P}(\exists i \neq j \leq n \text{ s.t. } \boldsymbol{x}_i = \boldsymbol{x}_j) \leq \sum_{i<j} P(\boldsymbol{x}_i = \boldsymbol{x}_j) \leq \binom{n}{2}\left(1 - \frac{2s}{d}\right)^d \leq e^{-2(s - \log n)}.$$

The claim directly follows. □

Let S_i be the support of $\boldsymbol{x}_i$. Since $\boldsymbol{x}_i(k), k \in d$ are i.i.d. Bernoulli random variables with parameter s/d, applying Chernoff bound (Proposition 1) to $|S_i| = \sum_{k=1}^d \boldsymbol{x}_i(k)$ directly gives us the following lower and upper deviation bounds for $|S_i|$.

Lemma 1. *For all $i \in [n]$ and any $\delta \geq 0$,*

$$\mathbb{P}(|S_i| \geq (1 + \delta)s) \leq \exp\left(-\frac{\delta^2 s}{2 + \delta}\right).$$

Lemma 2. *For all $i \in [n]$ and $0 < \delta < 1$,*

$$\mathbb{P}(|S_i| \leq (1 - \delta)s) \leq \exp\left(-\frac{\delta^2 s}{2}\right).$$

Define the event

$$\mathcal{E}_i := \left\{ \frac{s}{2} \leq |S_i| \leq 2s \right\}. \tag{4}$$

Lemma 2 and 1 immediately suggest that

$$\mathbb{P}(\mathcal{E}_i^c) \leq \exp\left(-\frac{s}{8}\right) + \exp\left(-\frac{s}{3}\right) \leq 2\exp\left(-\frac{s}{8}\right). \tag{5}$$

In the proofs, we use a lot of conditional arguments. Most frequently, $s/2 \leq |S_i| \leq 2s$, which happens with high probability from previous lemmas, is usually assumed. The following elementary inequality becomes handy in such situations.

Proposition 6. *For any two events A and B,*

$$\mathbb{P}(A) \leq \mathbb{P}(A \mid B) + \mathbb{P}(B^c).$$

Proof. By law of total probability,

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) = \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c) \leq \mathbb{P}(A \mid B) + \mathbb{P}(B^c),$$

where we used the fact that the probability measure is less than or equal to 1. □

For all vertex $i \in [n]$, let

$$\mathcal{F}_i := \left\{ \frac{1}{2^{t+2}} \cdot mq \leq |N_i| \leq 2^{t+2}e^t \cdot mq \right\}. \tag{6}$$

Lemma 3. *The following holds for all $i \in [n]$ when $n \geq (12t)^t m$,*

$$\mathbb{P}(\mathcal{F}_i^c) \leq 2\exp\left(-\frac{s}{8}\right) + 2\exp\left(-\frac{mq}{2^{t+6}}\right).$$

Proof. By definition the neighborhood size is given by $|N_i| = \sum_{j=1}^n \mathbb{1}\{j \in N_i\}$. Moreover we remark that conditionally on $\boldsymbol{x}_i$, the edges $\left(\mathbb{1}\{j \in N_i\}\right)_{j\in[n]}$ are independent and identically distributed. We denote by $p_i := \mathbb{P}(j \in N_i \mid \boldsymbol{x}_i)$ the edge probability. Hence we remark that conditionally on $\boldsymbol{x}_i$, $|N_i|$ is a sum of i.i.d. Bernoulli random variables with parameter p_i.

We first obtain upper and lower bounds for p_i conditionally the event $\mathcal{E}_i$ defined in (4) holding. For the rest of the proof, all probabilities are conditioned on $\mathcal{E}_i$ and we omit it from the condition for simplicity of presentation. We note that $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \sum_{k=1}^d \boldsymbol{x}_i(k)\boldsymbol{x}_j(k)$ is a $\mathrm{Binom}(|S_i|, \frac{s}{d})$ random variable. Recall that the edges are a subsample of the original graph with probability q. Hence the conditional probability that $j \in N_i$ is given by

$$\begin{aligned}\mathbb{P}(j \in N_i \mid \boldsymbol{x}_i) &= \mathbb{P}(j \in N_i \mid \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \geq t)\mathbb{P}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \geq t \mid \boldsymbol{x}_i) = q\mathbb{P}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \geq t \mid \boldsymbol{x}_i) \\ &= q\sum_{k=t}^{|S_i|} \binom{|S_i|}{k} \left(\frac{s}{d}\right)^k \left(1 - \frac{s}{d}\right)^{|S_i|-k} =: p_i.\end{aligned}$$

Then we can lower bound p_i by

$$\begin{aligned}p_i &= q\sum_{k=t}^{|S_i|} \binom{|S_i|}{k} \left(\frac{s}{d}\right)^k \left(1 - \frac{s}{d}\right)^{|S_i|-k} \geq q\binom{|S_i|}{t}\left(\frac{s}{d}\right)^t \left(1 - \frac{s}{d}\right)^{|S_i|-t} \\ &\overset{(a)}{\geq} q\left(\frac{|S_i|s}{td}\right)^t \left(1 - \frac{|S_i|s}{d}\right),\end{aligned}$$

where we used the elementary inequalities $\binom{n}{k} \geq (n/k)^k$ and $(1-x)^r \geq 1 - rx$ for $x \leq 1$ in (a). Hence, when $s/2 \leq |S_i| \leq 2s$ and $d \geq 4s^2$ (or equivalently $n \geq (4t)^t m$), we have

$$p_i \geq \frac{q}{2^t}\left(\frac{s^2}{td}\right)^t \left(1 - \frac{2s^2}{d}\right) \overset{(a)}{\geq} \frac{1}{2^{t+1}} \cdot \frac{mq}{n} \tag{7}$$

where to get (a) we used (2). We also have that for $s/2 \le |S_i| \le 2s$ and $d \ge 12s^2/t$ (or equivalently $n \ge (12)^t m$ according to (2)),

$$\begin{aligned} p_i &= q \sum_{k=t}^{|S_i|} \binom{|S_i|}{k} \left(\frac{s}{d}\right)^k \left(1 - \frac{s}{d}\right)^{|S_i|-k} \overset{(a)}{\le} q \sum_{k=t}^{|S_i|} \left(\frac{es|S_i|}{kd}\right)^k \le q \sum_{k=t}^{|S_i|} \left(\frac{2es^2}{td}\right)^k \\ &\le q \sum_{k=t}^{\infty} \left(\frac{2es^2}{td}\right)^k \le 2q \left(\frac{2es^2}{td}\right)^t \overset{(b)}{=} 2^{t+1} e^t \cdot \frac{mq}{n}, \end{aligned} \tag{8}$$

where we used $\binom{n}{k} \le (\frac{en}{k})^k$ in (a) and the equality (b) is by (2). Together with Proposition 1, we conclude that

$$\begin{aligned} \mathbb{P}(|N_i| \ge 2^{t+2} e^t \cdot mq) &= \mathbb{P}\left(|N_i| \ge 2^{t+2} e^t \cdot \frac{mq}{(n-1)p_i} \cdot (n-1)p_i\right) \\ &\overset{(a)}{\le} \mathbb{P}\left(|N_i| \ge \left(1 + \frac{n+1}{n-1}\right)(n-1)p_i\right) \le \exp\left(-\frac{(n+1)^2 p_i}{2n}\right) \\ &\overset{(b)}{\le} \exp\left(-\left(1 + \frac{1}{n}\right)^2 \frac{mq}{2^{t+2}}\right) \le \exp\left(-\frac{mq}{2^{t+2}}\right) \end{aligned}$$

where we used (8) in (a) and (7) in (b), and

$$\begin{aligned} \mathbb{P}\left(|N_i| \le \frac{1}{2^{t+2}} \cdot mq\right) &= \mathbb{P}\left(|N_i| \le \frac{1}{2^{t+2}} \cdot \frac{mq}{(n-1)p_i} \cdot (n-1)p_i\right) \\ &\overset{(c)}{\le} \mathbb{P}\left(|N_i| \ge \left(1 - \frac{n-2}{2n-2}\right)(n-1)p_i\right) \le \exp\left(-\frac{(n-2)^2 p_i}{8(n-1)}\right) \\ &\overset{(d)}{\le} \exp\left(-\frac{(n-2)^2 mq}{2^{t+4} n(n-1)}\right) \le \exp\left(-\frac{mq}{2^{t+6}}\right) \end{aligned}$$

where we used (7) in (c) and (8) in (d).

The lemma directly follows from Proposition 6 combined with (4). □

C PROOFS OF THE PERFECT RECOVERY RESULTS

Since $\hat{\pi}$ is the minimizer of the empirical risk ℓ, we have $\ell(\hat{\pi}) \le \ell(\pi^*)$. Without loss of generality, we assume that π^* is the identity transformation in the proofs hence $\pi^*(i) = i$ for all $i \in [n]$. By triangle inequality, we have

$$\ell(\hat{\pi}) \le \ell(\pi^*) = \sum_{i=1}^n \|\boldsymbol{z}_i - \boldsymbol{z}'_i\|^2 \le \sum_{i=1}^n (\|\boldsymbol{z}_i - \boldsymbol{x}_i\| + \|\boldsymbol{z}'_i - \boldsymbol{x}_i\|)^2 \le 2\sum_{i=1}^n \|\boldsymbol{z}_i - \boldsymbol{x}_i\|^2 + 2\sum_{i=1}^n \|\boldsymbol{z}'_i - \boldsymbol{x}_i\|^2. \tag{9}$$

We will show that with high probability both $\sum_{i=1}^n \|\boldsymbol{z}'_i - \boldsymbol{x}_i\|^2$ and $\sum_{i=1}^n \|\boldsymbol{z}_i - \boldsymbol{x}_i\|^2$ are equal to zero. This will directly imply that with high probability the empirical risk $\ell(\hat{\pi})$ is also zero.

Theorem 4 (Perfect recovery for vertex features)**.** *Let $G = (\boldsymbol{Y}, E) \sim \mathrm{NIRIG}(\boldsymbol{X}, \sigma, q)$ be defined as in Section 2 and let $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_d)$ be the output of the graph neural network described in Section 3. Then for $n \ge (12t)^t m$,*

$$\mathbb{P}(\exists i \ s.t.\ \boldsymbol{x}_i \ne \boldsymbol{z}_i) \le nd\left(3\exp\left(-\frac{mq}{5 \cdot 2^{t+7} s}\right) + 3\exp\left(-\frac{s}{175}\right) + \exp\left(-\frac{mq}{2^{t+7} s^2 \sigma^2}\right)\right).$$

By Proposition 5, with probability at least $1 - n^2 e^{-2s}$ all vertex features $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are unique. Therefore, π^* is the unique minimizer of (1). Since by Theorem 1 we know that $\ell(\hat{\pi}) = 0$ with high probability, we immediately have the following theorem by Proposition 6.

Theorem 5 (Prefect recovery for vertex matching)**.** *Let $\hat{\pi}$ be the solution of* (1)*. Then for $n \ge (12t)^t m$,*

$$\mathbb{P}(\hat{\pi} \ne \pi^*) \le n^2 \exp(-2s) + 2nd\left(3\exp\left(-\frac{mq}{5 \cdot 2^{t+7} s}\right) + 3\exp\left(-\frac{s}{175}\right) + \exp\left(-\frac{mq}{2^{t+7} s^2 \sigma^2}\right)\right).$$

The rest of this section is devoted to proving Theorem 4. The proof is centered around events defined as follows. For all $i \in [n]$ and $k \in [d]$, let

$$\mathcal{A}_{i,k} \coloneqq \biggl\{ \frac{s}{t|N_i|} \sum_{j \in N_i} \boldsymbol{x}_j(k) + \frac{s}{t|N_i|} \sum_{j \in N_i} \boldsymbol{\varepsilon}_j(k) \le \frac{1}{2} \biggr\}. \tag{10}$$

The next four lemmas concern events that will directly lead to bounds on $\mathcal{A}_{i,k}$.

Lemma 4. *For all $i \in [n]$,*

$$\mathbb{P}\biggl(\frac{s}{t|N_i|} \sum_{j \in N_i} \boldsymbol{\varepsilon}_j(k) \ge \frac{1}{4} \biggm| |N_i| \biggr) \le \exp\biggl(-\frac{t^2|N_i|}{32 s^2 \sigma^2} \biggr).$$

Proof. By definition of $\boldsymbol{\varepsilon}_j$, $\boldsymbol{\varepsilon}_j(k)$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ for $j \in N_i$. Hence,

$$\mathbb{P}\biggl(\frac{s}{t|N_i|} \sum_{j \in N_i} \boldsymbol{\varepsilon}_j(k) \ge \frac{1}{4} \biggm| |N_i| \biggr) = \Phi^c\biggl(\frac{t\sqrt{|N_i|}}{4s\sigma} \biggr).$$

By the upper tail bound of standard normal distribution (see, e.g., (Wainwright, 2019, (2.7))),

$$\Phi^c\biggl(\frac{t\sqrt{|N_i|}}{4s\sigma} \biggr) \le \exp\biggl(-\frac{t^2|N_i|}{32 s^2 \sigma^2} \biggr).$$

The claim directly follows. □

Lemma 5. *For all $i \in [n]$, when $n \ge 8^t m$,*

$$\mathbb{P}\biggl(\frac{s}{t|N_i|} \sum_{j \in N_i} \boldsymbol{x}_j(k) \ge \frac{1}{4} \biggm| \boldsymbol{x}_i(k) = 0, |N_i| \biggr) \le \exp\biggl(-\frac{t|N_i|}{24s} \biggr).$$

Proof. First we remark that knowing $\boldsymbol{x}_i(k) = 0$, the coordinates $(\boldsymbol{x}_j(k))_{j \in N_i}$ are i.i.d. $\mathrm{Bern}(s/d)$ random variables. Hence $\mathbb{E}[\sum_{j \in N_i} \boldsymbol{x}_j(k) \mid |N_i|] = s|N_i|/d$. Therefore by taking $\delta = \frac{td}{4s^2}$ in Proposition 1, we have

$$\mathbb{P}\biggl(\frac{s}{t|N_i|} \sum_{j \in N_i} \boldsymbol{x}_j(k) \ge \frac{1}{4} \biggm| \boldsymbol{x}_i(k) = 0, |N_i| \biggr) \le \exp\biggl(-\Bigl(\frac{td}{4s^2} - 1 \Bigr)^2 \cdot \Bigl(\frac{td}{4s^2} + 1 \Bigr)^{-1} \cdot \frac{s|N_i|}{d} \biggr).$$

For $n \ge 8^t m$, we have $\frac{td}{4s^2} \ge 2$. Hence,

$$\Bigl(\frac{td}{4s^2} - 1 \Bigr)^2 \cdot \Bigl(\frac{td}{4s^2} + 1 \Bigr)^{-1} \cdot \frac{s|N_i|}{d} \ge \Bigl(\frac{td}{8s^2} \Bigr)^2 \cdot \Bigl(\frac{3td}{8s^2} \Bigr)^{-1} \cdot \frac{s|N_i|}{d} = \frac{t|N_i|}{24s}.$$

The lemma directly follows. □

Lemma 6. *For all $i \in [n]$,*

$$\mathbb{P}\biggl(\frac{s}{t|N_i|} \sum_{j \in N_i} \Bigl(\frac{t}{s} - \boldsymbol{x}_j(k) \Bigr) \ge \frac{1}{4} \biggm| \boldsymbol{x}_i(k) = 1, |N_i| \biggr) \le \exp\biggl(-\frac{t|N_i|}{640s} \biggr) + \exp\biggl(-\frac{s}{175} \biggr).$$

Proof. We consider a vertex $j \in N_i$. Since $j \in N_i$, $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \sum_{k \in S_i} \boldsymbol{x}_j(k) \ge t$. Given S_i, $\bigl(\boldsymbol{x}_j(k)\bigr)_{k \in S_i}$ are identical (but not independent) Bernoulli random variables. Hence $\mathbb{P}(\boldsymbol{x}_j(k) = 1 \mid j \in N_i, k \in S_i) = \beta$ are the same for all $k \in S_i$. Since

$$\mathbb{E}\biggl[\sum_{k \in S_i} \boldsymbol{x}_j(k) \biggm| S_i, j \in N_i \biggr] = \sum_{k \in S_i} \mathbb{P}(\boldsymbol{x}_j(k) = 1 \mid j \in N_i, k \in S_i) = \beta|S_i| \ge t,$$

we have $\beta \geq t/|S_i|$. By Lemma 1 with $\delta = 1/4$, we have

$$\mathbb{P}\bigg(|S_i| \leq \frac{5}{4}s\bigg) \geq 1 - \exp\bigg(-\frac{s}{175}\bigg). \tag{11}$$

Now conditioned on $\boldsymbol{x}_i(k) = 1$, $\big(\boldsymbol{x}_j(k)\big)_{j \in N_i}$ are independent Bernoulli random variables with parameter β. By taking $\delta = 1 - \frac{3t}{4\beta}$ in Proposition 1, we obtain that

$$\begin{aligned}
&\mathbb{P}\bigg(\frac{s}{t|N_i|}\sum_{j \in N_i}\bigg(\frac{t}{s} - \boldsymbol{x}_j(k)\bigg) \geq \frac{1}{4} \;\bigg|\; \boldsymbol{x}_i(k) = 1, |N_i|, |S_i|\bigg) \\
&\quad = \mathbb{P}\bigg(\sum_{j \in N_i} \boldsymbol{x}_j(k) \geq \frac{3t|N_i|}{4s} = \frac{3t}{4s\beta}|N_i|\beta \;\bigg|\; \boldsymbol{x}_i(k) = 1, |N_i|, |S_i|\bigg) \\
&\quad \leq \exp\bigg(-\frac{1}{2}\bigg(1 - \frac{3t}{4s\beta}\bigg)^2 |N_i|\beta\bigg) \leq \exp\bigg(-\frac{1}{2}\bigg(1 - \frac{3|S_i|}{4s}\bigg)^2 \frac{t|N_i|}{|S_i|}\bigg).
\end{aligned}$$

Therefore,

$$\mathbb{P}\bigg(\frac{s}{t|N_i|}\sum_{j \in N_i}\bigg(\frac{t}{s} - \boldsymbol{x}_j(k)\bigg) \geq \frac{1}{4} \;\bigg|\; \boldsymbol{x}_i(k) = 1, |N_i|, |S_i| \geq \frac{5}{4}s\bigg) \leq \exp\bigg(-\frac{t|N_i|}{640s}\bigg).$$

The claim directly follows from applying Proposition 6. □

The following two lemmas are the major building blocks of our proof. Intuitively, the probability that the neuron in GNN is activated when the true signal is 1 or the neuron is deactivated when the true signal is 0 is very small if the parameters satisfy certain conditions.

Lemma 7. *For all $k \in [d]$, we have when $n \geq 8^t m$,*

$$\mathbb{P}(\mathcal{A}_{i,k}^c \mid \boldsymbol{x}_i(k) = 0, |N_i|) \leq \exp\bigg(-\frac{t|N_i|}{24s}\bigg) + \exp\bigg(-\frac{t|N_i|}{32s^2\sigma^2}\bigg).$$

Proof. We further define the following two events:

$$B_0 \coloneqq \bigg\{\frac{s}{t|N_i|}\sum_{j \in N_i} \boldsymbol{x}_j(k) \leq \frac{1}{4}\bigg\} \quad \text{and} \quad B_1 \coloneqq \bigg\{\frac{s}{t|N_i|}\sum_{j \in N_i} \boldsymbol{\varepsilon}_j(k) \leq \frac{1}{4}\bigg\}. \tag{12}$$

Then we have $\mathcal{A}_{i,k} \supset B_0 \cap B_1$. By a union bound, we obtain

$$\mathbb{P}(\mathcal{A}_{i,k}^c \mid \boldsymbol{x}_i(k) = 0, |N_i|) \leq \mathbb{P}(B_0^c \mid \boldsymbol{x}_i(k) = 0, |N_i|) + \mathbb{P}(B_1^c \mid \boldsymbol{x}_i(k) = 0, |N_i|).$$

Using Lemma 5, we have

$$\mathbb{P}(B_0^c \mid \boldsymbol{x}_i(k) = 0, |N_i|) \leq \exp\bigg(-\frac{t|N_i|}{24s}\bigg).$$

And Lemma 4 gives

$$\mathbb{P}(B_1^c \mid \boldsymbol{x}_i(k) = 0, |N_i|) = \mathbb{P}(B_1^c \mid |N_i|) \leq \exp\bigg(-\frac{t^2|N_i|}{32s^2\sigma^2}\bigg).$$

The lemma is proved by combining the above displays. □

Lemma 8. *For all $k \in [d]$,*

$$\mathbb{P}(\mathcal{A}_{i,k} \mid \boldsymbol{x}_i(k) = 1, |N_i|) \leq \exp\bigg(-\frac{t|N_i|}{640s}\bigg) + \exp\bigg(-\frac{s}{175}\bigg) + \exp\bigg(-\frac{t^2|N_i|}{32s^2\sigma^2}\bigg).$$

Proof. We similarly define the two events:

$$B_0 \coloneqq \bigg\{\frac{s}{t|N_i|}\sum_{j \in N_i}\bigg(\frac{t}{s} - \boldsymbol{x}_j(k)\bigg) \leq \frac{1}{4}\bigg\} \quad \text{and} \quad B_1 \coloneqq \bigg\{-\frac{s}{t|N_i|}\sum_{j \in N_i} \boldsymbol{\varepsilon}_j(k) \leq \frac{1}{4}\bigg\}.$$

Rearranging the terms, we have $\mathcal{A}_{i,k}^c \supset B_0 \cap B_1$. Hence by a union bound

$$\mathbb{P}(\mathcal{A}_{i,k} \mid \boldsymbol{x}_i(k) = 1, |N_i|) \le \mathbb{P}(B_0^c \mid \boldsymbol{x}_i(k) = 1, |N_i|) + \mathbb{P}(B_1^c \mid \boldsymbol{x}_i(k) = 1, |N_i|).$$

Using Lemma 6, we arrive at

$$\mathbb{P}(B_0^c \mid \boldsymbol{x}_i(k) = 1, |N_i|) \le \exp\biggl(-\frac{t|N_i|}{640s}\biggr) + \exp\biggl(-\frac{s}{175}\biggr).$$

And applying Lemma 4 to $-\boldsymbol{\varepsilon}_i$'s we have

$$\mathbb{P}(B_1^c \mid \boldsymbol{x}_i(k) = 1, |N_i|) \le \exp\biggl(-\frac{t^2|N_i|}{32s^2\sigma^2}\biggr).$$

We obtain the claim by putting together the above displays. □

The next lemma combines the previous two and proves an upper bound for making a mistake in the vertex features.

Lemma 9. *For all $i \in [n]$ we have*

$$\mathbb{P}(\boldsymbol{z}_i \neq \boldsymbol{x}_i) \le 3d\exp\biggl(-\frac{mq}{5\cdot 2^{t+7}s}\biggr) + 3d\exp\biggl(-\frac{s}{175}\biggr) + d\exp\biggl(-\frac{mq}{2^{t+7}s^2\sigma^2}\biggr).$$

Proof. We first note that by a union bound

$$\mathbb{P}(\boldsymbol{z}_i \neq \boldsymbol{x}_i) \le \sum_{i=1}^{d} \mathbb{P}(\boldsymbol{z}_i(k) \neq \boldsymbol{x}_i(k)).$$

By the law of total probability, we have

$$\begin{aligned}\mathbb{P}(\boldsymbol{z}_i(k) \neq \boldsymbol{x}_i(k)) &= \mathbb{P}(\boldsymbol{z}_i(k) = 1 \mid \boldsymbol{x}_i(k) = 0) \times \mathbb{P}(\boldsymbol{x}_i(k) = 0) \\ &\quad + \mathbb{P}(\boldsymbol{z}_i(k) = 0 \mid \boldsymbol{x}_i(k) = 1) \times \mathbb{P}(\boldsymbol{x}_i(k) = 1)) \\ &\le \max\{\mathbb{P}(\boldsymbol{z}_i(k) = 1 \mid \boldsymbol{x}_i(k) = 0), \mathbb{P}(\boldsymbol{z}_i(k) = 0 \mid \boldsymbol{x}_i(k) = 1)\}\end{aligned}$$

By Proposition 6, using Lemma 7 and 3, we obtain that

$$\begin{aligned}\mathbb{P}(\boldsymbol{z}_i(k) = 1 \mid \boldsymbol{x}_i(k) = 0) &\le \mathbb{P}\biggl(\boldsymbol{z}_i(k) = 1 \biggm| \boldsymbol{x}_i(k) = 0, |N_i| \ge \frac{mq}{2^{t+2}}\biggr) + \mathbb{P}\biggl(|N_i| \ge \frac{mq}{2^{t+2}}\biggr) \\ &\le 3\exp\biggl(-\frac{mq}{2^{t+7}s}\biggr) + \exp\biggl(-\frac{mq}{2^{t+7}s^2\sigma^2}\biggr) + 2\exp\biggl(-\frac{s}{8}\biggr).\end{aligned}$$

Similarly with Lemma 8,

$$\begin{aligned}\mathbb{P}(\boldsymbol{z}_i(k) = 0 \mid \boldsymbol{x}_i(k) = 1) &\le \mathbb{P}\biggl(\boldsymbol{z}_i(k) = 0 \biggm| \boldsymbol{x}_i(k) = 1, |N_i| \ge \frac{mq}{2^{t+2}}\biggr) + \mathbb{P}\biggl(|N_i| \ge \frac{mq}{2^{t+2}}\biggr) \\ &\le 3\exp\biggl(-\frac{tmq}{5\cdot 2^{t+7}s}\biggr) + 3\exp\biggl(-\frac{s}{175}\biggr) + \exp\biggl(-\frac{t^2mq}{2^{t+7}s^2\sigma^2}\biggr).\end{aligned}$$

Therefore, by combining the above two displays, we obtain that

$$\mathbb{P}(\boldsymbol{z}_i(k) \neq \boldsymbol{x}_i(k)) \le 3\exp\biggl(-\frac{mq}{5\cdot 2^{t+7}s}\biggr) + 3\exp\biggl(-\frac{s}{175}\biggr) + \exp\biggl(-\frac{mq}{2^{t+7}s^2\sigma^2}\biggr).$$

The claim directly follows. □

With Lemma 9 in place, Theorem 1 directly follows from a union bound:

$$\begin{aligned}\mathbb{P}(\exists i \text{ s.t. } \boldsymbol{x}_i \neq \boldsymbol{z}_i) &\le \sum_{i=1}^{n} \mathbb{P}(\boldsymbol{x}_i \neq \boldsymbol{z}_i) \\ &\le nd\biggl(3\exp\biggl(-\frac{mq}{5\cdot 2^{t+7}s}\biggr) + 3\exp\biggl(-\frac{s}{175}\biggr) + \exp\biggl(-\frac{mq}{2^{t+7}s^2\sigma^2}\biggr)\biggr).\end{aligned}$$

This concludes our proof for the perfect recovery.

D PROOFS OF THE IMPOSSIBILITY RESULTS FOR GRAPH NEURAL NETWORKS

Lemma 10. *For all vertices $i \in [n]$,*

$$\mathbb{P}(\boldsymbol{z}_i(k) = 1 \mid \boldsymbol{x}_i(k) = 0, \boldsymbol{X}, |N_i|) \geq \Phi^c\bigg(\frac{t\sqrt{|N_i|}}{2s\sigma}\bigg)$$

where $\Phi^c(x) \coloneqq 1 - \Phi(x)$ is the upper tail of standard normal distribution.

Proof. Define $\mathcal{A}_{i,k}$ as in (10) by

$$\mathcal{A}_{i,k} \coloneqq \bigg\{\frac{s}{t|N_i|}\sum_{j\in N_i}\boldsymbol{x}_j(k) + \frac{s}{t|N_i|}\sum_{j\in N_i}\boldsymbol{\varepsilon}_j(k) \leq \frac{1}{2}\bigg\}.$$

We remark that

$$\begin{aligned}
\mathbb{P}(\boldsymbol{z}_i(k) = 1 \mid \boldsymbol{x}_i(k) = 0, \boldsymbol{X}, |N_i|) &= \mathbb{P}(\mathcal{A}_{i,k}^c \mid \boldsymbol{x}_i(k) = 0, \boldsymbol{X}, |N_i|)\\
&= \mathbb{P}\bigg(\frac{s}{t|N_i|}\sum_{j\in N_i}\boldsymbol{x}_j(k) + \frac{s}{t|N_i|}\sum_{j\in N_i}\boldsymbol{\varepsilon}_j(k) \geq \frac{1}{2}\,\bigg|\, \boldsymbol{x}_i(k) = 0, \boldsymbol{X}, |N_i|\bigg)\\
&\overset{(a)}{\geq} \mathbb{P}\bigg(\frac{s}{t|N_i|}\sum_{j\in N_i}\boldsymbol{\varepsilon}_j(k) \geq \frac{1}{2}\,\bigg|\, \boldsymbol{x}_i(k) = 0, \boldsymbol{X}, |N_i|\bigg)\\
&\overset{(b)}{=} \mathbb{P}\bigg(\frac{s}{t|N_i|}\sum_{j\in N_i}\boldsymbol{\varepsilon}_j(k) \geq \frac{1}{2}\,\bigg|\, |N_i|\bigg) = \Phi^c\bigg(\frac{t\sqrt{|N_i|}}{2s\sigma}\bigg),
\end{aligned}$$

where (a) is due to the fact that $\frac{s}{t|N_i|}\sum_{j\in N_i}\boldsymbol{x}_j(k) \geq 0$ is nonnegative and (b) is by independence of the noise. □

Lemma 11. *For a pair of vertices i and j,*

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{\delta}_i(k) - \boldsymbol{\delta}_j(k) = 1, \boldsymbol{\delta}_i'(k) - \boldsymbol{\delta}_j'(k) = 1 \mid \mathcal{M}_{i,j}^k, |N_i|, |N_j|, |N_i'|, |N_j'|)\\
&\quad\geq \frac{1}{4}\Phi^c\bigg(\frac{t\sqrt{|N_i|}}{2s\sigma}\bigg)\Phi^c\bigg(\frac{t\sqrt{|N_i'|}}{2s\sigma}\bigg)\bigg(1 - \exp\bigg(-\frac{t|N_j|}{24s}\bigg) - \exp\bigg(-\frac{t|N_j'|}{24s}\bigg)\bigg).
\end{aligned}$$

Proof. By conditional independence of the random variables,

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{\delta}_i(k) - \boldsymbol{\delta}_j(k) = 1, \boldsymbol{\delta}_i'(k) - \boldsymbol{\delta}_j'(k) = 1 \mid \mathcal{M}_{i,j}^k, |N_i|, |N_j|, |N_i'|, |N_j'|)\\
&\quad= \mathbb{E}_{\boldsymbol{X}}[\mathbb{P}(\boldsymbol{\delta}_i(k) = 1 \mid \mathcal{M}_{i,j}^k, \boldsymbol{X}, |N_i|)\mathbb{P}(\boldsymbol{\delta}_i'(k) = 1 \mid \mathcal{M}_{i,j}^k, \boldsymbol{X}, |N_i'|)\\
&\qquad\times \mathbb{P}(\boldsymbol{\delta}_j(k) = 0, \boldsymbol{\delta}_j'(k) = 0 \mid \mathcal{M}_{i,j}^k, \boldsymbol{X}, |N_j|, |N_j'|)]\\
&\quad\geq \Phi^c\bigg(\frac{t\sqrt{|N_i|}}{2s\sigma}\bigg)\Phi^c\bigg(\frac{t\sqrt{|N_i'|}}{2s\sigma}\bigg)\mathbb{P}(\boldsymbol{\delta}_j(k) = 0, \boldsymbol{\delta}_j'(k) = 0 \mid \mathcal{M}_{i,j}^k, |N_j|, |N_j'|)
\end{aligned}$$

where we used Lemma 10 in the inequality.

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{\delta}_j(k) = 0, \boldsymbol{\delta}_j'(k) = 0 \mid \mathcal{M}_{i,j}^k, |N_j|, |N_j'|)\\
&\quad\geq \mathbb{P}\bigg(\frac{s}{t|N_j|}\sum_{l\in N_j}\boldsymbol{x}_l(k) \leq \frac{1}{4}, \frac{s}{t|N_j'|}\sum_{l\in N_j'}\boldsymbol{x}_l(k) \leq \frac{1}{4},\\
&\qquad\quad \frac{s}{t|N_j|}\sum_{l\in N_j'}\boldsymbol{\varepsilon}_l(k) \leq \frac{1}{4}, \frac{s}{t|N_j'|}\sum_{l\in N_j'}\boldsymbol{\varepsilon}_l'(k) \leq \frac{1}{4}\,\bigg|\, \mathcal{M}_{i,j}^k, |N_j|, |N_j'|\bigg)\\
&\quad= \mathbb{P}\bigg(\frac{s}{t|N_j|}\sum_{l\in N_j}\boldsymbol{\varepsilon}_l(k) \leq \frac{1}{4}\,\bigg|\, |N_j|\bigg)\mathbb{P}\bigg(\frac{s}{t|N_j'|}\sum_{l\in N_j'}\boldsymbol{\varepsilon}_l'(k) \leq \frac{1}{4}\,\bigg|\, |N_j|\bigg)\\
&\qquad\times \mathbb{P}\bigg(\frac{s}{t|N_j|}\sum_{l\in N_j}\boldsymbol{x}_l(k) \leq \frac{1}{4}, \frac{s}{t|N_j'|}\sum_{l\in N_j'}\boldsymbol{x}_l(k) \leq \frac{1}{4}\,\bigg|\, \boldsymbol{x}_j(k) = 0, |N_j|, |N_j'|\bigg).
\end{aligned}$$

Since $\frac{s}{t|N_j|}\sum_{l\in N_j}\boldsymbol{\varepsilon}_l(k)$ is a sum of independent centered Gaussian random variables,

$$\mathbb{P}\bigg(\frac{s}{t|N_j|}\sum_{l\in N_j}\boldsymbol{\varepsilon}_l(k)\le\frac{1}{4}\;\bigg|\;|N_j|\bigg)\ge\mathbb{P}\bigg(\frac{s}{t|N_j|}\sum_{l\in N_j}\boldsymbol{\varepsilon}_l(k)\le 0\;\bigg|\;|N_j|\bigg)=\frac{1}{2}.$$

The same holds for $\frac{s}{t|N_j'|}\sum_{l\in N_j'}\boldsymbol{\varepsilon}_l'(k)$. By a union bound and Lemma 5,

$$\begin{aligned}\mathbb{P}\bigg(&\frac{s}{t|N_j|}\sum_{l\in N_j}\boldsymbol{x}_l(k)\le\frac{1}{4},\frac{s}{t|N_j'|}\sum_{l\in N_j'}\boldsymbol{x}_l(k)\le\frac{1}{4}\;\bigg|\;\boldsymbol{x}_j(k)=0,|N_j|,|N_j'|\bigg)\\&\ge 1-\mathbb{P}\bigg(\frac{s}{t|N_j|}\sum_{l\in N_j}\boldsymbol{x}_l(k)\ge\frac{1}{4}\;\bigg|\;\boldsymbol{x}_j(k)=0,|N_j|\bigg)\\&\quad-\mathbb{P}\bigg(\frac{s}{t|N_j'|}\sum_{l\in N_j'}\boldsymbol{x}_l(k)\ge\frac{1}{4}\;\bigg|\;\boldsymbol{x}_j(k)=0,|N_j'|\bigg)\\&\ge 1-\exp\bigg(-\frac{t|N_j|}{24s}\bigg)-\exp\bigg(-\frac{t|N_j'|}{24s}\bigg).\end{aligned}$$

Putting the above displays together, we hence proved the lemma. □

Denote $i\bowtie j$ the event

$$\{\langle\boldsymbol{z}_i,\boldsymbol{z}_j'\rangle+\langle\boldsymbol{z}_j,\boldsymbol{z}_i'\rangle\ge\langle\boldsymbol{z}_i,\boldsymbol{z}_i'\rangle+\langle\boldsymbol{z}_j,\boldsymbol{z}_j'\rangle\}.$$

It is clear that when $i\bowtie j$ happens, perfect recovery is not possible since by swapping i and j, the loss ℓ in (1) is always not bigger. The following proposition shows this and therefore concludes the proof for impossibility of perfect recovery.

Proposition 7. *Assume that* $\min\{s,\frac{qm}{s}\}\ge c$ *and* $\frac{qm}{\sigma^2s^2}\le C$ *for constants* $c,C>0$*. Then there is a constant* $\delta(c,C)>0$ *such that for all* $\delta'\in[0,\delta(c,C)]$

$$\mathbb{P}(i\bowtie j)\ge\delta'.$$

Proof. The definition of $i\bowtie j$ reads

$$\langle\boldsymbol{z}_i,\boldsymbol{z}_i'\rangle+\langle\boldsymbol{z}_j,\boldsymbol{z}_j'\rangle\le\langle\boldsymbol{z}_i,\boldsymbol{z}_j'\rangle+\langle\boldsymbol{z}_j,\boldsymbol{z}_i'\rangle$$

which by definition of $\boldsymbol{z}_i,\boldsymbol{z}_j$ and of $\boldsymbol{z}_i',\boldsymbol{z}_j'$ means that

$$\langle\boldsymbol{x}_i+\boldsymbol{\delta}_i,\boldsymbol{x}_i+\boldsymbol{\delta}_i'\rangle+\langle\boldsymbol{x}_j+\boldsymbol{\delta}_j,\boldsymbol{x}_j+\boldsymbol{\delta}_j'\rangle\le\langle\boldsymbol{x}_i+\boldsymbol{\delta}_i,\boldsymbol{x}_j+\boldsymbol{\delta}_j'\rangle+\langle\boldsymbol{x}_j+\boldsymbol{\delta}_j,\boldsymbol{x}_i+\boldsymbol{\delta}_i'\rangle,$$

where $\boldsymbol{\delta}_i:=\boldsymbol{z}_i-\boldsymbol{x}_i$ and $\boldsymbol{\delta}_j:=\boldsymbol{z}_j-\boldsymbol{x}_j$. Rearranging the terms, we have

$$\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2+\langle\boldsymbol{x}_i-\boldsymbol{x}_j,\boldsymbol{\delta}_i-\boldsymbol{\delta}_j\rangle+\langle\boldsymbol{x}_i-\boldsymbol{x}_j,\boldsymbol{\delta}_i'-\boldsymbol{\delta}_j'\rangle\le-\langle\boldsymbol{\delta}_i-\boldsymbol{\delta}_j,\boldsymbol{\delta}_i'-\boldsymbol{\delta}_j'\rangle.$$

Observe that

$$\begin{aligned}\langle\boldsymbol{\delta}_i-\boldsymbol{\delta}_j,\boldsymbol{\delta}_i'-\boldsymbol{\delta}_j'\rangle=&\underbrace{\sum_{k\in S_i\cup S_j}(\boldsymbol{\delta}_i(k)-\boldsymbol{\delta}_j(k))(\boldsymbol{\delta}_i'(k)-\boldsymbol{\delta}_j'(k))}_{D_0}\\&+\underbrace{\sum_{k\in(S_i\cup S_j)^c}(\boldsymbol{\delta}_i(k)-\boldsymbol{\delta}_j(k))(\boldsymbol{\delta}_i'(k)-\boldsymbol{\delta}_j'(k))}_{D_1}.\end{aligned}$$

We will bound both terms on the right-hand side. In this goal, we recall the event $\mathcal{E}_i=\{\frac{s}{2}\le|S_i|\le 2s\}$ defined in (4). Using (5) we know that the event $\mathcal{E}_i\cap\mathcal{E}_j$ happens with probability at least $1-2e^{-s/8}$. We notice that when $cE_i\cap\mathcal{E}_j$ holds we have

$$\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\le|S_i|+|S_j|\le 4s.$$

By definition of $\boldsymbol{x}_i$'s and $\boldsymbol{z}_i$'s, we know that

$$\langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{\delta}_i - \boldsymbol{\delta}_j \rangle \le 2\sum_{k=1}^{d} |\boldsymbol{x}_i(k) - \boldsymbol{x}_j(k)| \overset{(a)}{=} 2\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \le 2(|S_i| + |S_j|) \le 8s$$

where to obtain (a) we used the fact that as $(\boldsymbol{x}_i(k))$ and $(\boldsymbol{x}_j(k))$ are Bernoulli random variables $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = \sum_{k=1}^{d} |\boldsymbol{x}_i(k) - \boldsymbol{x}_j(k)|$. Recall that S_i is the support of $\boldsymbol{x}_i$. First note that for every k we have $|\boldsymbol{\delta}_i(k) - \boldsymbol{\delta}_j(k)| \overset{a.s}{\le} 2$. Hence we obtain that on event $\mathcal{E}_i \cap \mathcal{E}_j$

$$|D_0| \le 4|S_i \cup S_j| \le 4(|S_i| + |S_j|) \le 16s.$$

Putting the above together, this implies that when $\mathcal{E}_i \cap \mathcal{E}_j$ holds we have

$$\big| \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 + \langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{\delta}_i - \boldsymbol{\delta}_j \rangle + \langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{\delta}'_i - \boldsymbol{\delta}'_j \rangle + D_0 \big| \le 28s. \tag{13}$$

We now aim at bounding D_1. In this goal, we let $\boldsymbol{b} \in \{-1, 0, 1\}^d$ be the vector such that

$$\boldsymbol{b}(k) := \begin{cases} (\boldsymbol{\delta}_i(k) - \boldsymbol{\delta}_j(k))(\boldsymbol{\delta}'_i(k) - \boldsymbol{\delta}'_j(k)) & \text{if } k \in (S_1 \cup S_2)^c. \\ 0 & \text{otherwise.} \end{cases}$$

Moreover for $k \in (S_1 \cup S_2)^c$, by definition we have $\boldsymbol{x}_i(k) = \boldsymbol{x}_j(k) = 0$. For ease of notation we define the events

$$\mathcal{M}_{i,j}^k := \{\boldsymbol{x}_i(k) = \boldsymbol{x}_j(k) = 0\}.$$

We will first see that $\mathbb{P}(\boldsymbol{b}(k) = 1 \mid \mathcal{M}_{i,j}^k) = \mathbb{P}(\boldsymbol{b}(k) = -1 \mid \mathcal{M}_{i,j}^k)$ for all $k \in (S_1 \cup S_2)^c$. We will then lower bound the probability of $\boldsymbol{b}(k) = \pm 1$ conditioned on a few events.

Recall the definition of $\mathcal{E}_i, \mathcal{E}_j$ in (4) and the events defined in (6), and let

$$\mathcal{F}_i := \left\{ \frac{1}{2^{t+2}} \cdot mq \le |N_i| \le 2^{t+2} e^t \cdot mq \right\} \quad \text{and} \quad \mathcal{F}'_i := \left\{ \frac{1}{2^{t+2}} \cdot mq \le |N'_i| \le 2^{t+2} e^t \cdot mq \right\}.$$

Further let $\mathcal{H}_{i,j} := \mathcal{E}_i \cap \mathcal{E}_j \cap \mathcal{F}_i \cap \mathcal{F}_j \cap \mathcal{F}'_i \cap \mathcal{F}'_j$.

As conditionally on $\mathcal{M}_{i,j}^k \cap \mathcal{H}_{i,j}$ the laws of $\boldsymbol{\delta}_i(k), \boldsymbol{\delta}_j(k) \in \{0, 1\}$ are the same, by symmetry we have

$$\mathbb{P}(\boldsymbol{\delta}_i(k) - \boldsymbol{\delta}_j(k) = 1 \mid \mathcal{M}_{i,j}^k, \mathcal{H}_{i,j}) = \mathbb{P}(\boldsymbol{\delta}_i(k) - \boldsymbol{\delta}_j(k) = -1 \mid \mathcal{M}_{i,j}^k, \mathcal{H}_{i,j}).$$

By using Lemma 11, we obtain that

$$\begin{aligned} p &:= \mathbb{P}(\boldsymbol{b}(k) = 1 \mid \mathcal{M}_{i,j}^k, \mathcal{H}_{i,j}) \\ &= \mathbb{P}(\boldsymbol{\delta}_i(k) - \boldsymbol{\delta}_j(k) = 1, \boldsymbol{\delta}'_i(k) - \boldsymbol{\delta}'_j(k) = 1 \mid \mathcal{M}_{i,j}^k, \mathcal{H}_{i,j}) \\ &\quad + \mathbb{P}(\boldsymbol{\delta}_i(k) - \boldsymbol{\delta}_j(k) = -1, \boldsymbol{\delta}'_i(k) - \boldsymbol{\delta}'_j(k) = -1 \mid \mathcal{M}_{i,j}^k, \mathcal{H}_{i,j}) \\ &\ge \frac{1}{2} \Phi^c \left(\frac{c_t \sqrt{mq}}{s\sigma} \right)^2 \left(1 - 2\exp\left(-\frac{C_t mq}{s} \right) \right) \end{aligned}$$

for constants $c_t, C_t > 0$ that only depend on t. Since for $k \in (S_i \cup S_j)^c$ the entries $\boldsymbol{x}_{i'}(k), i' \in N_i$ is independent of N_i hence $\boldsymbol{z}_i(k)$ only depend on the size of the neighborhood. Therefore, $\boldsymbol{b}(k)$'s are independent of each other. Applying Proposition 4 to $\sum_{k \in (S_1 \cup S_2)^c} \boldsymbol{b}(k)$ we obtain that

$$\mathbb{P}\left(\sum_{k \in (S_1 \cup S_2)^c} \boldsymbol{b}(k) \ge 28s \;\middle|\; \mathcal{H}_{i,j} \right) \ge c_\beta \exp\left(-C_\beta \frac{s^2}{p(d - 4s)} \right).$$

Further use the bound (13), we have that

$$\begin{aligned} \mathbb{P}(i \bowtie j) = \mathbb{P}(i \bowtie j \mid \mathcal{H}_{i,j}) \mathbb{P}(\mathcal{H}_{i,j}) &\ge \mathbb{P}\left(\sum_{k \in (S_1 \cup S_2)^c} \boldsymbol{b}(k) \ge 28s \;\middle|\; \mathcal{H}_{i,j} \right) \mathbb{P}(\mathcal{H}_{i,j}) \\ &\ge c_\beta \exp\left(-C_\beta \frac{s^2}{p(d - 4s)} \right) \mathbb{P}(\mathcal{H}_{i,j}). \end{aligned}$$

By a union bound, we know that $\mathbb{P}(\mathcal{H}_{i,j})$ is bounded away from 0 if for a constant $c > 0$,

$$\min\{s, mq\} \geq c.$$

Therefore, $\mathbb{P}(\bowtie)$ is bounded away from 0 if p is bounded away from 0 which is satisfied if

$$\frac{mq}{s} \geq c' \quad \text{and} \quad \frac{mq}{s^2\sigma^2} \leq C$$

for constants $c', C > 0$. The lemma is hence proved. □

E PROOFS OF THE IMPOSSIBILITY RESULTS FOR NOISY FEATURE ALIGNMENT

In Proposition 9 we will prove that if $\sigma^2 \geq 4s\left(1 + \frac{K}{n}\right)^{-2}$ then the probability of perfect recovery is bounded away from 0. We state the following information-theoretic proposition before we move to the proof.

Proposition 8. *Let $k \leq n$ be any arbitrary integer and $\pi \in \mathcal{P}(n)$ be an arbitrary permutation. Let $\pi(\boldsymbol{Y}) = (\boldsymbol{y}_{\pi(1)}, \ldots, \boldsymbol{y}_{\pi(k)})$ be the noisy observations of $\pi(\boldsymbol{X}) = (\boldsymbol{x}_{\pi(1)}, \ldots, \boldsymbol{x}_{\pi(k)})$ as defined in Definition 2. Choose $\boldsymbol{G} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_k) \in \mathbb{R}^{k \times d}$ to be an independent Gaussian random matrix with i.i.d. entries $\boldsymbol{g}_i(j) \sim \mathcal{N}(0, \sigma^2)$. Then, the total variation distance satisfies*

$$\mathrm{TV}(\pi(\boldsymbol{Y}), \boldsymbol{G} \mid \boldsymbol{X}) \leq \frac{1}{2\sigma}\|\mathrm{vec}(\boldsymbol{X})\|$$

where we denoted

$$\mathrm{TV}(\pi(\boldsymbol{Y}), \boldsymbol{G} \mid \boldsymbol{X}) := \sup_{A \in \mathcal{B}(\mathbb{R}^{n \times d})} |\mathbb{P}(\pi(\boldsymbol{Y}) \in A \mid \boldsymbol{X}) - \mathbb{P}(\boldsymbol{G} \in A)|.$$

Proof. By Pinsker's inequality, we know that

$$\mathrm{TV}(\pi(\boldsymbol{Y}), \boldsymbol{G} \mid \boldsymbol{X}) \leq \sqrt{\frac{1}{2}\mathrm{KL}(\pi(\boldsymbol{Y}) \parallel \boldsymbol{G} \mid \boldsymbol{X})}.$$

We remark that by definition of $\pi(\boldsymbol{Y})$, we have $\mathrm{vec}(\pi(\boldsymbol{Y})) \sim \mathcal{N}(\mathrm{vec}(\boldsymbol{X}_\pi), \sigma^2 \boldsymbol{I}_{kd})$. Hence using the formula for the KL divergence between two Gaussian vectors (see, e.g., (Wainwright, 2019, Exercise 15.13(b))), we have

$$\mathrm{KL}(\pi(\boldsymbol{Y}) \parallel \boldsymbol{G} \mid \boldsymbol{X}) = \frac{1}{2\sigma^2}\|\mathrm{vec}(\pi(\boldsymbol{X}))\|^2 = \frac{1}{2\sigma^2}\|\mathrm{vec}(\boldsymbol{X})\|^2.$$

The claim is hence proved. □

With the previous proposition in place, we are ready to prove the results. First denote by $i \bowtie j$ the following event:

$$\{\langle \boldsymbol{y}_i, \boldsymbol{y}'_j \rangle + \langle \boldsymbol{y}_j, \boldsymbol{y}'_i \rangle \geq \langle \boldsymbol{y}_i, \boldsymbol{y}'_i \rangle + \langle \boldsymbol{y}_j, \boldsymbol{y}'_j \rangle\}.$$

Note that $i \bowtie j$ is equivalent to

$$\{\|\boldsymbol{y}_i - \boldsymbol{y}'_j\|^2 + \|\boldsymbol{y}_i - \boldsymbol{y}'_j\|^2 \leq \|\boldsymbol{y}_i - \boldsymbol{y}'_i\|^2 + \|\boldsymbol{y}_j - \boldsymbol{y}'_j\|^2\}.$$

Hence when $i \bowtie j$ happens, perfect recovery is not possible since if we swap i and j, the loss ℓ in (3) is not increased.

Proposition 9. *Let $\tilde{\pi}$ be the solution of (3). If $\sigma^2 \geq 2s\left(1 + \frac{K}{n}\right)^{-2}$ for any $K > 4$ we have that*

$$\mathbb{P}(\tilde{\pi} = \pi^*) \leq e^{-\frac{K-4}{4}}.$$

Hence if $\sigma^2 \gg s$ we have that the probability of perfect recovery will converge to 0.

Proof. First by triangle inequality and Proposition 8 we have

$$\begin{aligned}&\mathrm{TV}((\boldsymbol{y}_1, \boldsymbol{y}_2), (\boldsymbol{y}_2, \boldsymbol{y}_1) \mid \boldsymbol{x}_{1:2})\\ &\quad\le \mathrm{TV}((\boldsymbol{y}_1, \boldsymbol{y}_2), (\boldsymbol{g}_1, \boldsymbol{g}_2) \mid \boldsymbol{x}_{1:2}) + \mathrm{TV}((\boldsymbol{y}_2, \boldsymbol{y}_1), (\boldsymbol{g}_1, \boldsymbol{g}_2) \mid \boldsymbol{x}_{1:2})\\ &\quad\le \frac{1}{\sigma}(\|\boldsymbol{x}_1\|^2 + \|\boldsymbol{x}_2\|^2)^{1/2}.\end{aligned}$$

This directly implies that if we write the following set

$$A \coloneqq \{(\boldsymbol{a}, \boldsymbol{b}) : \langle \boldsymbol{a}, \boldsymbol{y}_2' \rangle + \langle \boldsymbol{b}, \boldsymbol{y}_1' \rangle \ge \langle \boldsymbol{a}, \boldsymbol{y}_1' \rangle + \langle \boldsymbol{b}, \boldsymbol{y}_2' \rangle\}$$

then we have

$$|\mathbb{P}((\boldsymbol{y}_1, \boldsymbol{y}_2) \in A \mid \boldsymbol{x}_{1:2}) - \mathbb{P}((\boldsymbol{y}_2, \boldsymbol{y}_1) \in A \mid \boldsymbol{x}_{1:2})| \le \frac{1}{\sigma}(\|\boldsymbol{x}_1\|^2 + \|\boldsymbol{x}_2\|^2)^{1/2}.$$

Now by Jensen inequality we have

$$\begin{aligned}&|\mathbb{P}((\boldsymbol{y}_1, \boldsymbol{y}_2) \in A) - \mathbb{P}((\boldsymbol{y}_2, \boldsymbol{y}_1) \in A)|\\ &\quad\le \mathbb{E}[|\mathbb{P}((\boldsymbol{y}_1, \boldsymbol{y}_2) \in A \mid \boldsymbol{x}_{1:2}) - \mathbb{P}((\boldsymbol{y}_2, \boldsymbol{y}_1) \in A \mid \boldsymbol{x}_{1:2})|]\\ &\quad\le \frac{1}{\sigma}\mathbb{E}[(\|\boldsymbol{x}_1\|^2 + \|\boldsymbol{x}_2\|^2)^{1/2}] \le \frac{1}{\sigma}\sqrt{\mathbb{E}[\|\boldsymbol{x}_1\|^2] + \mathbb{E}[\|\boldsymbol{x}_2\|^2]} \overset{(a)}{=} \frac{\sqrt{2s}}{\sigma}\end{aligned} \tag{14}$$

where to get (a) we used the fact that $\|\boldsymbol{x}_1\|^2, \|\boldsymbol{x}_2\|^2 \sim \mathrm{Binom}(d, \frac{s}{d})$.

Now note that for continuous random variables $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ and $(\boldsymbol{y}_1', \boldsymbol{y}_2')$,

$$\mathbb{P}(\langle \boldsymbol{y}_1, \boldsymbol{y}_2' \rangle + \langle \boldsymbol{y}_2, \boldsymbol{y}_1' \rangle = \langle \boldsymbol{y}_1, \boldsymbol{y}_1' \rangle + \langle \boldsymbol{y}_2, \boldsymbol{y}_2' \rangle) = 0$$

The events $\{(\boldsymbol{y}_1, \boldsymbol{y}_2) \in A\}$ and $\{(\boldsymbol{y}_2, \boldsymbol{y}_1) \in A\}$ are disjoint except for a zero-probability event. This implies that

$$1 = \mathbb{P}((\boldsymbol{y}_1, \boldsymbol{y}_2) \in A) + \mathbb{P}((\boldsymbol{y}_2, \boldsymbol{y}_1) \in A) \overset{(a)}{\le} 2\mathbb{P}((\boldsymbol{y}_1, \boldsymbol{y}_2) \in A) + \frac{\sqrt{2s}}{\sigma}$$

where (a) is a consequence of (14). Hence when $\sigma^2 \ge 2s\big(1 + \frac{K}{n}\big)^{-2}$ for some $\delta > 0$, we obtain that

$$\mathbb{P}((\boldsymbol{y}_1, \boldsymbol{y}_2) \in A) \ge \frac{K}{2n}.$$

This implies that

$$\mathbb{P}(\tilde{\pi}(1) \ne 1 \text{ or } \tilde{\pi}(2) \ne 2) \ge \frac{K}{2n}.$$

Using the fact that the random variables $(\mathbb{1}\{\tilde{\pi}(i) \ne i \text{ or } \tilde{\pi}(i+1) \ne i+1\})_{\substack{i \text{ odd}\\ i \le n-1}}$ are independent we obtain that

$$\begin{aligned}\mathbb{P}(\exists i \text{ s.t. } \tilde{\pi}(i) \ne i) &= \mathbb{P}(\exists i \text{ odd s.t. } \tilde{\pi}(i) \ne i \text{ or } \tilde{\pi}(i+1) \ne i+1)\\ &\ge 1 - \left(1 - \frac{K}{2n}\right)^{n/2-2} \ge 1 - \exp\left(-\frac{K-4}{4}\right).\end{aligned}$$

Hence we proved the desired result. □

The following proposition complements the previous proposition and shows that when $\sigma^2 \le s$ perfect recovery is still not possible if $\frac{\sigma^2\sqrt{d}}{s}$ is sufficiently large.

Proposition 10. *Suppose that $\frac{1}{4}\sigma^2 \le s \le \frac{d}{2}$. Assume that $\tilde{\pi}$ is the solution of* (3)*. If $\sigma^2 \ge \frac{\sqrt{21}s}{\sqrt{d}}$ and then*

$$\mathbb{P}(\tilde{\pi} \ne \pi^*) \ge \frac{1}{8}.$$

If $s \gg b_n$ and $\sigma^2 \gg \frac{s}{\sqrt{db_n}}$ for a sequence (b_n) diverging to infinity at a rate slower than $b_n = O(\log n)$ then

$$\lim_{n\to\infty} \mathbb{P}(\tilde{\pi} = \pi^*) = 0.$$

Proof. Recall that $i \bowtie j$ if

$$\langle \boldsymbol{y}_i, \boldsymbol{y}_i' \rangle + \langle \boldsymbol{y}_j, \boldsymbol{y}_j' \rangle \le \langle \boldsymbol{y}_i, \boldsymbol{y}_j' \rangle + \langle \boldsymbol{y}_j, \boldsymbol{y}_i' \rangle;$$

which by definition of $\boldsymbol{y}_i, \boldsymbol{y}_j$ and of $\boldsymbol{y}_i', \boldsymbol{y}_j'$ means that

$$\langle \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i, \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i' \rangle + \langle \boldsymbol{x}_j + \boldsymbol{\varepsilon}_j, \boldsymbol{x}_j + \boldsymbol{\varepsilon}_j' \rangle \le \langle \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i, \boldsymbol{x}_j + \boldsymbol{\varepsilon}_j' \rangle + \langle \boldsymbol{x}_j + \boldsymbol{\varepsilon}_j, \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i' \rangle.$$

Rearranging the terms yields $i \bowtie j$ if and only if

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 + \langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{\varepsilon}_i - \boldsymbol{\varepsilon}_j + \boldsymbol{\varepsilon}_i' - \boldsymbol{\varepsilon}_j' \rangle \le -\langle \boldsymbol{\varepsilon}_i - \boldsymbol{\varepsilon}_j, \boldsymbol{\varepsilon}_i' - \boldsymbol{\varepsilon}_j' \rangle.$$

Define the events

$$\begin{aligned} A_0 &:= \Big\{ -\langle \boldsymbol{\varepsilon}_i - \boldsymbol{\varepsilon}_j, \boldsymbol{\varepsilon}_i' - \boldsymbol{\varepsilon}_j' \rangle \ge (a_1 + a_2) s \Big\}, \\ A_1 &:= \Big\{ \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \le a_1 s \Big\}, \\ A_2 &:= \Big\{ \langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{\varepsilon}_i - \boldsymbol{\varepsilon}_j + \boldsymbol{\varepsilon}_i' - \boldsymbol{\varepsilon}_j' \rangle \le a_2 s \Big\}. \end{aligned} \tag{15}$$

We first remark that

$$\begin{aligned} \mathbb{P}(i \bowtie j) &\ge \mathbb{P}(A_0 \cap A_1 \cap A_2) = \mathbb{P}(A_0) - \mathbb{P}(A_0 \cap A_1^c) - \mathbb{P}(A_0 \cap A_1 \cap A_2^c) \\ &\ge \mathbb{P}(A_0) - \mathbb{P}(A_1^c) - \mathbb{P}(A_1 \cap A_2^c) \ge \mathbb{P}(A_0) - \mathbb{P}(A_1^c) - \mathbb{P}(A_2^c \mid A_1). \end{aligned}$$

Hence to obtain the desired result we will bound $\mathbb{P}(A_0)$ from below and $\mathbb{P}(A_1^c), \mathbb{P}(A_2^c \mid A_1)$ from above.

We first focus on the first result. In this goal, we remark that $\langle \boldsymbol{\varepsilon}_i - \boldsymbol{\varepsilon}_j, \boldsymbol{\varepsilon}_i' - \boldsymbol{\varepsilon}_j' \rangle = \sum_{k=1}^d (\boldsymbol{\varepsilon}_i(k) - \boldsymbol{\varepsilon}_j(k))(\boldsymbol{\varepsilon}_i'(k) - \boldsymbol{\varepsilon}_j'(k))$ is a sum of i.i.d. random variables. Hence to lower bound $\mathbb{P}(A_0)$ we can use the central limit theorem. We choose $a_1 = 7$ and $a_2 = 14$ in (15). Moreover, for all $k \in [d]$ we define

$$\boldsymbol{b}(k) := (\boldsymbol{\varepsilon}_i(k) - \boldsymbol{\varepsilon}_j(k))(\boldsymbol{\varepsilon}_i'(k) - \boldsymbol{\varepsilon}_j'(k)).$$

By definition of $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\varepsilon}_j$, we know that

$$\begin{aligned} \mathbb{E}[\boldsymbol{b}(k)] &= 0, \\ \mathbb{E}[\boldsymbol{b}(k)^2] &= \mathbb{E}[(\boldsymbol{\varepsilon}_i(k) - \boldsymbol{\varepsilon}_j(k))^2]\mathbb{E}[(\boldsymbol{\varepsilon}_i'(k) - \boldsymbol{\varepsilon}_j'(k))^2] = 4\sigma^4, \\ \mathbb{E}[|\boldsymbol{b}(k)|^3] &= \mathbb{E}[|\boldsymbol{\varepsilon}_i(k) - \boldsymbol{\varepsilon}_j(k)|^3]\mathbb{E}[|\boldsymbol{\varepsilon}_i'(k) - \boldsymbol{\varepsilon}_j'(k)|^3] = \frac{64\sigma^6}{\pi}. \end{aligned}$$

Therefore, by Berry–Esseen theorem (Tyurin, 2010), we have

$$\left| \mathbb{P}(A_0) - \Phi^c\left(\frac{21s}{2\sigma^2\sqrt{d}} \right) \right| \le \frac{4}{\pi\sqrt{d}}.$$

Since $d \ge (21s/\sigma^2)^2$ and $s \ge 4\sigma^2$, we obtain that

$$\mathbb{P}(A_0) \ge \Phi^c\left(\frac{21s}{2\sigma^2\sqrt{d}} \right) - \frac{4}{\pi\sqrt{d}} \ge \Phi^c\left(\frac{1}{4} \right) - \frac{4}{42\pi} \ge \frac{3}{10} - \frac{1}{30} = \frac{4}{15}.$$

We now aim to upper bound $\mathbb{P}(A_1^c)$. Since

$$\mathbb{E}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = 2d \cdot \frac{s}{d} \cdot \left(1 - \frac{s}{d}\right) = 2s\left(1 - \frac{s}{d}\right),$$

using $\delta = 6$ in Proposition 1 yields

$$\mathbb{P}(A_1^c) \le \exp\left(-\frac{\delta^2 \mathbb{E}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2 + \delta} \right) = \exp\left(-9s\left(1 - \frac{s}{d}\right) \right) \le \exp\left(-\frac{9}{2} \right)$$

for $d \geq 2s$.

Finally we remark that conditional on $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, the random variable $\langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{\varepsilon}_i - \boldsymbol{\varepsilon}_j + \boldsymbol{\varepsilon}'_i - \boldsymbol{\varepsilon}'_j \rangle$ is distributed as $\mathcal{N}(0, 4\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\sigma^2)$. Hence, by Gaussian concentration, we obtain that

$$\mathbb{P}(A_2^c \mid \boldsymbol{x}_i, \boldsymbol{x}_j) \leq \exp\left(-\frac{49s^2}{2\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\sigma^2}\right) \leq \exp\left(-\frac{98s}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}\right).$$

Then, we have

$$\mathbb{P}(A_2^c \mid A_1) \leq \exp(-7).$$

Putting them together, we have when $d \geq (21s)^2/\sigma^4$

$$\mathbb{P}(i \bowtie j) \geq \frac{4}{15} - \exp\left(-\frac{9}{2}\right) - \exp(-7) \geq \frac{1}{4}.$$

Now by a union bound argument we know that

$$\mathbb{P}(i \bowtie j) \leq \mathbb{P}(\hat{\pi}(i) \neq i) + \mathbb{P}(\hat{\pi}(j) \neq j) = 2\mathbb{P}(\hat{\pi}(i) \neq i).$$

This directly implies the first result.

We will now establish the second one. In this goal we note that as we have assumed that $s \gg b_n$ we know that

$$\max\left(\frac{\sqrt{sb_n}}{16(1 + \sqrt{s/b_n})}, \frac{\sqrt{sb_n}}{2(2\sqrt{s/b_n} + 1)}\right) \sim \max\left(\frac{b_n}{16}, \frac{b_n}{4}\right),$$
$$\frac{11(1 + \sigma + \sqrt{s/b_n})^2 sb_n}{2d\sigma^4} \sim \frac{11s^2}{2d\sigma^4}.$$

Moreover we have assumed that $b_n \gg \frac{s^2}{d\sigma^4}$ hence there exist constants $C_1, C_2 > 0$ such that for n large enough we have

$$\exp\left(-\frac{C_2^2\sqrt{sb_n}}{16(C_1 + \sqrt{s/b_n})}\right) + \exp\left(-\frac{\sqrt{sb_n}C_1^2}{2(2\sqrt{s/b_n} + C_1)}\right)$$
$$\geq \frac{1}{2}\exp\left(-\frac{11(C_1 + C_2\sigma + \sqrt{s/b_n})^2 sb_n}{2d\sigma^4}\right)$$

and

$$\exp\left(-\frac{11(C_1 + C_2\sigma + \sqrt{s/b_n})^2 sb_n}{2d\sigma^4}\right) \leq 1 - \frac{\sqrt{3}}{2}.$$

Set $a_1 \coloneqq C_1\frac{\sqrt{b_n}}{\sqrt{s}} + 1$ and $a_2 \coloneqq C_2\sigma\frac{\sqrt{b_n}}{\sqrt{s}}$. Using Proposition 3 with $u = (a_1 + a_2)s/\sqrt{d}$ we have

$$\begin{aligned}
\mathbb{P}(A_0) &\geq \left(1 - \exp\left(-\frac{(a_1 + a_2)^2 s^2}{d\sigma^4}\right)\right)^2 \exp\left(-\frac{11(a_1 + a_2)^2 s^2}{2d\sigma^4}\right) \\
&\geq \left(1 - \exp\left(-\frac{(C_1 + C_2\sigma + \sqrt{s/b_n})^2 sb_n}{d\sigma^4}\right)\right)^2 \exp\left(-\frac{11(C_1 + C_2\sigma + \sqrt{s/b_n})^2 sb_n}{2d\sigma^4}\right) \\
&\sim \left(1 - \exp\left(-\frac{s^2}{d\sigma^4}\right)\right)^2 \exp\left(-\frac{11s^2}{2d\sigma^4}\right).
\end{aligned}$$

Write $\delta = a_1 - 1$ then we have

$$\begin{aligned}
\mathbb{P}(A_1^c) &\leq \exp\left(-\frac{\delta^2\mathbb{E}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2 + \delta}\right) = \exp\left(-\frac{(a_1 - 1)^2}{1 + a_1}s\left(1 - \frac{s}{d}\right)\right) \\
&\overset{(a)}{\leq} \exp\left(-\frac{(a_1 - 1)^2}{2(1 + a_1)}s\right) \leq \exp\left(-\frac{\sqrt{sb_n}C_1^2}{2(2\sqrt{s/b_n} + C_1)}\right) \sim \exp\left(-\frac{b_nC_1^2}{4}\right)
\end{aligned}$$

where (a) is due to the assumption that $d \geq 2s$. Now once again, conditionally on $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, the random variable $\langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{\varepsilon}_i - \boldsymbol{\varepsilon}_j + \boldsymbol{\varepsilon}'_i - \boldsymbol{\varepsilon}'_j \rangle$ is distributed as $\mathcal{N}(0, 4\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\sigma^2)$. Hence we also have

$$\mathbb{P}(A_2^c \mid \boldsymbol{x}_i, \boldsymbol{x}_j) \leq \exp\biggl(-\frac{a_2^2 s^2}{16\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\sigma^2}\biggr)$$

and therefore

$$\mathbb{P}(A_2^c \mid A_1) \leq \exp\biggl(-\frac{a_2^2 s}{16 a_1 \sigma^2}\biggr) \leq \exp\biggl(-\frac{C_2^2\sqrt{s b_n}}{16(C_1 + \sqrt{s/b_n})}\biggr) \sim \exp\biggl(-\frac{C_2^2 b_n}{16}\biggr).$$

Hence, for n large enough we have

$$\begin{aligned}\mathbb{P}(i \bowtie j) \geq{}& \exp\biggl(-\frac{11(C_1 + C_2\sigma + \sqrt{s/b_n})^2 s b_n}{2d\sigma^4}\biggr)\\ &\times \biggl(\biggl(1 - \exp\biggl(-\frac{(C_1 + C_2\sigma + \sqrt{s/b_n})^2 s b_n}{d\sigma^4}\biggr)\biggr)^2 - \frac{1}{2}\biggr)\\ \geq{}& \frac{1}{4}\exp\biggl(-\frac{11(C_1 + C_2\sigma + \sqrt{s/b_n})^2 s b_n}{2d\sigma^4}\biggr) \sim \frac{1}{4}\exp\biggl(-\frac{11 s^2}{2d\sigma^4}\biggr).\end{aligned}$$

Moreover for n large enough we have

$$\begin{aligned}\mathbb{P}(\exists i, j \text{ s.t. } i \bowtie j) &\geq \mathbb{P}(\exists \text{ odd } i \leq n-1 \text{ s.t. } i \bowtie (i+1))\\ &\geq 1 - (1 - \mathbb{P}(1 \bowtie 2))^{\lfloor (n-1)/2 \rfloor}\\ &\geq 1 - \biggl(1 + \frac{n}{8}\exp\biggl(-\frac{11(C_1 + C_2\sigma + \sqrt{s/b_n})^2 s b_n}{2d\sigma^4}\biggr)\biggr)^{-1}.\end{aligned}$$

We have assumed that $\frac{s^2}{2d\sigma^4} \ll b_n = O(\log(n))$ which implies that

$$\biggl(1 + \frac{n}{8}\exp\biggl(-\frac{11(C_1 + C_2\sigma + \sqrt{s/b_n})^2 s b_n}{2d\sigma^4}\biggr)\biggr)^{-1} \sim \frac{1}{4}\exp\biggl(-\frac{11 s^2}{2d\sigma^4}\biggr) = o(1).$$

Hence we obtain the desired result. □

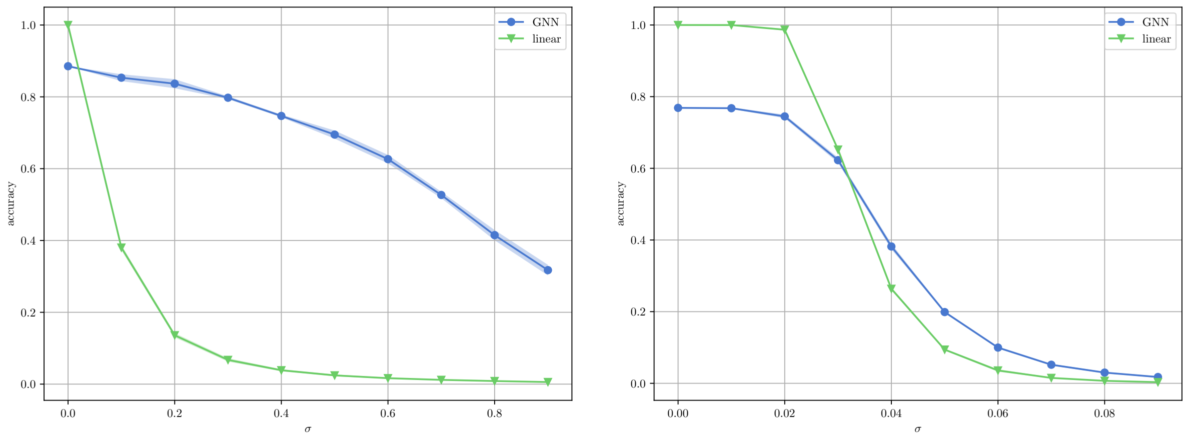
F ADDITIONAL EXPERIMENTS

We include experiments on two additional datasets: Marvel Universe Social Graph (`https://syntagmatic.github.io/marvel/`) and PubMed Diabetes (Namata et al., 2012). The Marvel dataset consists of Marvel characters (heroes) as nodes in the graph. The node features are the comic issues in which the hero appeared. Two heroes are connected by an edge if they appeared in the same comic issue. This follows our definition of a random intersection graph. The PubMed dataset is created similarly to Cora and CiteSeer, but with more articles and fewer features. Each publication is described by a TF/IDF weighted word vector. A summary of the two datasets can be found in Table 3.

Table 3: Summary of datasets.

Dataset	# Vertices	# Edges	# Features
Marvel	$6,444$	$574,467$	$12,849$
PubMed	$19,717$	$88,648$	500

We again add independent Gaussian noise to the node features and find the matching using the linear method and the GNN. The results are shown in Figure 5. Similar trends can be observed in the plots: the linear method fails when the feature noise is large, while the GNN can tolerate larger amounts of noise with the help of the graph structure.



(a) Marvel

(b) PubMed

Figure 5: Impact of the noise parameter on real-world datasets.