When Can Two Unlabeled Networks Be Aligned Under Partial Overlap?

Ehsan Kazemi*, Lyudmila Yartseva* and Matthias Grossglauser*

***Abstract*— Network alignment refers to the problem of matching the vertex sets of two unlabeled graphs, which can be viewed as a generalization of the classic graph isomorphism problem. Network alignment has applications in several fields, including social network analysis, privacy, pattern recognition, computer vision, and computational biology. A number of heuristic algorithms have been proposed in these fields. Recent progress in the analysis of network alignment over stochastic models sheds light on the interplay between network parameters and matchability.**

In this paper, we consider the alignment problem when the two networks overlap only partially, i.e., there exist vertices in one network that have no counterpart in the other. We define a random bigraph model that generates two correlated graphs $G_{1,2}$; it is parameterized by the expected node overlap t^2 and by the expected edge overlap s^2. We define a cost function for structural mismatch under a particular alignment, and we identify a threshold for perfect matchability: if the average node degrees of $G_{1,2}$ grow as $\omega\left((s^{-2}t^{-1}\log(n)\right)$, then minimization of the proposed cost function results in an alignment which (i) is over exactly the set of shared nodes between G_1 and G_2, and (ii) agrees with the true matching between these shared nodes. Our result shows that network alignment is fundamentally robust to partial edge and node overlaps.

I. Introduction

Graph data captures relationships among entities, which is a central abstraction in many fields, including the social sciences, biology, information security, pattern recognition, machine vision, and networking. In many data analysis applications, information from different sources has to be merged into an integrated data model. This is notoriously difficult, because entity names or features from different sources are often unreliable and/or incompatible. When merging graph data, one remedy is to rely on structural information rather than on explicit vertex labels or vertex features to match two (or several) graphs. This network reconciliation problem has received significant attention recently: Social networks can be aligned by structural information [3], [4], [10], [13], [17], [19], [21], [26], with applications in network de-anonymization [11], [12], [18], [20], [25]; protein-interaction network matching allows us to find proteins with common functions in different species [14], [15], [22]; graph matching has many applications in pattern recognition and machine vision [5], e.g., finding similar images in a database by matching segment-adjacency graphs [7], [16], [24].

Network alignment[1] can be viewed as a generalization of the classic graph-isomorphism problem. Graph isomorphism is hard in general and is in NP (but not known to be in NP-complete). For specific classes of graphs, more is known: for example, for the Erdös-Rényi random graph $G(n,p)$ [8] the threshold function for asymmetry is known to be $p = \log(n)/n$ [1].

However, finding the exact graph isomorphism can be (exponentially) complex in the worst case.[2] In addition, in the scenarios considered here, the two graphs are subject to noise and uncertainties, and are not exactly isomorphic [5]. To address the above two issues, several heuristics have been proposed [5], for example based on a notion of graph edit distance [9]. In general, performance guarantees and a characterization of feasible classes of graphs to be matched by such heuristics have been elusive.

Recent work [21] has taken an information-theoretic angle and shown conditions on the parameters of a random bigraph model when perfect matching is possible. This model generates two correlated $G(n,ps)$ random graphs, with a similarity parameter $0 \leq s \leq 1$. When $s < 1$, with high probability the two graphs are not isomorphic, but [21] establishes a threshold function for p such that the correct alignment can nevertheless be identified. The threshold is proportional to $c(s)\log(n)/n$, where the function $c(s)$ is a penalty due to the dissimilarity of the two graphs. In summary, their work shows conditions where graph structure fundamentally contains sufficient information to find alignments, if computational resources are unlimited.

However, they make several strong and unrealistic assumptions, including that the vertex sets of the two graphs are of the same size, and that a full matching between these sets can be found. In most practical scenarios, node overlap would be only partial. For example, when reconciling two social networks, we should be able to allow for users of one network not to be users of the other. To the best of our knowledge, it is an open question to what extent partial overlap of the node sets hampers the feasibility of network alignment. We address this question in this paper.

Contributions We make the following contributions in this paper.

(a) First, we extend the random bigraph model of [21] to generate two Erdös-Rényi random graphs whose vertex sets overlap only partially. The model has two parameters (t and s) to control vertex overlap and edge overlap, respectively.

(b) Second, our main result is a sufficient condition on the graph density (or average vertex degree) and on the amount of noise for perfect matching. A perfect matching amounts to (i) filtering out nodes without counterparts in both G_1 and

*École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, firstname.lastname@epfl.ch. L. Yartseva was partially supported by the Swiss National Science Foundation (SNSF) under grant 200021-149826/1.

[1]Network alignment is also known as graph matching or network reconciliation in the literature.

[2]The class of graphs that appear the most challenging is thought to be the strongly regular graphs [23].

G_2, and (ii) correctly match the remaining nodes that are present in both graphs.

(c) Third, we formulate network alignment as an optimization problem over the space of all possible partial matchings between the two node sets. We show scaling conditions such that minimizing a cost function identifies the true matching with high probability. While the optimization formulation does not lend itself to a scalable algorithm, our results delineate the boundary between what is fundamentally possible and impossible.

This paper is structured as follows. In Section II, we introduce our model for generating correlated graphs with partial vertex overlap, and state our main result. In Section III, we prove the result. Section IV concludes the paper. Some technical details are relegated to appendices.

II. Model and Conditions for Perfect Matching

In this section, we first state the graph matching problem formally. Then, to formalize a partial overlap in the vertex sets of the graphs, we present a random bigraph model that generates two correlated Erdös-Rényi random graphs. We introduce a cost function for structural mismatch under a given candidate alignment of the two graphs. Finally, we state the main theorem of this paper. Our theorem shows that under surprisingly mild conditions, minimizing this cost function finds the correct matching with high probability.

A. Graph Matching

Assume we are given two graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$, which may represent, for example, two social networks (e.g., G_1 is Facebook, G_2 is LinkedIn). We know that some users have profiles in several social networks. In this paper, we study the graph matching problem, which refers to inferring the alignment of the common users of the networks G_1 and G_2 by structural information only.

The graph-matching problem is defined formally as follows. Given the two graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$, the goal is to find a matching between the nodes in $V_0 = V_1 \cap V_2$, where V_0 (we define $n_0 = |V_0|$) is the set of vertices common to both graphs. We call this true hidden matching π_0. We assume that, without loss of generality, $V_{1,2} \subset [n] = \{1, \ldots, n\}$ and denote $n_1 = |V_1|$, $n_2 = |V_2|$. Next, we define the set of all possible matchings Π from graph G_1 to G_2.

Definition 1: Π is the set of all *partial matchings* π from the vertex set V_1 to V_2. A partial matching π is a subset of $V_1 \times V_2$ such that any node in $V_1 = \{1, \ldots, n_1\}$ and $V_2 = \{1, \ldots, n_2\}$ is matched to at most one node in the other graph.

Thus, the *identity* hidden matching π_0 is the set of couples of nodes that are sampled in both graphs G_1 and G_2, i.e., $\pi_0 = \{[u, u] : u \in V_0\}$. Further, if node $v_1 \in V_1$ is matched to node $v_2 \in V_2$, we say $v_2 = \pi(v_1)$ and $v_1 = \pi^{-1}(v_2)$. For a pair of nodes $e = (u, v)$ we define $\pi(e) = (\pi(u), \pi(v))$. Let us define $V_{1,2}(\pi)$ as the sets of vertices in $V_{1,2}$ that are matched by π, and $E_{1,2}(\pi)$ as the sets of matched edges (an edge is matched if both endpoints are matched). For a node u, we say $\pi(u)$ is *null* (denoted by $\pi(u) = \emptyset$) if either u is not sampled ($u \notin V_1$) or u is not matched (i.e., $u \in V_1$ but $u \notin V_1(\pi)$). Similarly, for a node v, we say $\pi^{-1}(v)$ is *null* ($\pi^{-1}(v) = \emptyset$) if $v \notin V_2$ or $v \notin V_2(\pi)$. For a pair $e = (u, v)$, $\pi(e)$ is defined to be null (denoted by $\pi(e) = \emptyset$) if either $\pi(u) = \emptyset$ or $\pi(v) = \emptyset$. Similarly, $\pi^{-1}(e) = \emptyset$ if either $\pi^{-1}(u) = \emptyset$ or $\pi^{-1}(v) = \emptyset$.

Definition 2: For a matching π we define (i) $|\pi|$ as the size of matching π, (ii) l as the number of correctly matched couples of the form $[i, i]$ and, (iii) $k = |\pi| - l$ as the number of wrongly matched couples. Let Π_k^l represent a class of matchings of size $|\pi| = l + k \leq \min\{n_1, n_2\}$ with l correctly matched couples. Note that the sets Π_k^l partition the set Π of all partial matchings.

For example, Fig. 1 shows the identity matching $\pi_0 \in \Pi_0^7$ and the matching $\pi \in \Pi_6^2$ from V_1 to V_2.

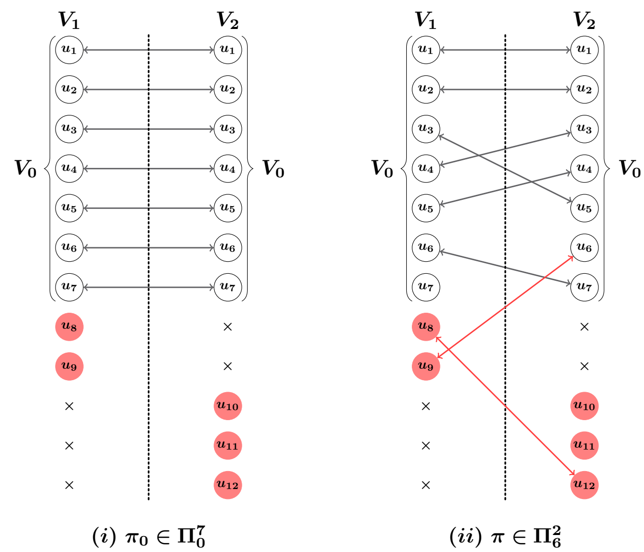

Fig. 1: *Examples of two matchings: (i) The true matching $\pi_0 \in \Pi_0^7 = \{[u_1, u_1], \ldots, [u_7, u_7]\}$, and (ii) the matching $\pi \in \Pi_6^2$. White nodes are sampled in both graphs, while red nodes are sampled in only one but not the other.*

B. Random Bigraph Model

We study the graph-matching problem under a random bigraph model. This model assumes that graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ are sampled from an Erdös-Rényi ($G(n, p)$) [8] graph $G(V, E)$ as follows: First, the generator graph $G(V, E)$ is sampled from the probability space of $G(n, p)$ graphs with n nodes, where each of the $\binom{n}{2}$ possible edges exists independently with probability $0 < p \leq 1$; Second, vertex sets $V_{1,2}$ are sampled independently from the vertex set V with probability t, i.e., $\mathrm{P}(u \in V_1) = \mathrm{P}(u \in V_2) = t$ for all $u \in V$. Third, the edge sets $E_{1,2}$ are sampled from those edges in E whose both endpoints are sampled in $V_{1,2}$; this means that each edge is in $E_{1,2}$ independently with probability s.

We refer to this model as the $G(n, p; t, s)$ bigraph model. This model is inspired by [21], but considers a more challenging and realistic scenario where the two graphs have partially overlapping vertex sets (this is modeled by the node sampling process).

C. Perfect Matchability Under Structural Mismatch

We now define a cost function that quantifies the structural mismatch between the two graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ under a given partial matching π. The cost function has two terms Φ_π and Ψ_π:

- Mismatched edges:

$$\Phi_\pi = \sum_{e \in E_1(\pi)} 1_{\{\pi(e) \notin E_2\}} + \sum_{e \in E_2(\pi)} 1_{\{\pi^{-1}(e) \notin E_1\}}.$$

- Unmatched edges: $\Psi_\pi = \Psi_\pi^1 + \Psi_\pi^2$, where Ψ_π^1 and Ψ_π^2 are the number of unmatched edges in E_1 and E_2, respectively. More precisely, we define

$$\Psi_\pi^1 = |\{e \in E_1 \backslash E_1(\pi)\}| \text{ and } \Psi_\pi^2 = |\{e \in E_2 \backslash E_2(\pi)\}|.$$

The cost function is a weighted sum of Φ_π and Ψ_π:

$$\Delta_\pi = \Phi_\pi + \alpha \Psi_\pi. \quad (1)$$

Our approach consists in minimizing the cost function Δ_π over all possible partial matchings π. There is a tradeoff between the two cost terms (1): adding node couples to the matching π cannot decrease Φ_π (and it can increase even for correct couples because of edge sampling), while Ψ_π cannot increase. The parameter α controls this tradeoff: with $\alpha = 0$, the trivial empty matching minimizes Δ_π; with $\alpha > 1$ the optimal matching is always of the largest possible size $\min\{n_1, n_2\}$, because the increase in Φ_π when adding a couple to π is smaller than the decrease in $\alpha\Psi_\pi$. Below, we identify constraints on α and provide an appropriate value such that with high probability, matching found by minimizing Δ_π is the correct partial matching π_0.

We now state the main result of the paper.

Theorem 3: In the $G(n, p; t, s)$ bigraph model with $\frac{\log n}{ns^3t^2} \ll p \ll 1$, there exists a value of α such that with high probability

$$\pi_0 = \operatorname*{argmin}_\pi \Delta_\pi. \quad (2)$$

Before proving Theorem 3, we provide some context for the result.

Expressed in terms of the expected degree $npst$ of the two observable graphs $G_{1,2}$, the threshold is $\log(n)/s^2t$ for perfect matchability.

The dependence on n is tight. To see this, consider the intersection graph $G_0 = G(V_0, E_1 \cap E_2)$. Its expected degree is nps^2t^2.[3] If this is asymptotically less than $\log nt^2$, then G_0 has symmetries w.h.p. (which in fact stem from isolated vertices [2]). In this case, the correct matching cannot be determined uniquely. To see this, assume that an oracle reveals, separately for G_1 and for G_2, the set of nodes and edges without counterpart. These sets contain no useful information to estimate π_0 over the common nodes, because of the independence assumptions in the model. Essentially, given an oracle, G_0 is a sufficient statistic for π_0, whose symmetries would preclude inferring π_0.

Based on this argument, the dependence on t is tight as well, while there is a gap of a factor of s between the achievability result in Theorem 3 and the trivial lower bound based on G_0. It is not clear whether the upper or lower bound is loose with respect to s.

With $t = 1$, we can recover the achievability result of Pedarsani and Grossglauser [21] up to a constant. Note that this is not trivial, as their problem formulation minimizes a cost function[4] over the set $\{\Pi_k^l : k + l = n\}$, while here we minimize over the larger set $\{\Pi_k^l : k + l \leq n\}$. Our result shows, then, that there is asymptotically no penalty for not knowing *a priori* the overlap set V_0.

The cost function Δ_π with $\alpha = 1$ is similar to a simple graph edit distance between G_1 and G_2. Suppose we wanted to find the cheapest way to transform the unlabeled graph G_1 into G_2 through edge additions and deletions. Then the number of operations is exactly Δ_π. Our conditions on α (discussed in detail within the proof) show that minimizing this edit distance does not work. Instead, the tradeoff between penalizing mismatched mapped edges and unmapped edges needs to be controlled more finely through an appropriate choice of α that depends on p and s.

The result is for the Erdös-Rényi random graph model with uniform sampling. This parsimonious model is a poor approximation of most real networks, which have salient properties not shared with random graphs (skewed degree distribution, clustering, community structure, etc.). However, we conjecture that network alignment for random graphs is harder than for real graphs, because the structural features of real networks make nodes more distinguishable than in random graphs. Our results suggest that even for the difficult case of random graphs, network alignment is fundamentally easy given sufficient computational power.

III. Proof of Theorem 3

We provide a brief sketch followed by the detailed proof. Let S be the number of matchings $\pi \in \Pi$ such that $\Delta_\pi - \Delta_{\pi_0} \leq 0$. Following the Markov inequality, as S is a non-negative integer-valued random variable, we have $\mathrm{P}[S \geq 1] \leq \mathrm{E}[S]$. We will prove that, under the conditions of Theorem 3,

$$\mathrm{P}[S \geq 1] \leq \mathrm{E}[S] = \sum_{\pi \in \Pi} \mathrm{P}(\Delta_\pi - \Delta_{\pi_0} \leq 0) \to 0. \quad (3)$$

The main complication of the proof stems from the fact that the random variables Δ_π and Δ_{π_0} are correlated in a complex way, because they are both functions of the random vertex and random edge sets $V_{1,2}$ and $E_{1,2}$. Both Δ_π and Δ_{π_0} can be written as sums of Bernoulli random variables. The main challenge in the proof is to decompose the difference $\Delta_\pi - \Delta_{\pi_0}$ into components that are mutually independent and can be appropriately bounded.

For this, we first partition the node sets V_1 and V_2 with respect to how they are mapped by π and π_0. This node

[3] To be precise, $(n-1)ps^2t^2$; we sometimes omit lower-order terms for readability.

[4] Identical to ours with $\alpha = 0$.

partition induces an edge partition. Elements of some parts of the edge partition contribute equally to Δ_π and Δ_{π_0} and can be ignored. The remaining parts can be further subdivided into linear structures (specifically, chains and cycles) with only internal and short-range correlation. Finally, this leads to the desired decomposition of the sums of Bernoullis, which is fine enough to apply standard concentration arguments to Δ_π and Δ_{π_0} individually, and to then stochastically bound their difference.

Proof of Theorem 3 We consider the contribution of edges (or potential edges) to the terms Δ_π and Δ_{π_0} as a random variable in the $G(n, p; t, s)$ probability space. More precisely, for a pair of nodes $u, v \in V_1$ and their images under the matching π (i.e., $\pi(u), \pi(v)$) we look at the probability of having/not having an edge between these nodes in $G_{1,2}$. From now on, a *pair* e represents a possible edge $e = (u, v)$ which, based on the realization of the $G(n, p; t, s)$ bigraph random model, might have or not have an actual edge between the nodes u and v.

Let us call the set of all pairs in G_1 as V_1^2 (here, we slightly abuse the notation, meaning $\binom{V_1}{2}$). The set V_2^2 is defined similarly. We define, by analogy, the set of matched pairs $V_1^2(\pi)$ as the set of all the pairs $(u, v) \in \binom{V_1(\pi)}{2}$. Also, the set $V_2^2(\pi)$ is defined similarly.

The term Φ_π counts the number of edges in both graphs that are matched to a nonexistent edge in the other graph. More precisely, the contribution of pair $e \in V_1^2(\pi)$ and its image $\pi(e) \in V_2^2(\pi)$ to Φ_π is $\phi(e) = |1_{\{e \in E_1(\pi)\}} - 1_{\{\pi(e) \in E_2(\pi)\}}|$. Note that pairs e and $\pi(e)$ contribute to Φ_π if and only if exactly one of them exists in G_1 or G_2. Also, for $e \in V_1^2 \setminus V_1^2(\pi)$, we define $\psi_1(e) = 1_{\{e \in E_1 \setminus E_1(\pi)\}}$ which represents the contribution of pair e to Ψ_π^1. This indicator term is equal to 1 if the edge between unmatched pair e in G_1 exists. Similarly, for $e \in V_2^2 \setminus V_2^2(\pi)$, we define $\psi_2(e) = 1_{\{e \in E_2 \setminus E_2(\pi)\}}$. To sum up, we can write Δ_π as

$$\Delta_\pi = \sum_{e \in V_1^2(\pi)} \phi(e) + \alpha \left[\sum_{e \in V_1^2 \setminus V_1^2(\pi)} \psi_1(e) + \sum_{e \in V_2^2 \setminus V_2^2(\pi)} \psi_2(e) \right]. \quad (4)$$

In order to compute contributions of pairs to Δ_π and Δ_{π_0}, we first partition the vertices in the set $V_1 \cup V_2$ based on the matchings π and π_0. Then we partition the node pairs with respect to this node partition.

A. Node Partition

We partition the nodes in $V_1 \cup V_2$ into the following five parts based on the matching π:

(i) $\checkmark(\pi)$ is the set of nodes that are matched correctly by π, i.e., $\checkmark(\pi) = \{u \in V_1 \cup V_2 | \pi(u) = u\}$.
(ii) $\rightarrow(\pi)$ is the set of nodes that are matched in the graph G_1, but π^{-1} is null for them, i.e., $\rightarrow(\pi) = \{u \in V_1 \cup V_2 | \pi(u) \neq \emptyset, \pi^{-1}(u) = \emptyset\}$.
(iii) $\leftarrow(\pi)$ is the set of nodes that are matched in the graph G_2, and π is null for them, i.e., $\leftarrow(\pi) = \{u \in V_1 \cup V_2 | \pi(u) = \emptyset, \pi^{-1}(u) \neq \emptyset\}$.
(iv) $\leftrightarrow(\pi)$ is the set of nodes that are matched in both graphs $G_{1,2}$, but wrongly, i.e., $\leftrightarrow(\pi) = \{u \in V_1 \cup V_2 | \pi(u) \neq \{u, \emptyset\}, \pi^{-1}(u) \neq \emptyset\}$.
(v) $\times(\pi)$ is the set of nodes which are null in both graphs $G_{1,2}$ under the matching π, i.e., $\times(\pi) = \{u \in V_1 \cup V_2 | \pi(u) = \emptyset, \pi^{-1}(u) = \emptyset\}$.

In the matching π_0 all the nodes in V_0 are matched correctly and the other nodes are left unmatched; therefore, only the two sets $\checkmark(\pi_0)$ and $\times(\pi_0)$ are nonempty. The pairwise intersections of the partitions under the two matchings π and π_0 are defined in Table I. For an example of these pairwise intersections, see Table II.

π_0 \ π	$\checkmark$	$\leftrightarrow$	$\rightarrow$	$\leftarrow$	$\times$
$\checkmark$	$\mathcal{C}$	$\mathcal{W}$	$\mathcal{L}$	$\mathcal{R}$	$\mathcal{S}$
$\times$	$\emptyset$	$\emptyset$	$\mathcal{Q}$	$\mathcal{X}$	$\mathcal{U}$

TABLE I: *Partition of the nodes in $V_1 \cup V_2$ into eight sets based on the pairwise intersections of partition of the nodes in $V_1 \cup V_2$ under π and π_0.*

π_0 \ π	$\checkmark$	$\leftrightarrow$	$\rightarrow$	$\leftarrow$	$\times$
$\checkmark$	u_1, u_2	u_3, u_4, u_5, u_6	$\emptyset$	u_7	$\emptyset$
$\times$	$\emptyset$	$\emptyset$	u_8, u_9	u_{12}	u_{10}, u_{11}

TABLE II: *Example of partition of the nodes $V_1 \cup V_2$ of the graphs $G_{1,2}$ from Fig. 1.*

B. Edge Partition

We now partition the set of pairs based on the classes of nodes which are defined in Table I. A pair e contributes equally to Δ_π and Δ_{π_0} if it is matched in the same way by π and π_0 (i.e., $\pi_0(e) = \pi(e)$), or if it is null in both. The following sets are those pairs that contribute equally to Δ_π and Δ_{π_0}, and consequently, their contributions will cancel-out in the difference $\Delta_\pi - \Delta_{\pi_0}$:

1) Pairs between the nodes in the set $\mathcal{C}$. These pairs are present in both graphs and their endpoints are matched correctly by both π and π_0. For example, in Fig. 1, the pair (u_1, u_2) is matched to the same pair by matchings π_0 and π.
2) Pairs in G_1 between $\mathcal{U} \cap V_1$ (i.e., the nodes in V_1 which are unmatched by π and not sampled in V_2) and V_1 contribute equally to both Ψ_π and Ψ_{π_0}. Similarly, for the pairs in $(\mathcal{U} \cap V_2) \times V_2$ in the graph G_2. Note that these pairs are present in only one of the graphs. As an example, in Fig. 1, the pairs (u_{10}, u_{11}), (u_{10}, u_{12}) and (u_{10}, u_2) in graph G_2 are matched neither under π nor under π_0.
3) Pairs e between $\mathcal{Q}$ and $\mathcal{S} \cup \mathcal{R}$ in the graph G_1 contribute equally to both Ψ_π and Ψ_{π_0} by a term $\psi_1(e)$. Similarly, the pairs between $\mathcal{X}$ and $\mathcal{S} \cup \mathcal{L}$ in the graph G_2 contribute a term $\psi_2(e)$ under both matchings π and π_0. Note that these pairs are present only in one of

the graphs. In Fig. 1, (u_7, u_8) and (u_7, u_9) provide two examples of pairs in this class from graph G_1.

Let Z_π and Z_{π_0} denote the contribution of these pairs to Δ_π and Δ_{π_0}, respectively. By definition $Z_\pi = Z_{\pi_0}$. Call $\mathcal{E}$ the set of all the remaining pairs that are matched differently under π and π_0. Note that $\mathcal{E}$ depends on both matchings π and π_0. As for each instance of the $G(n, p; t, s)$ bigraph model the matching π_0 is fixed, for simplicity of notation we drop the dependence on π_0 and define $X_\pi = \Delta_\pi - Z_\pi$ and $Y_\pi = \Delta_{\pi_0} - Z_{\pi_0}$. Here X_π and Y_π represent the sums of indicator terms over the contribution of pairs in the set $\mathcal{E}$ under matchings π and π_0, respectively. To wrap up, we have

$$\Delta_\pi - \Delta_{\pi_0} = (X_\pi + Z_\pi) - (Y_\pi + Z_{\pi_0}) = X_\pi - Y_\pi. \quad (5)$$

The next step of the proof is to find a lower-bound for $X_\pi - Y_\pi$. In order to compute contributions of pairs from the set $\mathcal{E}$ to different indicator terms in X_π and Y_π, we partition this set into the following subclasses:

1) The set of pairs present in only one of the graphs $G_{1,2}$ and matched by π. Note that at least one of the endpoints of these pairs are not sampled in either $V_{1,2}$. Therefore, these pairs are not matched by π_0. These pairs are divided into the two following sets:
   - $\mathcal{E}_{\emptyset,M*} = \{(i, j) \in (\mathcal{Q} \times V_1(\pi))\}$ is the set of pairs that contribute a $\psi_1(e)$ to $\Psi^1_{\pi_0}$ and a $\phi(e)$ to Φ_π.
   - $\mathcal{E}_{\emptyset,*M} = \{(i, j) \in (\mathcal{X} \times V_2(\pi))\}$ is the set of pairs that contribute a $\psi_2(e)$ to $\Psi^2_{\pi_0}$ and a $\phi\left(\pi^{-1}(e)\right)$ to Φ_π.

   For example, in Fig. 1, we have $(u_3, u_8) \in \mathcal{E}_{\emptyset,M*}$ and $(u_1, u_{12}) \in \mathcal{E}_{\emptyset,*M}$.
2) The set of pairs present in both graphs $G_{1,2}$ but unmatched by π in at least one of the graphs. These pairs can be further partitioned into three subclasses:
   - $\mathcal{E}_{M,M\emptyset} = \{(i, j) \in \mathcal{L} \times (\mathcal{C} \cup \mathcal{W} \cup \mathcal{L})\}$ is the set of pairs that are matched in G_1 and unmatched in G_2. A pair $e \in \mathcal{E}_{M,M\emptyset}$ contributes to a $\phi(e)$ to Φ_{π_0} and Φ_π, and $\psi_2(e)$ to Ψ^2_π.
   - $\mathcal{E}_{M,\emptyset M} = \{(i, j) \in \mathcal{R} \times (\mathcal{C} \cup \mathcal{W} \cup \mathcal{R})\}$ is the set of pairs that are matched in G_2 and unmatched in G_1.
   - $\mathcal{E}_{M,\emptyset\emptyset} = \{(i, j) \in (\mathcal{S} \times V_0) \bigcup (\mathcal{L} \times \mathcal{R})\}$ is the set of pairs that are unmatched by π in both graphs. These pairs contribute to a $\phi(e)$ to Φ_{π_0}, and $\psi_2(e)$ to both Ψ^1_π and Ψ^2_π.

   In Fig. 1, the unmatched pair (u_4, u_7) in G_1 is matched by π only in G_2, i.e., $(u_4, u_7) \in \mathcal{E}_{M,\emptyset M}$.
3) $\mathcal{E}_{M,MM} = \{(i, j) \in \mathcal{W} \times (\mathcal{C} \cup \mathcal{W})\}$ is the set of pairs that are present and matched, but wrongly, by π in both graphs $G_{1,2}$. These pairs are matched differently by π and π_0. The pairs in the set $\mathcal{E}_{M,MM}$ contribute to a $\phi(e)$ in Φ_{π_0}, and contribute to terms $\phi(e)$ and $\phi\left(\pi^{-1}(e)\right)$ in Φ_π. For example, in Fig. 1, the pairs (u_1, u_3) and (u_4, u_5) which are matched differently by π_0 and π belong to the set $\mathcal{E}_{M,MM}$.

   Note that this is not generally true. Indeed, transpositions[5] in π contribute equally to both Φ_π and Φ_{π_0}. We have at most $\lfloor k/2 \rfloor$ pairs of this type, because the number of wrongly matched couples is k. To be precise, we do not consider these pairs in the set $\mathcal{E}_{M,MM}$.

Now, let us define the sizes of the described sets as follows: $m_1 = |\mathcal{E}_{\emptyset,M*} \cup \mathcal{E}_{\emptyset,*M}|$, $m_{2,1} = |\mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,\emptyset M}|$, $m_{2,2} = |\mathcal{E}_{M,\emptyset\emptyset}|$, $m_2 = m_{2,1} + m_{2,2}$ and $m_3 = |\mathcal{E}_{M,MM}|$. Also, we define $m = m_1 + m_2 + m_3$.

C. Indicator Terms and Expected Values

In Lemma 4, the two terms X_π and Y_π are expressed as sums of indicator terms (Bernoulli random variables) over the pairs in $\mathcal{E}$.

Lemma 4: For X_π we have:

$$X_\pi = \sum_{e \in \mathcal{E}_{\emptyset,M*} \cup \mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,MM}} \phi(e) + \alpha \left[\sum_{e \in \mathcal{E}_{M,\emptyset M} \cup \mathcal{E}_{M,\emptyset\emptyset}} \psi_1(e) + \sum_{e \in \mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,\emptyset\emptyset}} \psi_2(e) \right], \quad (6)$$

where $\phi(e) \sim Be\left(2ps(1 - ps)\right)$ and $\psi_1(e), \psi_2(e) \sim Be(ps)$. For Y_π we have:

$$Y_\pi = \sum_{e \in \mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,\emptyset M} \cup \mathcal{E}_{M,\emptyset\emptyset} \cup \mathcal{E}_{M,MM}} \phi(e) + \alpha \left[\sum_{e \in \mathcal{E}_{\emptyset,M*}} \psi_1(e) + \sum_{e \in \mathcal{E}_{\emptyset,*M}} \psi_2(e) \right], \quad (7)$$

where $\phi(e) \sim Be\left(2ps(1 - s)\right)$, and $\psi_1(e), \psi_2(e) \sim Be(ps)$.

Proof: First, note that $\mathcal{E}_{\emptyset,M*} \cup \mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,MM} = \mathcal{E} \cap V_1^2(\pi)$ is the set of all matched pairs from G_1 which are in the set $\mathcal{E}$. Remember that by (5) the term X_π is the sum of indicators in Δ_π over pairs in the set $\mathcal{E}$. Thus, we get the first term in the right hand side of (6). Each pair e (same is true for $\pi(e)$) exists in each of the graphs $G_{1,2}$ with probability ps; therefore $\phi(e) = Be\left(2ps(1 - ps)\right)$. Second, we compute the number of terms $\psi_{1,2}(e)$ that contribute to X_π. These are (i) pairs of type $\mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,\emptyset M}$ that contribute to either Ψ^1_π or Ψ^2_π, and (ii) pairs of type $\mathcal{E}_{M,\emptyset\emptyset}$ that contribute to both Ψ^1_π and Ψ^2_π. The probability of a pair e to have an actual edge $e \in E_{1,2}$ is ps, hence $\psi_1(e), \psi_2(e) \sim Be(ps)$.

Y_π is the contribution of the pairs in the set $\mathcal{E}$ to Δ_{π_0}. For each pair e matched by π_0 and π, $e \in \mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,\emptyset M} \cup \mathcal{E}_{M,\emptyset\emptyset} \cup \mathcal{E}_{M,MM}$ there is an indicator $\phi(e)$ in Y_π. Note that this $\phi(e)$ is an indicator of the event that e is sampled in G_1 and $\pi(e) = e$ is not sampled in G_2 (or vice versa). Thus $\phi(e) = Be\left(2ps(1 - s)\right)$. The argument for $\psi_1(e), \psi_2(e)$ is the same as for X_π. This proves (7). ■

In the next corollary, we compute the expected values of X_π and Y_π.

[5]A pair (u, v) is a transposition under π if $\pi(u) = v$ and $\pi(v) = u$.

Corollary 5: For X_π and Y_π we have

$$\mathrm{E}[X_\pi] = \left(m_3 + \frac{m_1 + m_{2,1}}{2}\right) 2ps(1-ps) + \alpha m_{2,1} ps + 2\alpha m_{2,2} ps.$$

$$\mathrm{E}[Y_\pi] = (m_2 + m_3) 2ps(1-s) + \alpha m_1 ps.$$

Proof: Note that the term $\phi(e)$, which is defined as $\phi(e) = |1_{\{e \in E_1(\pi)\}} - 1_{\{\pi(e) \in E_2(\pi)\}}|$, depends to pairs e and $\pi(e)$ from graphs G_1 and G_2, respectively. Also, as the matching π is an injective function, each pair $e \in V_1^2$ can be matched to at most one pair from V_2^2. This is generally true for pairs $e \in V_2^2$ from G_2. Therefore, the number of pairs from graph G_1 which contribute to the $\{\phi(e)\}$ terms is equal to the number of pairs from graph G_2 which contribute to these terms, i.e., $|\mathcal{E}_{\emptyset,M*} \cup \mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,MM}| = |\mathcal{E}_{\emptyset,*M} \cup \mathcal{E}_{M,\emptyset M} \cup \mathcal{E}_{M,MM}|$. Remember that $|\mathcal{E}_{\emptyset,M*} \cup \mathcal{E}_{\emptyset,*M}| = m_1$ and $|\mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,\emptyset M}| = m_2$. To sum up, number of $\{\phi(e)\}$ terms which contribute to X_π (defined precisely in Lemma 4) is $m_3 + \frac{m_1 + m_{2,1}}{2}$. The rest comes directly from the definitions of m_1, m_2 and m_3. ■

In the following lemma, we prove that the expected value fof X_π is larger than the expected value of Y_π.

Lemma 6: If $1 - ps > \alpha > 1 - s$, then $\mathrm{E}[X_\pi] > \mathrm{E}[Y_\pi]$.

Proof: From Corollary 5, we have $\mathrm{E}[X_\pi] > ps\left((1-ps)m_1 + 2\alpha m_2 + 2(1-ps)m_3\right) > \mathrm{E}[Y_\pi]$ if the following inequalities hold: (i) $(1-ps) > \alpha$, (ii) $\alpha > (1-s)$, and (ii) $(1-ps) > (1-s)$. Note that if the first two inequalities hold, then the third inequality holds. ■

D. Correlation Structure

Lemma 6 guarantees that for any $\pi \neq \pi_0$, $\mathrm{E}[\Delta_\pi] > \mathrm{E}[\Delta_{\pi_0}]$. In the following, we demonstrate that X_π and Y_π, as sums of correlated Bernoulli random variables, concentrate around their means.

Due to the edge sampling process, the presence of edges between the nodes in V_0 is correlated in the two graphs G_1 and G_2. For example, consider an event $\phi(e)$ that is a function of edges $e \in G_1$ and $\pi(e) \in G_2$. Furthermore, assume $\pi(e)$ is sampled and matched in the graph G_1. Then, the presence of $\pi(e)$ in G_1 is correlated with the presence of $\pi(e)$ in G_2. Therefore, the two terms $\phi(e)$ and $\phi(\pi(e))$ are correlated. By the same lines of reasoning, if $\pi^2(e)$ is sampled and matched in G_1, the two terms $\phi(\pi(e))$ and $\phi\left(\pi^2(e)\right)$ are correlated, and so on. Thus, terms Φ_π and Ψ_π are the sums of correlated Bernoulli random variables.

To address these correlations, we first define *chains* and *cycles* of pairs under the alignment π. We call a sequence of different pairs $(e_1, \cdots, e_i \cdots, e_q)$ a *chain* if (i) $\pi^{-1}(e_1) = \emptyset$, i.e., e_1 is either unmatched or not sampled in G_2; (ii) $\pi(e_q) = \emptyset$, i.e., e_q is either unmatched or not sampled in G_1; and (iii) $\pi(e_i) = e_{i+1}$ for $1 \leq i < q$, i.e., each pair in a chain is the image of the previous pair in that chain under the alignment π. In Fig. 2b, the sequence $((u_3, u_9), (u_5, u_6), (u_4, u_7))$ is an example of a chain of length three. Also, we call a sequence of different pairs $(e_1, \cdots, e_i, \cdots, e_q)$ a *cycle* if (i) $\pi(e_i) = e_{i+1}$ for $1 \leq i < q$; and (ii) $\pi(e_q) = e_1$. As an example, see the cycle $((u_2, u_3), (u_2, u_5), (u_2, u_4))$ in Fig. 3a.

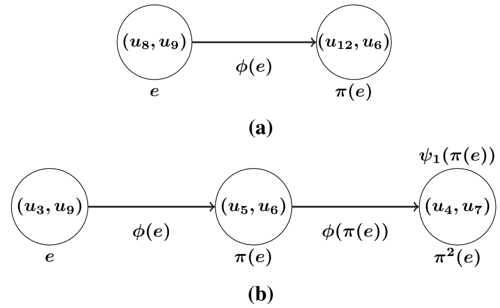

Fig. 2: *(a) Example of a chain with length one from the matching π from Fig. 1. (b) Example of a chain with length three from the matching π from Fig. 1: The term $\psi_1(\pi(e))$ corresponds to the contribution of pair (u_2, u_6) in the graph G_1. In this chain, the term $\phi(\pi(e))$ is correlated with the two terms $\phi(e)$ and $\psi_1(\pi(e))$.*

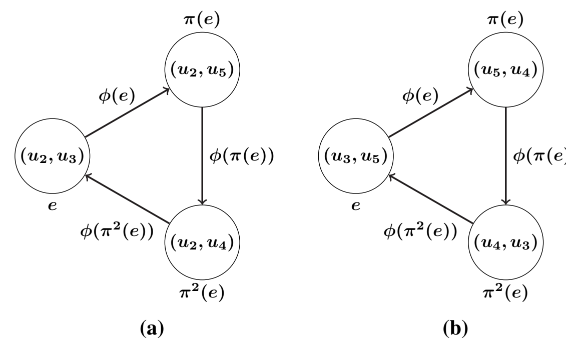

Fig. 3: *Examples of two cycles from the matching π from Fig. 1: Pairs generate a cycle of dependent terms. In these cycles, the terms $\phi(e), \phi(\pi(e))$ and $\phi\left(\pi^2(e)\right)$ are correlated pairwise.*

Following the discussion above, we state Lemmas 7 and 8. In Lemma 7, we (i) partition all the pairs of $\mathcal{E}$ into chains and cycles; and (ii) demonstrate contributions of these pairs to different indicator terms. In Lemma 8, we characterize correlations between terms in the induced sequence of indicators.

Lemma 7: All the pairs in the set $\mathcal{E}$ can be partitioned into chains and cycles, where they induce sequences of indicator terms as follows:

For each cycle $(e_1, \cdots, e_i, \cdots e_q), 1 \leq i < q$, its pairs contribute to the induced sequence of indicator terms $(\phi(e_1), \cdots \phi(e_i), \cdots \phi(e_q))$.

For each chain $(e_1, \cdots e_i, \cdots e_q), 1 \leq i < q$, its pairs contribute to one of the following five types of induced sequences of indicator terms:

1) $e_1 \in \mathcal{E}_{\emptyset,M*}$ and $e_q \in \mathcal{E}_{\emptyset,*M}$, these pairs contribute to the induced sequence of indicator terms $(\phi(e_1), \cdots \phi(e_i), \cdots \phi(e_{q-1}))$.
2) $e_1 \in \mathcal{E}_{\emptyset,M*}$ and $e_q \in \mathcal{E}_{M,\emptyset M}$, these pairs contribute to the induced sequence of indicator terms $(\phi(e_1), \cdots \phi(e_i), \cdots \phi(e_{q-1}), \psi_1(e_q))$.

3) $e_1 \in \mathcal{E}_{M,M\emptyset}$ and $e_q \in \mathcal{E}_{\emptyset,*M}$, these pairs contribute to the induced sequence of indicator terms $(\psi_2(e_1), \phi(e_1), \cdots \phi(e_i), \cdots \phi(e_{q-1}))$.
4) if $e_1 \in \mathcal{E}_{M,M\emptyset}$ and $e_q \in \mathcal{E}_{M,\emptyset M}$, these pairs contribute to the induced sequence of indicator terms $(\psi_2(e_1), \phi(e_1), \cdots \phi(e_i), \cdots \phi(e_{q-1}), \psi_1(e_q))$.
5) $e_1 \in \mathcal{E}_{M,\emptyset\emptyset}$, here we have a chain of length one. The edge e_1 contributes to the induced sequence of indicator terms $(\psi_2(e_1), \psi_1(e_1))$.

Lemma 8: For sequences of induced indicator terms from partitions in Lemma 7, we have

- All the induced indicators ϕ/ψ associated with different chains and cycles are mutually independent.
- For a chain, each indicator ϕ/ψ is correlated with at most the preceding and subsequent indicators in the induced sequence.
- For a cycle, each indicator ϕ/ψ is correlated with at most the preceding and subsequent indicators in the induced sequence, and $\phi(e_1)$ is correlated with $\phi(e_q)$.

For details regarding the correctness of this partition, their induced indicator terms and the correlation arguments refer to Appendix II.

From Lemma 8, we know that each term $\phi(e)$ (or $\psi_{1,2}(e)$) is correlated with at most two of its neighbors (e.g., see Figs. 2 and 3). We associate a label 0 or 1 with all the induced $\phi(e)$ and $\psi_{1,2}(e)$ terms by alternating marks. We obtain a marking that all the indicators with the same mark are independent. This is not generally true for the terms at start and end of cycles with odd number of indicators. See the discussions on how to handle these special cases, and detailed computation of the concentration bounds in Appendix III.

Next, based on this marking procedure, we split X_π into two sums of independent random variables and derive concentration bounds for each sum.

E. Concentration

We define $\mu_1 = \mathrm{E}[X_\pi]$ and $\mu_2 = \mathrm{E}[Y_\pi]$ and apply a union bound for the difference $X_\pi - Y_\pi$ (5).

$$\mathrm{P}\left[X_\pi - Y_\pi \leq 0\right] \leq \mathrm{P}\left[X_\pi < \frac{\mu_1+\mu_2}{2}\right] + \mathrm{P}\left[Y_\pi > \frac{\mu_1+\mu_2}{2}\right]. \tag{8}$$

We use the following bounds for the concentration of X_π and Y_π around their means (See Lemma 13 from Appendix III).

$$\mathrm{P}\left[X_\pi < \frac{\mu_1+\mu_2}{2}\right] \leq 2\exp\left(-\frac{(\mu_1-\mu_2)^2}{96\mu_1}\right),$$
$$\mathrm{P}\left[Y_\pi > \frac{\mu_1+\mu_2}{2}\right] \leq \exp\left(-\frac{(\mu_1-\mu_2)^2}{12\mu_1}\right).$$

Next, we lower-bound $\mu_1 - \mu_2$ to estimate (8) as follows. Assume $\alpha' = \min\left((1-ps-\alpha), (\alpha-(1-s))\right)$, then from Corollary 5 we have $\mu_1 - \mu_2 \geq \alpha' ps(m_1+m_2+m_3) \geq ps\alpha' m$. Also, note that $\mu_1 \leq 2mps$ and $\mu_2 \leq 2mps$. To sum up, we have

$$\mathrm{P}[X_\pi - Y_\pi \leq 0] \leq 3\exp\left(-\alpha'^2 \frac{mps}{192}\right). \tag{9}$$

Thus the expected number of matchings $\pi \neq \pi_0$ such that $\Delta_\pi \leq \Delta_{\pi_0}$ is

$$E(S) \leq \sum_{k,l} |\Pi_k^l| \mathrm{P}[X_\pi - Y_\pi \leq 0]$$
$$\leq \sum_{k,l} |\Pi_k^l| 3\exp\left(-\frac{\alpha'^2}{192} psm\right).$$

To finalize our proof, it remains to find a lower bound for m (as the number of node pairs in the set $\mathcal{E}$) and an upper bound for $|\Pi_k^l|$.

Lemma 9: We have

1) if $k \leq n_0 - l$, then $m > \frac{(n_0-l)(n_0-2)}{2}$ and $|\Pi_k^l| < n^{3(n_0-l)}$.
2) if $k > n_0 - l$, then $m > \frac{k(n_0-2)}{2}$ and $|\Pi_k^l| < n^{3k}$.

Proof: First, we upper-bound the number of matchings in the set Π_k^l. Assume we first choose l nodes from n_0 nodes in the set V_0 that are matched correctly. Then, we choose k other nodes from the remaining nodes of V_1 and V_2. Also, there are at most $k!$ possible matchings between these k chosen nodes. Therefore,

$$|\Pi_k^l| \leq \binom{n_0}{l}\binom{n_1-l}{k}\binom{n_2-l}{k} k! \leq n_0^{n_0-l} n_1^k n_2^k. \tag{10}$$

Based on the value of k we consider two different cases:

- if $k \leq n_0 - l$, then $|\Pi_k^l| < n^{3(n_0-l)}$. By definition, $m = |\mathcal{E}|$ is the number of pairs which are matched differently by π and π_0. This includes the set of pairs between any sampled node $v_1 \in V_0$ and any node $v_2 \in V_0$ matched differently by π and π_0. Note that these pairs are all the present pairs and there are $m_2 + m_3$ of them. Also, we should consider the pairs that contribute equally to both terms due to transpositions. Thus we have $m \geq \binom{n_0-l}{2} + (n_0-l)l - \lfloor \frac{k}{2} \rfloor \geq \frac{(n_0-l)(n_0-2)}{2}$.
- if $k > n_0 - l$, then $|\Pi_k^l| < n^{3k}$. Here note that the set $\mathcal{E}$ includes all the pairs between any sampled node $v_1 \in V_0$ and any node $v_2 \in V_1(\pi) \cup V_2(\pi)$ which are matched differently by π and π_0. Again, we should consider transpositions. We compute the number of pairs matched by π as $m \geq m_3 + m_1 \geq \binom{k}{2} + kl - \lfloor \frac{k}{2} \rfloor$. After that, if $k \geq n_0$, we have the statement immediately; otherwise, we use $l > n_0 - k$, and obtain $m \geq \binom{k}{2} + k(n_0-k) - \lfloor \frac{k}{2} \rfloor \geq \frac{k(n_0-2)}{2}$.

■

Now, we find an upper bound for $\mathrm{E}[S]$ based on the above cases.

(1) If $k \leq n_0 - l$: we define $i = n_0 - l$. Using the facts that $m > \frac{(n_0-l)(n_0-2)}{2}$, $k \leq n$ and $|\Pi_k^l| < n^{3(n_0-l)}$, we obtain

$$\mathrm{E}[S] \leq \sum_{k,l} 3\exp\left(i\left(3\log n - ps\frac{\alpha'^2}{384}(n_0-2)\right)\right)$$
$$\leq \sum_{i=1}^{n_0} 3\exp\left((3i+1)\log n - ips\frac{\alpha'^2}{384}(n_0-2)\right).$$

(2) If $k > n_0 - l$: using the facts that $m > \frac{k(n_0-2)}{2}$ and $|\Pi_k^l| < n^{3k}$, we obtain

$$\mathrm{E}[S] \leq \sum_{k,l} 3\exp\left(k\left(3\log n - ps\frac{\alpha'^2}{384}(n_0-2)\right)\right)$$
$$\leq \sum_{k=1}^{n} 3\exp\left((3k+1)\log n - kps\frac{\alpha'^2}{384}(n_0-2)\right).$$

The geometric sum goes to 0 if the first term goes to 0. Thus given that $ps \gg \frac{1536}{\alpha'^2}\frac{\log n}{n_0}$, we obtain $\mathrm{E}[S] \to 0$. We obtain $n_0 = nt^2(1+o(1))$ from a Chernoff bound and get $ps \gg \frac{1536}{\alpha'^2}\frac{\log n}{nt^2}$.[6]

To conclude the proof of Theorem 3, we choose $\alpha = \frac{(1-ps)+(1-s)}{2} = 1 - \frac{s(1+p)}{2}$; then $\alpha' = \frac{s(1+p)}{2}$ and we derive the final bound $ps \gg \frac{\log n}{ns^2t^2}$.

IV. Conclusion

In this paper, we address the problem of matching two unlabeled graphs by their edge structure alone. We propose a stochastic model for generating two correlated graphs with partial node and edge overlaps. More precisely, we introduce the $G(n,p;t,s)$ bigraph generator model, where $G(n,p)$ is the underlying ground-truth graph, and t and s are two parameters that control the similarities of the vertex and edge sets, respectively. We take an information-theoretic perspective, in that we ignore computational limitations and identify sufficient conditions such that a combinatorial optimization problem yields the correct answer with high probability.

We give conditions on the graph density p, and prove that within these conditions the true partial matching between the node sets of the two graphs can be inferred with zero error. The conditions on the node and edge similarity parameters t and s are quite benign: essentially, the average node degree has to grow as $\omega\left(\frac{\log(n)}{s^2t}\right)$.

Beyond establishing the scaling relation of network alignment in the presence of partial node overlap, the structure of the cost function suggests heuristics for efficient algorithms. In particular, the cost function takes the form of a graph edit distance, but with a tradeoff between the two types of error (mismatch and map-to-null) that is quite delicate to control (through the parameter α). We therefore expect our model and result to be useful in the development and tuning of matching heuristics.

Acknowledgments

The authors would like to thank Hamed Hassani and Jefferson Elbert Simões for feedback on drafts of this paper.

References

[1] L. Babai, P. Erdös, and S. M. Selkow. Random graph isomorphism. *SIAM J. Comput.*, 9(3):628–635, 1980.
[2] B. Bollobás. *Random Graphs*. Cambridge University Press, 2001.
[3] C. F. Chiasserini, M. Garetto, and E. Leonardi. De-anonymizing scale-free social networks by percolation graph matching. In *Proc. of IEEE INFOCOM 2015*, Hong Kong, Apr 2015.
[4] C. F. Chiasserini, M. Garetto, and E. Leonardi. Impact of Clustering on the Performance of Network De-anonymization. In *Proc. of ACM COSN 2015*, Palo Alto, CA, USA, Nov 2015.

[6] For any $\alpha \in [1-s, 1-ps]$.

[5] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, 18(03):265–298, 2004.
[6] D. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.
[7] A. Egozi, Y. Keller, and H. Guterman. A probabilistic approach to spectral graph matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):18–27, Jan 2013.
[8] P. Erdös and A. Rényi. On Random Graphs I. *Publ. Math. Debrecen*, 6:290–297, 1959.
[9] M.-L. Fernández and G. Valiente. A graph distance metric combining maximum common subgraph and minimum common supergraph. *Pattern Recognition Letters*, 22(6):753–758, 2001.
[10] P. Foggia, G. Percannella, and M. Vento. Graph matching and learning in pattern recognition in the last 10 years. *International Journal of Pattern Recognition and Artificial Intelligence*, 28(01), 2014.
[11] S. Ji, W. Li, N. Z. Gong, P. Mittal, and R. Beyah. On your social network de-anonymizablity: Quantification and large scale evaluation with seed knowledge. In *Proc. of the Network and Distributed System Security (NDSS) Symposium*, San Diego, CA, USA, Feb 2015.
[12] S. Ji, W. Li, P. Mittal, X. Hu, and R. Beyah. Secgraph: A uniform and open-source evaluation system for graph data anonymization and de-anonymization. In *Proc. of USENIX Security Symposium 2015*, Washington, D.C., USA, Aug 2015.
[13] E. Kazemi, S. H. Hassani, and M. Grossglauser. Growing a graph matching from a handful of seeds. *Proc. of the VLDB Endowment*, 8(10):1010–1021, 2015.
[14] B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. of the National Academy of Sciences*, 100(20):11394–11399, 2003.
[15] G. Klau. A new graph-based method for pairwise global network alignment. *BMC Bioinformatics*, 10(Suppl 1):S59, 2009.
[16] D. Knossow, A. Sharma, D. Mateus, and R. Horaud. Inexact matching of large and sparse graphs using laplacian eigenvectors. In *Graph-Based Representations in Pattern Recognition*, pages 144–153. Springer, 2009.
[17] N. Korula and S. Lattanzi. An efficient reconciliation algorithm for social networks. *Proc. of the VLDB Endowment*, 7(5):377–388, 2014.
[18] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *Proc. of ACM WSDM 2010*, New York City, USA, Feb 2010.
[19] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Proc. of IEEE Symposium on Security and Privacy 2009*, Oakland, CA, USA, May 2009.
[20] P. Pedarsani, D. R. Figueiredo, and M. Grossglauser. A bayesian method for matching two similar graphs without seeds. In *Proc. of IEEE Communication, Control, and Computing (51st Annual Allerton Conference) 2013*, Monticello, IL, USA, Oct 2013.
[21] P. Pedarsani and M. Grossglauser. On the privacy of anonymized networks. In *Proc. of ACM SIGKDD 2011*, San Diego, CA, USA, Aug 2011.
[22] R. Singh, J. Xu, and B. Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *PNAS*, 105(35):12763–12768, 2008.
[23] D. A. Spielman. Faster isomorphism testing of strongly regular graphs. In *Proc. of ACM STOC 1996*, Philadephia, Pen., USA, May 1996.
[24] L. Torresani, V. Kolmogorov, and C. Rother. Feature correspondence via graph matching: Models and global optimization. In *Computer Vision–ECCV 2008*, pages 596–609. Springer, 2008.
[25] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel. A practical attack to de-anonymize social network users. In *Proc. of IEEE Symposium on Security and Privacy 2010*, Oakland, CA, USA, May 2010.
[26] L. Yartseva and M. Grossglauser. On the performance of percolation graph matching. In *Proc. of ACM COSN 2013*, Boston, MA, USA, Oct 2013.

Appendix I
Concentration lemmas

Lemma 10: [Chernoff-Hoeffding bound [6]]
Let $X \triangleq \sum_{i=1}^{n} X_i$ where $X_i, 1 \leq i \leq n$, are independently

distributed in $[0,1]$. Then for $\epsilon > 0$,

$$\mathrm{P}\left[X > (1+\epsilon)\mathrm{E}[X]\right] \leq \exp\left(-\frac{\epsilon^2}{3}\mathrm{E}[X]\right),$$

$$\mathrm{P}\left[X < (1-\epsilon)\mathrm{E}[X]\right] \leq \exp\left(-\frac{\epsilon^2}{2}\mathrm{E}[X]\right).$$

APPENDIX II
PARTITION OF NODE PAIRS INTO CHAINS AND CYCLES

In this appendix, we provide the detailed proof for Lemmas 7 and 8. More precisely, we prove that the set chains and cycles correctly partition the pairs in set $\mathcal{E}$, and we characterize the dependence structure of the indicators within this partition.

First, note that each pair $e \in \mathcal{E}_{\emptyset,M*}$ is present only in G_1, thus it contributes only to one $\phi(e)$ indicator term. Consider the chain $(e, \pi(e), \dots \pi^c(e))$ when c is the smallest number such that $\pi^{c+1}(e)$ is null. This case happens in one of the two following cases:

- if $\pi^c(e) \in \mathcal{E}_{\emptyset,*M}$ then $\pi^c(e)$ is matched and exists only in G_2. Therefore, this chain of pairs induces the sequence $\left(\phi(e), \cdots, \phi\left(\pi^{c-1}(e)\right)\right)$ of indicator terms. Fig. 2a is an example of such a chain under the matching π from Fig. 1.
- if $\pi^c(e) \in \mathcal{E}_{M,\emptyset M}$ then $\pi^c(e)$ exists in both graphs but is matched only in G_2. Therefore, this chain induces the sequence $(\phi(e), \cdots, \psi_1(\pi^c(e)))$ of indicator terms. Fig. 2b is an example of such a chain under the matching π from Fig. 1.

Second, each pair $e \in \mathcal{E}_{M,M\emptyset}$ is present in both G_1 and G_2, but is matched only in G_1, thus it contributes to terms $\phi(e)$ and $\psi_2(e)$. Consider the chain $(e, \pi(e), \dots \pi^c(e))$ when c is the smallest number such that $\pi^{c+1}(e)$ is null. This case happens in one of the two following cases:

- if $\pi^c(e) \in \mathcal{E}_{\emptyset,*M}$, then $\pi^c(e)$ is matched and exists only in the graph G_2. Therefore, this chain induces the sequence of $\left(\psi_2(e), \phi(e), \cdots, \phi\left(\pi^{c-1}(e)\right)\right)$ of indicator terms.
- if $\pi^c(e) \in \mathcal{E}_{M,\emptyset M}$, then $\pi^c(e)$ exists in both graphs but is matched only in the graph G_2. Therefore, this chain induces the sequence of $(\psi_2(e), \phi(e), \cdots, \psi_1(\pi^c(e)))$ of indicator terms.

Now we formulate a cycle/chain partition process as follows: First, for each pair, we build a chain as described above; second, for each pair $e \in \mathcal{E}_{M,M\emptyset}$ we also build a chain; third, for each pair of type $e \in \mathcal{E}_{M,\emptyset\emptyset}$ we build another chain $(\psi_1(e), \psi_2(e))$.

Note that the first two types of chains are duals of each other: For each chain of pairs which ends with a pair $e \in \mathcal{E}_{\emptyset,*M}$ or $e \in \mathcal{E}_{M,\emptyset M}$, we can build the same chain of pairs backwards; starting from e and applying π^{-1} instead of π. Based on this observation, we compute that there are $m_1 + m_2$ pairs that start or end a chain. Thus, the fourth step is to partition the remaining, unvisited pairs that all have type $\mathcal{E}_{M,MM}$ (note that they are sampled and matched by π in both graphs).

For each unvisited pair e, the unvisited pair $\pi(e)$ also has type $\mathcal{E}_{M,MM}$ (otherwise $\pi(e)$ and e belong to some chain and, hence, e is visited), thus the pairs e and $\pi(e)$ are not null. To build a cycle, we start with a pair e and build the sequence $(e, \cdots, \pi^c(e))$, where c is the smallest number such that $\pi^c(e) = e$. We continue until there are no more unvisited pairs. Note that each indicator of a pair belongs to at most one chain or cycle because π is an injective function from V_1^2 to V_2^2. Fig. 3 provides examples of cycles of pairs under the matching π from Fig. 1.

Note that pairs induced by transpositions generate cycles of length two, i.e., for a pair $e = (u,v)$ with $\pi(u) = v$ and $\pi(v) = v$ the cycle $(\phi(e), \phi(\pi(e))$ is generated where $\pi^2(e) = e$.

Remember that we defined the indicator terms as follows: (i) $\phi(e) = |1_{\{e \in E_1(\pi)\}} - 1_{\{\pi(e) \in E_2(\pi)\}}|$; (ii) $\psi_1(e) = 1_{\{e \in E_1 \setminus E_1(\pi)\}}$; and (iii) $\psi_2(e) = 1_{\{e \in E_2 \setminus E_2(\pi)\}}$. From the definition, it is clear that for two node pairs $e_i \neq e_j$, we have $\psi_1(e_i) \perp\!\!\!\perp \psi_2(e_j)$. Also, if $e_j \notin \{e_i, \pi(e_i)\}$, then $\phi(e_i) \perp\!\!\!\perp \psi_1(e_j), \psi_2(e_j)$. Further, if $e_j, \pi(e_j) \notin \{e_i, \pi(e_i)\}$, then $\phi(e_i) \perp\!\!\!\perp \phi(e_j)$.

Following these independence arguments, we simply can conclude that indicators associated with different chains and cycles are mutually independent, and these indicators are correlated only with their precedent and subsequent terms in induced sequences.

APPENDIX III
MARKING INDICATORS

In this appendix we show (i) that there is an efficient algorithm for marking the indicator terms to break the dependency between them; and (ii) based on this marking strategy, we derive a bound for the concentration of X_π around its expected value.

In Lemmas 7 and 8 we defined induced sequences of indicators terms and characterized their correlation. Now we mark each indicator with alternating $0/1$ in such a way that the indicators with the same mark are independent; except for the case when the beginning and the end of a cycle of odd length have the same mark. Another requirement is that for each type of indicator, i.e., (i) indicators $\phi(e)$ and (ii) start/end indicators $\psi_{1,2}(e)$ at least a constant fraction of indicators should be marked with 0 and a constant fraction of them with 1.

For a sequence of indicators $(\phi(e_1), \cdots \phi(e_i), \cdots \phi(e_q))$ induced by a cycle (See Lemma 7), we start with a pair $\phi(e_1)$ and mark it with $m(\phi(e_1)) = 0$. Next, we mark $\phi(e_2)$ with 1, $\phi(e_3)$ with 0 and so on. We continue the next sequence with a new mark (if we ended with 1 then we start with 0 and vice versa) until there are no more cycles.

For a sequences induced by chains, it is slightly more complicated. First, note that we can iteratively mark a sequence from the beginning or from the end. Second, we remind the reader that all the indicators induced by $e = e_1/e_q$ beginning/end of the chain are either $\phi(e)$ for $e \in \mathcal{E}_{\emptyset,M*} \cup \mathcal{E}_{\emptyset,*M}$ or $\psi(e)$ for $e \in \mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,\emptyset M} \cup \mathcal{E}_{M,\emptyset\emptyset}$.

Now, let us mark all the sequences of indicators induced by chains while doing the following four steps iteratively:

1) Take the sequence that starts/ends with an indicator $\phi(e)$ and mark $\phi(e)$ with $m(\phi(e)) = 0$ next we mark $\phi(\pi(e))$ (or $\phi(\pi^{-1}(e))$) with 1 , $\phi(\pi^2(e))$ with 0 and so on.
2) Take the sequence that starts/ends with an indicator $\psi(e)$ and mark $\psi(e)$ with $m(\psi(e)) = 0$ next we mark $\phi(\pi(e))$ (or $\phi(\pi^{-1}(e))$) with 1, $\phi(\pi^2(e))$ with 0 and so on.
3) Take the sequence that starts/ends with an indicator $\phi(e)$ and mark $\phi(e)$ with $m(\phi(e)) = 1$ next we mark $\phi(\pi(e))$ (or $\phi(\pi^{-1}(e))$) with 0 , $\phi(\pi^2(e))$ with 1 and so on.
4) Take the sequence that starts/ends with an indicator $\psi(e)$ and mark $\psi(e)$ with $m(\psi(e)) = 1$ next we mark $\phi(\pi(e))$ (or $\phi(\pi^{-1}(e))$) with 0, $\phi(\pi^2(e))$ with 1 and so on.

If we do not have more sequences that starts/ends with an indicator of one of the types, we continue marking the remaining sequences alternating a start mark with 0 or 1.

Lemma 11: There exists a marking of the indicators $\{\phi(e) \cup \psi_{1,2}(e)\}$ with 0/1 labels such that

1) at least $\frac{1}{3}$ indicators of pairs $\{\mathcal{E}_{\emptyset,M*} \cup \mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,MM}\}$ gets mark 0 and at least $\frac{1}{3}$ gets mark 1.
2) at least $\frac{1}{6}$ indicators $\{\psi_1(e)\}$, $\{\psi_2(e)\}$ of sets of pairs $\{\mathcal{E}_{M,\emptyset M} \cup \mathcal{E}_{M,\emptyset\emptyset}\}$, $\{\mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,\emptyset\emptyset}\}$ respectively gets mark 0 and at least $\frac{1}{6}$ gets mark 1.
3) if $m(\phi(e_1)) = m(\phi(e_2))$ and $e_1 \neq \pi^c(e_2)$ for some $c \geq 0$, then the indicators $\phi(e_1)$ and $\phi(e_2)$ are independent. The same is true for $\psi_{1,2}$ terms.

Proof: We start by proving the second clause of the lemma. At each iteration, out of eight considered start/end indicators (four starts and four ends) at least two and at most six have type ψ. Out of these six, at least one is marked with 0 at step two and at least one with 1 at step four (which exactly amounts to at least $\frac{1}{6}$ of the considered subset). If we are in the case of no more chains starting/ending from an indicator ϕ, we mark every second chain-start with 0. In this case, at least $\frac{1}{4}$ of indicators of type ψ is marked with 0. The same argument is true for mark 1.

Now we proof the first clause. Consider indicators $\{\phi(e)\}$ of pairs $\{\mathcal{E}_{\emptyset,M*} \cup \mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,MM}\}$. For the indicators induced by cycles, we start numbering with 0, and alternating 0 and 1. Thus approximately (depending if we stopped at 0 or 1) half of pairs is marked with 0 and the rest is marked with 1. For the chains, at least $\frac{1}{6}$ start/end indicators of type ϕ marked with 1 and same for mark 0 (The argument here is the same as for indicators of pairs of type ψ.). For internal indicators, as we alternate the start counter at each iteration, at least $\frac{1}{3}$ of the indicators is marked with 0 and at least $\frac{1}{3}$ of the indicators is marked with 1.

The last independence statement follows directly from the definition of the chains and cycles. ∎

For simplicity, we write $m(e) = 0/1$ meaning $m(\phi(e)) = 0/1$ or $m(\psi(e)) = 0/1$.

Using this marking, we split the X_π into two sums: $X_\pi = S_1 + S_2$ where

$$S_1 = \sum_{\substack{e \in \mathcal{E}_{\emptyset,M*} \cup \mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,MM} \\ m(e)=0}} \phi(e) + \alpha \left[\sum_{\substack{e \in \mathcal{E}_{M,\emptyset M} \cup \mathcal{E}_{M,\emptyset\emptyset} \\ m(e)=0}} \psi_1(e) + \sum_{\substack{e \in \mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,\emptyset\emptyset} \\ m(e)=0}} \psi_2(e) \right]$$

and

$$S_2 = \sum_{\substack{e \in \mathcal{E}_{\emptyset,M*} \cup \mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,MM} \\ m(e)=1}} \phi(e) + \alpha \left[\sum_{\substack{e \in \mathcal{E}_{M,\emptyset M} \cup \mathcal{E}_{M,\emptyset\emptyset} \\ m(e)=1}} \psi_1(e) + \sum_{\substack{e \in \mathcal{E}_{M,M\emptyset} \cup \mathcal{E}_{M,\emptyset\emptyset} \\ m(e)=1}} \psi_2(e) \right]$$

Lemma 12: We have

$$\mathrm{E}[S_1] \geq \frac{\mathrm{E}[X_\pi]}{6} \text{ and } \mathrm{E}[S_2] \geq \frac{\mathrm{E}[X_\pi]}{6}.$$

Proof: This follows directly from Lemma 11 and linearity of expectation. ∎

Lemma 13: Denote by $\mu_1 = \mathrm{E}[X_\pi]$ and by $\mu_2 = \mathrm{E}[Y_\pi]$.

$$\mathrm{P}[X_\pi < \frac{\mu_1 + \mu_2}{2}] \leq 2\exp(-\frac{(\mu_1 - \mu_2)^2}{96\mu_1})$$

$$\mathrm{P}[Y_\pi > \frac{\mu_1 + \mu_2}{2}] \leq \exp(-\frac{(\mu_1 - \mu_2)^2}{12\mu_1})$$

Proof: As $X_\pi = S_1 + S_2$, then

$$\begin{aligned} &\mathrm{P}[X_\pi < (1-\epsilon)\mu_1] \\ &\quad \leq \mathrm{P}\left[S_1 < (1-\epsilon)\mathrm{E}[S_1] \bigcup S_2 < (1-\epsilon)\mathrm{E}[S_2]\right] \\ &\quad \leq \mathrm{P}\left[S_1 < (1-\epsilon)\mathrm{E}[S_1]\right] + \mathrm{P}\left[S_2 < (1-\epsilon)\mathrm{E}[S_2]\right]. \end{aligned}$$

We prove that $\mathrm{P}[S_1 < (1-\epsilon)\mathrm{E}[S_1])$ (the proof for S_2 is similar). As the result of Lemma 11, all the terms in S_1 are independent except the case where in a cycle the beginning and the end indicators have the same mark. For those cycles $\phi(e_1), \ldots, \phi(e_c)$, we introduce a new variable $W_{e_1} = \frac{\phi(e_1)+\phi(e_c)}{2}$ and for the rest of indicators we define $W_e = \frac{\phi(e)}{2}$. Note that if $W = \sum W_{e_i}$, then $2W = S_1$ and all W_e terms are independent.

$$\begin{aligned} &\mathrm{P}\left[S_1 < (1-\epsilon)\mathrm{E}[S_1]\right] = \mathrm{P}\left[\sum W_i < (1-\epsilon)\mathrm{E}[W]\right] \\ &\leq \exp\left(-\frac{\mathrm{E}[W]\epsilon^2}{2}\right) \text{ (by a Chernoff-Hoeffding bound 10)} \\ &\leq \exp\left(-\frac{\mathrm{E}[S_1]\epsilon^2}{4}\right) \leq \exp\left(-\frac{\mathrm{E}[X_\pi]\epsilon^2}{24}\right) \text{ (by Lemma 12)} \end{aligned}$$

To sum up, we put $\epsilon = \frac{\mu_1 - \mu_2}{2\mu_1}$, note that $\frac{\mu_1+\mu_2}{2} = \mu_1 - \frac{\mu_1-\mu_2}{2} = \mu_1(1 - \frac{\mu_1-\mu_2}{2\mu_1})$, and obtain the bound for X_π. For μ_2 we can write similarly $\frac{\mu_1+\mu_2}{2} = \mu_2(1+\frac{\mu_1-\mu_2}{2\mu_1})$. The result for Y_π follows directly from a Chernoff bound because all the terms are independent. ∎