Assembling a Network out of Ambiguous Patches

Lyudmila Yartseva[1] , Jefferson Elbert Simões[2] and Matthias Grossglauser[1]

***Abstract*— Many graph mining and network analysis problems rely on the availability of the full network over a set of nodes. But inferring a full network is sometimes non-trivial if the raw data is in the form of many small *patches* or subgraphs, of the true network, and if there are ambiguities in the identities of nodes or edges in these patches. This may happen because of noise or because of the nature of data; for instance, in social networks, names are typically not unique. *Graph assembly* refers to the problem of reconstructing a graph from these many, possibly noisy, partial observations. Prior work suggests that graph assembly is essentially impossible in regimes of interest when the true graph is Erdős-Rényi . The purpose of the present paper is to show that a modest amount of clustering is sufficient to assemble even very large graphs.**

We introduce the $G(n, p; q)$ random graph model, which is the random closure over all open triangles of a $G(n, p)$ Erdős-Rényi , and show that this model exhibits higher clustering than an equivalent Erdős-Rényi . We focus on an extreme case of graph assembly: the patches are small (1-hop egonets) and are unlabeled. We show that in realistic regimes, graph assembly is fundamentally feasible, because we can identify, for every edge e, a subgraph induced by its neighbors that is unique and present in every patch containing e. Using this result, we build a practical algorithm that uses canonical labeling to reconstruct the original graph from noiseless patches. We also provide an achievability result for noisy patches, which are obtained by edge-sampling the original egonets.

I. Introduction

Network data describes relationships between entities, which has many downstream uses for inference and prediction tasks. For example, community detection can reveal social communities or security threats [7]; centrality measures can reveal influences or weaknesses in an organization [6]; source detection algorithms can reveal the instigator of a rumor or patient zero of an epidemic [16]. Obviously, a precondition for such statistical methods to perform well is that the network obtained from raw data is reliable.

This work is about inferring a network from raw data where node identities are ambiguous or even absent. This is a difficult and relevant problem because networks often need to be *assembled* from a large set of observations in the form of small subgraphs. For example, the structure of an IP network can be derived from the routing/forwarding tables of all the routers in a domain; and a scientific co-authorship or co-citations network is the union over the subgraphs revealed by each paper. If the node identities in such a collection of subgraphs are reliable and unambiguous, assembling the true network is trivial: a good estimate for the true network is the union of the subgraphs. If every edge in the true network appears in at least one subgraph, then this estimator is indeed exact.

Unfortunately, in practice, we cannot always rely on unambiguous node identities from one observed subgraph to another. Consider several scenarios where network assembly under node ambiguity is necessary. For instance, suppose we are given a corpus of text with many different authors, each describing social interactions and transactions among their social contacts. Each author might use ambiguous identifiers for the protagonists, e.g., first name, nickname, or some descriptive reference. From this we want to reconstruct the full social network. This situation arises in social network analysis, such as in digital humanities [12]. Another example stems from efforts to anonymize sensitive network information. If a full social network cannot be released, out of concern that this network could be deanonymized, one protection mechanism that has been used in the literature is the release of all the 1-hop egonets of this network, with all node identities withheld [5].

In this paper, we are chiefly interested in the following problem: if labels provide little or no information to disambiguate nodes, to what extent is the structure of observed subgraph sufficient for reassembly? This problem is relatively unexplored. Although heuristic algorithms are presented in [9], [20], no guarantees are provided. In the field of database mining, particularly entity resolution, several questions of ambiguities in data are well studied [4]. However they are mostly based on the similarity between labels of the entities and rely on structural information as a secondary means. There has also been some work in the field of pattern discovery [1], but the authors focus on the problem of approximate labeling of the nodes and look for patterns that minimize the cost of such labeling, rather than using the graph structure. These results are oriented overall by the design of constrained algorithms, rather than by the investigation of theoretical feasibility.

Under partial or full node ambiguity, reassembling the true graph from small subgraphs (called *patches*) is an interesting statistical and computational problem. It is related to the reconstruction conjecture formulated by Kelly [14], which addresses the question of a graph G being uniquely identifiable by all its subgraphs obtained by deleting one vertex from G. A closely related problem was considered most recently by Mossel et al. [17], who are also interested in graph assembly problem, but for low clustered graphs, such as an Erdős-Rényi or random regular graphs. They analyze the size of the neighborhoods and state theoretical thresholds for the feasibility of the assembly. They find that the size of the neighborhoods has to be quite large for assembly to be

[1]School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
firstname.lastname@epfl.ch
[2]Systems Engineering and Computer Science Program, COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
elbert@land.ufrj.br

feasible.

Real networks are very different from random graphs. In particular, they tend to have a clustering coefficient that is much higher than a random graph of corresponding density. This implies that there is richer local structure, i.e., short cycles including triangles. The purpose of the present paper is to show that this local structure can be exploited in reassembly, and to successfully stitch together small neighborhoods. We introduce a random graph model that is, to the best of our knowledge, novel and potentially of independent interest. The model generates a graph whose edge set is formed by the random closures of open triangles (three nodes with two incident edges) of an underlying ER graph. We use this model to prove that relatively sparse graphs can be assembled from small distance-1 neighborhoods.

Main Results and Outline

To explain the assembly problem, we begin by assuming that we are given a collection of patches. We restrict our problem to the case where each patch is a local neighborhood of a center node, called 1-egonet. The network assembly problem is to (i) infer the true graph from the collection of patches and (ii) map each node in each patch to the correct node.

As mentioned above, if the true graph is Erdős-Rényi and the patches are 1-egonets around every node, assembly is almost always impossible [17]. In this case, it means the set of 1-egonets collapses into a small set of classes. For sparse reghimes that is np being a constant, the number of hops of the neighborhood r should be $\log n$. For denser reghimes $np \gg \log^2 n$ assembly requires at least $r = 3$-hop egonets. The idea is that in this case each node has unique neighbors-degree sequence, thus this node is identifiable in other egonets if it is connected to a center. We claim that $r = 3$ is required because of the lack of transitivity (short cycles) in such graphs and the assembly is feasible even for $r = 1$ if the graph is clustered. In Section III, we introduce a new random graph model $G(n, p; q)$ of independent interest, where we generate random closures with probability q over an Erdős-Rényi $G(n, p)$ generator. This process of triangle closure is known to be natural property of most real networks, such as social networks [15].

In our model, each node of the generator graph generates some connected community around assuring clustered structure of the network. We show that this graph has an asymptotically larger clustering coefficient[1] (equal $\frac{q}{np}$) than the $G(n, p)$ with the same average degree. And we show that particular induced subgraphs contain a much denser structure; as a result providing patterns for assembling patches.

The assembly problem can rely only on subgraph isomorphism relationships among the patches. At first sight, it might seem that assembly would be extremely challenging, especially if the patches are small. In Section IV, we show instead that, under some mild assumptions, structural information in the patches is sufficient for assembly to succeed. To prove this result, we focus on the induced subgraph over the set of common neighbors of two adjacent nodes u and v. If these subgraphs are not isomorphic to any other such neighborhood graph, then it can be used as a fingerprint to find adjacent nodes. The proposed approach results in a simple and effectively tractable algorithm of network assembly, in Section IV-C.

The last model accounts for more realistic scenarios, where noise is introduced into observations by removing some connections. In Section V, we characterize the amount of noise our model can tolerate in order to still make correct graph assembly feasible based only on the structure. We find that, in this case, the density of the original graph can be similar in magnitude as in the noiseless case, with a small penalty that is a function of the amount of noise introduced.

[1]The clustering coefficient of a node u is the density of the subgraph induced by its neighbors; assumed to be 0 if u is a singleton.

II. Formal Statement of the Network Assembly Problem

In its most general form, the graph assembly problem takes as input a finite collection of graphs called *patches*; these patchers have been extracted from a larger graph that we call *master graph*. The labels of vertices in each patch bears little or no resemblance at all to their original labels in the master graph, and the problem consists of putting these pieces together in an assembled graph $\hat{G}$.

In this work, we will consider a specific variation of this problem, where each patch is created by extracting the egonet around each vertex in master graph. The *egonet*, or 1-egonet of a vertex v in a graph G, denoted G_v, is the induced subgraph generated by v and its neighbors in G — we say that v is the *center* of this egonet. We will further assume that, for each egonet in the patch collection, the identity of v is either kept intact or somehow inferrable, but all other identities are removed.

To accurately model this problem mathematically, we will need a few definitions:

Definition 1 (Egonet extraction): Let G be a graph with $V(G) = [n]$ for some $n \in \mathbb{N}$, and edge set $E(G)$.

- The *egonet collection* of G is the indexed family of graphs $\{G_v\}_{v\in[n]}$;
- An *anonymized egonet collection* of G is a set of graphs $\mathcal{P} = \{G'_v = f_v(G_v)\}_{v\in[n]}$, where $f_v : V(G_v) \to [|V(G_v)|]$ is a bijection such that $f_v(v) = 1$; the functions $\{f_v\}_{v\in[n]}$ are called the *anonymization functions*.

Note that f_v relabels every vertex in G_v arbitrarily, except for v, that is forcefully assigned the label 1. This means that, as long as the indices of each graph in the collection are known, the identities of the respective centers are also known. This relabeled version of G_v is denoted by G'_v.

Definition 2 (Egonet collection assembly): Let $\mathcal{P} = \{G'_v\}_{v\in[n]}$ be a collection of graphs, such that $V(G'_v) = [n_v]$ for some $n_v \in \mathbb{N}$. An *assembly* of $\mathcal{P}$ is a pair $(\hat{G}, \{a_v\}_{v\in[n]})$, where $\hat{G}$ is a graph (called *assembled graph*) with $V(\hat{G}) = [n]$, and each $a_v : [n_v] \to [n]$ is an injective function such that $a_v(1) = v$.

An assembly determines not only which graph $\hat{G}$ is ultimately obtained, but also how each vertex in each egonet of our collection is mapped to $\hat{G}$. This is enough for us to formally state the *egonet assembly* problem:

- *Input:* an anonymized egonet collection $\mathcal{P} = \{G'_v\}_{v\in[n]}$;
- *Output:* an assembly $(\hat{G}, \{a_v\}_{v\in[n]})$ of $\mathcal{P}$.

Clearly, we would like to have the assembled graph $\hat{G}$ equal to the master graph G. Indeed, if $\mathcal{P} = \{G'_v\}$ is an anonymized egonet collection of G — that is, for every $v \in [n]$, $G'_v = f_v(G_v)$ — it is not difficult to see that $(G, \{f_v^{-1}\})$ is a valid assembly of $\mathcal{P}$. The interesting theoretical question is whether G is the only graph for which there is such assembly. If this is the case, then the problem of egonet assembly is *feasible*.

III. $G(n,p;q)$ Model

In many real networks, neighborhoods of nodes are highly connected (i.e., have high clustering coefficient). For example, in social networks, friends of any given person are more likely to know each other. This behavior is called triadic closure [19]. We would like to address the question of how a graph's clustering coefficient improves the feasibility of assembly.

For this purpose, we introduce the $G(n,p;q)$ random graph model and analyze its properties. The $G(n,p;q)$ model is defined via an intermediate Erdős-Rényi random graph $G_p(V,E_p) \sim G(n,p)$. The graph $G(V,E) \sim G(n,p;q)$ contains a random subset of all the possible closures of connected triples in G_p. Our goal is to obtain a model that is mathematically tractable (akin to the Erdős-Rényi model [10]), but possesses a higher clustering coefficient.

For convenience, we denote by $P_e = \mathbf{1}_{\{e\in E_p\}}$ the indicators of edges in G_p, with $Q_e = \mathbf{1}_{\{e\in E\}}$ being the indicators of edges in our final graph G. We refer to edges in E_p as *p-edges* and edges in E as *q-edges*.

Define the set of independent $\mathsf{Be}(q)$ random variables $\{T_t\}_{t\in V^3}$, with the restriction $T_{u,g,v} = T_{v,g,u}$. Let t be a connected triple (u,g,v) in G_p, i.e., a pair of incident edges $(u,g),(g,v) \in E_p$. The idea is that, for each such connected triple, we apply triadic closure with probability q (hence each term q-edge), so that the edge $(u,v) \in E$ if and only if $T_{u,g,v} = 1$ for at least one $g \in V \setminus \{u,v\}$ that is connected to both u and v by p-edges. We call

$$S_e = S_{(u,v)} = \{g \in V \ : \ P_{u,g}P_{g,v}T_{u,g,v}\}$$

the set of generators for an edge (u,v). Thus, there is an edge $e \in E$ iff it has at least one generator. Note that E_p and E need not be disjoint.

Remark 1: The following facts hold:

1) For any $e \in \binom{V}{2}$, $|S_e| = \mathsf{Bi}\big(n, p^2q\big)$;[2]
2) For any $e \in \binom{V}{2}$, $Q_e = \mathsf{Be}\big(1-(1-p^2q)^n\big)$.

Some additional useful definitions are as follows: for any $u \in V$, the *neighborhood* N_u of u is the set of vertices adjacent to u in G (thus, via q-edges), with $d_u = |N_u|$ its degree, and the *p-neighborhood* N_u^p of u is the set of vertices adjacent to u in G_p (via p-edges).

We show some key properties of this model. Let c_u be the clustering coefficient of node u.

[2]Since $n \to \infty$, from now, we ommit constant substractions and write $n-1, n-2, \dots$ as n.

Proposition 3: Let $u \in V$ be arbitrary. If $np \to \infty$, $n^2p^3 \to 0$ and q is fixed, then:

- $\mathbb{E}[|E|] \simeq \frac{n^3p^2q}{2}$;
- $\mathbb{E}[d_u] \simeq (np)^2q$;
- $\mathbb{E}[c_u] \simeq \dfrac{q}{np}$.

Proof: See Appendix VIII. ∎

Consider for comparison an Erdős-Rényi random graph $G(n,p')$ with the same expected density. It has an edge probability $p' = np^2q$, average degree of $(np)^2q$, and its expected clustering coefficient is therefore np^2q, which is asymptotically smaller than for the $G(n,p;q)$ model (since $n^2p^3 \to 0$ and $q/np \gg np^2q$).

Another interesting feature of the $G(n,p;q)$ model is that, for a rather general regime of p, all edges have a very limited number of generators.

Lemma 4: For $np \to \infty, n^5p^6 \to 0$ and fixed q, w.h.p.[3], all edges have at most two generators.

Proof: It is enough to show that the expected number of edges with three or more generators vanishes, as this implies the result by the first moment method.

Let p_k denote the probability that an arbitrary edge (u,v) has precisely k generators. Recall from Remark 1 that the generator set of an edge has size $\mathsf{Be}\big(n,p^2q\big)$. This implies

$$p_k = \binom{n}{k}(p^2q)^k(1-p^2q)^{n-k} \le (np^2q)^k.$$

For any edge e, the probability that it has at least 3 generators is given by

$$\begin{aligned}\mathbb{P}(|S_e| \ge 3\}) &= \sum_{k\ge3} p_k \le \sum_{k\ge3}(np^2q)^k \\ &= (np^2q)^3\frac{1}{(1-np^2q)} \simeq (np^2q)^3\end{aligned}$$

where the last steps follow from the convergence of the geometric series for large enough n — note that our hypothesis imply that $np^2 \to 0$. Finally

$$\begin{aligned}\mathbb{E}[\{(u,v) : S(u,v) \ge 3\}] &= \sum_{u,v}\mathbb{P}(|S_e| \ge 3\}) \lesssim \sum_{u,v}(np^2q)^3 \\ &= n^2(np^2q)^3 \simeq n^5p^6q^3 = o(1).\end{aligned}$$

∎

In further we consider the following more restrictive regimes on p and q: $(np)^5p \to 0$, fixed q and $npq^2 = 12\log n + \omega(1)$. These assure the average degree d be larger than $\Omega(\log^2 n)$. The smaller values of q can be considered with more tedious analysis, however even large (const) values of q assure only small increase in clustering coefficient. We demonstrate that this small increase is sufficient to assure feasibility of assembly.

IV. Assembly of Noiseless Egonets

Our goal in this section is to demonstrate that for a certain regime of the parameters p and q, it is feasible to reassemble a collection of noiseless 1-hop egonets extracted from a $G(n,p;q)$ random graph. For this, we will characterize a

[3]*With high probability*, i.e., with probability that tends to 1 as $n \to \infty$.

number of properties that this random graph possesses with a high probability, and these properties will naturally lead to a very intuitive algorithm for reassembling the given egonets.

The intuition behind the result is as follows. Let us assume for a moment that the edges of the master graph are uniquely labeled, and that this labeling is preserved through patch generation process — that is, edges in egonets that correspond to the same edge on the master graph are given the same label. In this case, it is straightforward to reidentify the nodes. For instance, if the edge (u, v) is assigned the unique label 35, then there will be an edge labeled 35 in the egonet of u, which means its other endpoint must be v, since no edge to another node is assigned the label 35. Analogously, we can identify u on the egonet of v.

This observation means that the problem can be solved, as long as we can assign such a consistent labeling to edges between all egos and its respective neighbors. However, we must assign these labels by looking only at the structure of the egos and nothing else. Fortunately, under condition that either u or v is the ego-center, any edge (u, v) has a structural feature that is preserved by the egonet extraction process. This feature is the *induced subgraph of common neighbors* of u and v, which we denote $H_{u,v}$. Note that this feature is symmetric by nature, thus $H_{u,v} \sim H_{v,u}$[4]. As the main result of this section, we show that, for a $G(n, p; q)$ random graph, any two edges have non-isomorphic subgraphs of common neighbors. Therefore this feature acts as a fingerprint for all edges in a graph.

Further we formulate the main result of this section, provide key lemmas and follow by the proof of the result.

Theorem 5: Let G be a graph with node set $[n]$ and unique edge fingerprints, and let $\mathcal{P} = \{G'_v\}_v$ be an anonymized egonet collection extracted from G. There exists an assembly algorithm that builds $\hat{G}$ from the input $\mathcal{P}$ and $V(\hat{G}) = V(G)$ and $E(\hat{G}) = E(G)$.

A. Structural Properties of Patches

To determine when all edges in a $G(n, p; q)$ random graph indeed have unique (up to isomorphism) subgraphs of common neighbors, we must first characterize the structure of these subgraphs. We start by determining the node set of $H_{u,v}$, which we call $N_{u,v}$.

Lemma 6: If G is sampled from $G(n, p; q)$ with $np \to \infty$, $(np)^5 p \to 0, q$ is fixed, then for any fixed $u, v \in G$ such that u is adjacent to v, the following statements hold w.h.p.:

- For each $x \in N_{u,v}$, there exists $g \in S(u, v) \cap S(u, x) \cap S(v, x)$ – i.e., all the edges of the uxv triangle have at least one common generator;
- $|N_{u,v}| = \mathsf{Bi}\big(n, |S(u, v)|pq^2\big)$ and $\mathbb{E}\Big[|N_{u,v}|\Big||S(u, v)|\Big] = |S(u, v)|np^2q$.

See Figure 1 for an illustration of $H_{u,v}$.

Proof: See Appendix IX. ∎

We can now easily characterize the edges between the nodes of the neighborhoods $N_{u,v}$. For any $x, y \in N_{u,v}$, by Lemma 6 there exists $g_1, g_2 \in S(u, v)$ such that $P_{g_1,x} = 1$ and $P_{g_2,y} = 1$. If $g_1 = g_2$, then this triangle is closed

[4]here $\sim$ means graph isomorphism

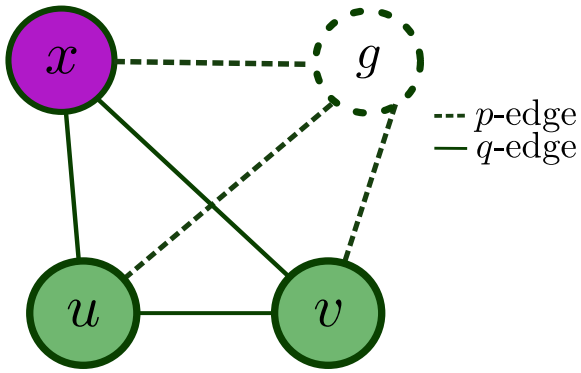

Fig. 1. $x \in N_{u,v}$ is a common neighbour of u and v.

with probability q (independently for each pair x, y), otherwise they are connected with probability np^2q. Recalling Lemma 4 each edge has at most two generators, thus,

Corollary 7: Under the conditions of Lemma 6, w.h.p., one of these cases holds:

1) $|S(u, v)| = 1$ and $H_{u,v}$ is a single Erdős-Rényi graph $G(\mathsf{Bi}\big(n, pq^2\big), q)$;
2) $|S(u, v)| = 2$ and $H_{u,v}$ consists of two Erdős-Rényi graphs $G(\mathsf{Bi}\big(n, pq^2\big), q)$, with each crossing edge existing independently with probability np^2q.

Note that $np^2q \ll q$, hence in the latter case, the two Erdős-Rényi graphs have dense structure, but are very loosely connected.

B. Uniqueness of Edge Fingerprints

We are now ready to prove our key result of this section.

Theorem 8: Let G be a $G(n, p; q)$ random graph, with $(np)^5 p \to 0$, fixed q and $npq^2 = 12 \log n + \omega(1)$. Then, w.h.p., for any pairwise adjacent nodes u, v and $\hat{u}, \hat{v}$, either $\{u, v\} = \{\hat{u}, \hat{v}\}$ or $H_{u,v}$ is not isomorphic to $H_{\hat{u},\hat{v}}$.

Proof: Denote by W the number of quadruples $(u, v, \hat{u}, \hat{v})$, with u and v adjacent, $\hat{u}$ and $\hat{v}$ are adjacent and $(\hat{u}, \hat{v}) \neq (u, v), (v, u)$, such that $H_{u,v}$ is isomorphic to $H_{\hat{u},\hat{v}}$. By the first moment method, it is enough to show that $\mathbb{E}[W] \to 0$. We note that $\mathbb{E}[W] = \sum_{u,v,\hat{v},\hat{u}} \mathbb{P}(H_{u,v} \sim H_{\hat{u},\hat{v}})$.

Now, we fix $u, v, \hat{v}$ and $\hat{u}$ and split our analysis into cases. We consider the most complex case in detail and omit lengthy and similar computations for other cases.

1) $|S(u, v)| = |S(\hat{u}, \hat{v})| = 1$:
   a) $S(u, v) = S(\hat{u}, \hat{v}) = \{g\}$, where g is the common generator of (u, v) and $(\hat{u}, \hat{v})$:
      i) $u = \hat{u}$ and $v \neq \hat{v}$ (or, analogously, $u \neq \hat{u}$ and $v = \hat{v}$):
      Note that any vertex $x \in V$ is in both $H_{u,v}$ and $H_{u,\hat{v}}$ according to the following criteria:
      - $x \in H_{u,v}$ iff $P_{g,x}T_{u,g,x}T_{v,g,x} = 1$
      - $x \in H_{u,\hat{v}}$ iff $P_{g,x}T_{u,g,x}T_{\hat{v},g,x} = 1$

      Let J be the subgraph induced by $\{x \in V : P_{g,x}T_{u,g,x} = 1\}$. By the criteria above, any node that is not in J cannot belong to either $H_{u,v}$ or $H_{u,\hat{v}}$. Note that J is an Erdős-Rényi random graph $G(\mathsf{Bi}(n, pq), q)$ (each node $x \in V$ satisfies $P_{g,x}T_{u,g,x} = 1$ independently with probability

pq, and any two nodes $x, y \in J$ are adjacent if and only if $T_{x,g,y} = 1$, which happens with probability q independently).
Furthermore, each node $x \in J$ belongs to $H_{u,v}$ or $H_{u,\hat{v}}$ if $T_{v,g,x} = 1$ or $T_{\hat{v},g,x} = 1$, and these conditions hold independently with probability q. This means $H_{u,v}$ and $H_{u,\hat{v}}$ are obtained from J by sampling each node independently with probability q. To bound the probability that these two graphs are isomorphic, we apply Lemma 15 with $m = \mathsf{Bi}(n, pq)$ and $t = q$ and any fixed $0 < \delta < 1$, obtaining:

$$\mathbb{P}(H_{u,v} \sim H_{u,\hat{v}}) \leq \exp\big(m \log m - m^2 c\big) + 2\exp\left(-\frac{\delta^2 mq}{2}\right)$$

for $c = (1-\delta)^2 q^2 (1-q) \log c_1$ and $c_1 = (q^2 + (1-q)^2)^{-1} \in (1, 2]$. Note, however, that m is a random variable. We can apply the Chernoff bound (see Appendix 13) to bound $npq(1-\delta) \leq m \leq npq(1+\delta)$. Thus:

$$\begin{aligned}\mathbb{P}(H_{u,v} \sim H_{\hat{u},\hat{v}}) \leq \exp\big(&npq(1+\delta)\log npq(1+\delta) \\ &- (npq)^2(1-\delta)^2 c\big) \\ &+ 4\exp\left(-\frac{\delta^2 npq^2}{3}\right).\end{aligned}$$

ii) $u \neq \hat{u}$ and $v \neq \hat{v}$. The case is analogous to the previous one, except that both $H_{u,v}$ and $H_{\hat{u},\hat{v}}$ are obtained by node sampling the graph $J' = N_s^p$; J' is a $G(\mathsf{Bi}(n,p), q)$ random graph and each node is sampled with probability q^2 to obtain both $H_{u,v}$ and $H_{\hat{u},\hat{v}}$. Thus, by Lemma 15:

$$\begin{aligned}\mathbb{P}(H_{u,v} &\sim H_{\hat{u},\hat{v}}) \\ &\leq \exp\big(npq^2(1+\delta)\log npq^2(1+\delta) \\ &\quad - (npq)^2(1-\delta)^2 c\big) \\ &\quad + 4\exp\left(-\frac{\delta^2 npq^2}{3}\right).\end{aligned}$$

b) $S(u,v) \neq S(\hat{u},\hat{v})$. Denote $S(u,v) = \{g_1\}$ and $S(\hat{u},\hat{v}) = \{g_2\}$. Since $N^p_{g_1} \cap N^p_{g_2} = \emptyset$ holds w.h.p, it also holds that $N_{u,v} \cap N_{\hat{v},\hat{u}} = \emptyset$ (by Lemma 6). Then, by the reasoning similar to that in Lemma 15, $\mathbb{P}(H_{u,v} \sim H_{\hat{u},\hat{v}}) \leq m!(q^2 + (1-q)^2)^{\binom{m}{2}}$.

2) $|S(u,v)| = 2$. Denote $S(u,v) = \{g_1, g_2\}$. In this case, $H_{u,v}$ consists of two weakly connected Erdős-Rényi graphs $H^1_{u,v} \cup H^2_{u,v}$, and similarly for $H_{\hat{u},\hat{v}}$. Thus we can estimate the probability of $H^i_{u,v}$ being isomorphic to the analogous component of $H_{\hat{u},\hat{v}}$ and thus reduce the problem to the previous case.

Using the loosest bound of the previous cases, we can bound $\mathbb{E}[W]$:

$$\begin{aligned}\mathbb{E}[W] \leq n^4\Big(&\exp\big(npq(1+\delta)\log npq(1+\delta) \\ &- (npq)^2(1-\delta)^2 c\big) \\ &+ 4\exp\left(-\frac{\delta^2 npq^2}{3}\right)\Big)\end{aligned}$$

$$\begin{aligned}&= \exp\big(4\log n + npq(1+\delta)\log npq(1+\delta) \\ &\quad - (npq)^2(1-\delta)^2 c\big) \\ &\quad + 4\exp\left(4\log n - \frac{\delta^2 npq^2}{3}\right)\end{aligned}$$

the last summand dominates and thus whole sum goes to 0 if $p \geq \frac{12\log n + \omega(1)}{\delta^2 q^2 n}$. Note that if $q \in (0,1)$ and $\delta < 1$, c is constant. ∎

C. Feasibility of Egonet Assembly

The results leading to Theorem 8 enable us to design a simple assembly algorithm that works as follows. Let $\mathcal{P} = \{G'_v\}$ be the anonymized egonet collection of a graph $G = (V, E)$, which is the graph we want to obtain at the end of the assembly process; and also assume that all edges in G have unique fingerprints, that is, for any two distinct edges $(u,v), (\hat{u},\hat{v}) \in E$, $H_{u,v}$ and $H_{v,u}$ are not isomorphic.

G must have $[n]$, the index set of $\mathcal{P} = \{G'_v\}_v$, as its node set, so we begin by setting $[n]$ as the vertex set of our assembled graph $\hat{G}$. To construct its edge set $E(\hat{G})$, choose a node $u \in [n]$. We know that u is present in egonet G'_u and has been assigned the label 1 in G_u. Take a node $j \in G'_u$ other than 1. Edge $(1, j)$ is the image of some edge (u, v) in G, and the subgraph of G'_u induced by 1, j and their common neighbors is a relabeled version of $H_{u,v}$. Extract this fingerprint from G'_u and search for a second edge, in a different egonet, with an isomorphic fingerprint. Since fingerprints of edges in G are unique, there will be exactly one such edge, say $(1, k)$ on the egonet $G'_{u'}$, and both of them must have originated from the same edge on the master graph. The labels of this edge must be the egonet centers u and u' of the two matching edges, so we add the edge (u, u') to $E(\hat{G})$. Repeat this for all egonets until they are exhausted, at which point the algorithm terminates. We will call this the *fingerprint assembly algorithm*.

If all edges in G have unique fingerprints, this algorithm will always reassemble G correctly:

Proof: [Theorem 5] Assume $E(\hat{G}) \neq E(G)$. Then, one of the following cases must hold:

- There is a pair of vertices (u, v) such that $(u, v) \in E(G) \setminus E(\hat{G})$. $(u, v) \in E(G)$ originates two image edges in $\{G_v\}$, $(1, j) \in E(G'_u)$ and $(1, k) \in E(G'_v)$ for some j, k, with the same fingerprint in their respective patches as (u, v). Since no other edge in G has such fingerprint, these two images must be matched by the algorithm. This implies (u, v) will be added to $E(\hat{G})$, which contradicts $(u, v) \notin E(\hat{G})$.
- There is a pair of vertices (u, v) such that $(u, v) \in E(\hat{G}) \setminus E(G)$. Since (u, v) has been added to $E(\hat{G})$ by the algorithm, then there is an edge $(1, j) \in E(G'_u)$ and an edge $(1, k) \in E(G'_v)$ with matching fingerprints, where $u \neq v$. $(1, j)$ in G'_u is the image of an edge $(u, \hat{v}) \in E(G)$ for some $\hat{v} \in [n]$, and $(1, k)$ in G_v is the image of an edge $(v, \hat{u}) \in E(G)$ for some $\hat{u} \in [n]$, from which they have extracted their fingerprints. Since G has unique edge fingerprints, $(u, \hat{v})$ and $(v, \hat{u})$ must be the same edge, which implies $\hat{v} = v$, $\hat{u} = u$ and, as a result, $(u, v) \notin E(\hat{G})$, which contradicts $(u, v) \notin E(G)$.

■

Corollary 9: Let G be a $G(n,p;q)$ random graph with $(np)^5 p \to 0$, fixed q and $npq^2 = 12\log n + \omega(1)$, and let $\mathcal{P}$ be an anonymized egonet collection extracted from G. If $\hat{G}$ is the output graph of the fingerprint assembly algorithm with input $\mathcal{P}$, then $E(\hat{G}) = E(G)$ w.h.p.

Note that this algorithm requires $\binom{|E|}{2}$ checks for graph isomorphism. This is, in general, a computationally expensive procedure even after recent improvements, with the best known algorithm having quasi-polynomial time complexity [3]. With an oracle for the graph isomorphism problem, the average case complexity of this algorithm drops to around $|E|(npq^2) + |E|^2$, from the subgraph extraction process and the checks for graph isomorphism, respectively. Any technique for optimizing the graph isomorphism step, such as applying approximate graph isomorphism techniques, can be used to reduce its running time. Additionally, if this step is solved by constructing an isomorphism whenever possible, one can use the information given by this isomorphism to further reduce the number of fingerprints comparisons.

Our implementation of this algorithm uses canonical labeling methods to check for subgraph isomorphism. A *canonical labeling* is a labeling of the graph's vertices that uniquely captures the structure of the graph, and two graphs are isomorphic if and only if their canonical forms are precisely equal. The problems of canonization and isomorphism are similar in both theory and algorithm design, even though it is not known whether they are poly-time equivalent [2].

Our implementation has an additional optimization step: Instead of searching through all edges in the egonets looking for edges with isomorphic fingerprint, we convert each fingerprint to an integer value. These integers are extracted from the canonical form of the fingerprint and are therefore graph invariants. Afterwards these edges are stored in a hash map where we use the corresponding integerfingerprints as the search key; thus, reducing the pairwise search for isomorphic fingerprints to a scan over the hash map for edges with matching keys. This optimization reduces the algorithm complexity from $\binom{|E|}{2}$ checks for isomorphism to $|E|$ calculations of canonical forms, at the cost of $|E|$ additional graph-to-integer conversions. Although eventual hash collisions can in principle insert noise in our fingerprint comparison, we do not expect such collisions to be frequent. Additional graph invariants, such as number of edges, can also be extracted from the fingerprints to disambiguate even further in case of eventual collisions, but we choose not to use them in our implementation.

We implement the algorithm by using the canonical labeling procedure from the Bliss library [11]. This library provides us with a hash calculation procedure, that we use to convert fingerprints to integer values. Additional collisions can result from this, and the same mitigation techniques described previously can also be applied here. We ran a set of experiments for finite graphs sampled from $G(n,p;q)$ model and found out that the algorithm can restore all the edges with precision 1.

We do not focus on developing the most efficient algorithm in this paper, hence we do not set up an extensive experiment set with different theoretical and practical models. One of the interesting future directions would be to consider real and artificial noisy data-models and to develop an approximate assembly algorithm for this. Here we are more interested in the feasibility of graph reconstruction from very poor additional information. And the experiments fully support the theoretical results: for graphs sampled from $G(n,p;q)$ model the edge fingerprints are unique, thus assuring feasibility of assembly.

V. Assembly of Noisy Egonets

In realistic scenarios, we often deal with imperfect patches. For instance, the observations of a user's circle in social networks can be noisy. In contrast with the noiseless case, perfect(no edge mismatch) assembly can no longer be expected by default. Rather, we should expect that, in low-noise scenarios, the correct assembly has a small number of edge mismatches, due to the correlation induced in the patch collection by the true graph. Therefore, we intuitively expect the correct assembly to have minimum edge inconsistency.

To evaluate this hypothesis, we consider the following variation of our problem. For each egonet G_u around a node u, extracted from a master graph G, we generate a noisy observation G_u^* by keeping the original node set but sampling edges independently with probability s. As in the noiseless case, the problem is to assembly a master graph G from an anonymized collection $\mathcal{P} = \{f_v(G_u^*)\}$.

In order to show that the hypothesis is true under certain conditions, we prove a result analogous to Theorem 8. This result can be expressed in terms of the *edge mismatch* between two graphs.

Definition 10: Let $H_1(V_1,E_1)$ and $H_2(V_2,E_2)$ be two graphs with $|V_1| = |V_2|$ and let π be a bijection between V_1 and V_2. The *edge mismatch* of mapping H_1 and H_2 by π, denoted by $\Delta(H_1,H_2,\pi)$, is given by: $\Delta(H_1,H_2,\pi) = \sum_{(u,v)\in\binom{V_1}{2}} \mathbf{1}_{\{(u,v)\in E_1 \otimes (\pi(u),\pi(v))\in E_2\}}$

Furthermore, for two neighboring nodes u and v, we denote by $H_{u,v}^*$ the subgraph of G_u^* induced by common neighbors of u and v. Note that $H_{u,v}^*$ and $H_{v,u}^*$ have the same node sets[5]. However, since both G_u^* and G_v^* are noisy egonet-observations, it does not hold in general that $H_{u,v}^* = H_{v,u}^*$, which differs from the noiseless case, where $H_{u,v} = H_{v,u}$ by construction. Rather, $H_{u,v}^*$ and $H_{v,u}^*$ are both subgraphs of $H_{u,v}$.

Lemma 11: Let G be a random graph generated by the $G(n,p;q)$ model with $\frac{\log n}{s^3} \ll np \ll n^{1/5}$ and q be fixed. Then, w.h.p., for any pairwise-adjacent nodes u,v and $\hat{u},\hat{v}$, either $\{u,v\} = \{\hat{u},\hat{v}\}$ or $\Delta(H_{u,v}^*, H_{v,u}^*, \pi_0) < \Delta(H_{u,v}^*, H_{\hat{v},\hat{u}}^*, \pi)$ for any bijection π (with π_0 being the identity mapping over N_{uv}).

Proof: Analogously to Theorem 8, denote by W the number of quadruples $(u,v,\hat{u},\hat{v})$, with u and v adjacent, $\hat{u}$ and $\hat{v}$ are adjacent and $(\hat{u},\hat{v}) \neq (u,v),(v,u)$, such that $\Delta(H_{u,v}^*, H_{v,u}^*, \pi_0) \geq \Delta(H_{u,v}^*, H_{\hat{v},\hat{u}}^*, \pi)$ for some bijection π. We show that $\mathbb{E}[W] \to 0$.

Fix u, v, $\hat{u}$ and $\hat{v}$ and consider only the case $S(u,v) = S(\hat{u},\hat{v}) = \{g\}$, $u = \hat{u}$ and $v \neq \hat{v}$ — other cases as

[5]In the regimes further considered, these graphs remain connected

broken down in the proof of Theorem 8 will be omitted, as they yield stricter bounds. Our goal is to bound $\mathbb{P}\Big(\Delta(H^*_{u,v}, H^*_{v,u}, \pi_0) \geq \Delta(H^*_{u,v}, H^*_{\hat{v},\hat{u}}, \pi) \text{ for some } \pi\Big)$.

Denote $J = \{x \in V \text{ s.t. } P_{g,x}T_{u,g,x} = 1\}$. Nodes in J will belong to $H^*_{u,v}, H^*_{\hat{v},u}$ independently with probability q, with $H^*_{v,u}$ and $H^*_{u,v}$ sharing all nodes. Furthermore, all edges between nodes in these sets will show up in the corresponding graphs with probability s, independently. Thus, all three graphs can be seen as a two-step sample of the subgraph induced by J, with the first step removing nodes and the second one removing edges, and $H^*_{u,v}$ and $H^*_{v,u}$ sharing the first sampling step. Please refer to Figure 3 in Appendix X for an illustration of this process.

We further assume that $|H^*_{u,v}| = |H^*_{\hat{v},u}| = m$, as otherwise there are no bijections between the node sets of these graphs and the quadruple is not counted in W by default. We can now apply Lemma 16 from Appendix X, with G in the statement equal to J in this construction, some $\delta \in (0,1)$ and $t = q$. Note that J is an Erdős-Rényi random graph $G(m,q)$.

$$\begin{aligned}
&\mathbb{P}\big(\exists\pi : \Delta(H^*_{u,v}, H^*_{v,u}, \pi_0) > \Delta(H^*_{u,v}, H^*_{\hat{v},u}, \pi)\big) \\
&\leq 2\sum_{k=2}^{\infty} m^k \exp\left(k\left(\log m - \frac{mps^3}{16}\right)\right) \\
&\quad + \exp\left(-\frac{\delta^2 m(1-q)}{2}\right)
\end{aligned}$$

Recall that, by Lemma 6, $m = |N_{u,v}| = \mathsf{Bi}\big(n, pq^2\big)$. Chernoff bound (Lemma 13) implies $(1-\delta)npq^2 \leq m \leq (1+\delta)npq^2$, so summing over all quadruples yields

$$\begin{aligned}
\mathbb{E}[W] &\leq \sum_{u,v,\hat{v},\hat{u}} \Bigg[2\sum_{k=2}^{\infty} \exp\big(k\big(\log npq^2(1+\delta) \\
&\qquad - \frac{npq^2 s^3(1-\delta)}{16}\Big)\Big) \\
&\qquad + 3\exp\left(-\frac{\delta^2 npq^2(1-q)}{3}\right)\Bigg] \\
&\leq 2\sum_{k=2}^{\infty} \exp\big(4\log n + k\big(\log npq^2(1+\delta) \\
&\qquad - \frac{npq^2 s^3(1-\delta)}{16}\Big)\Big) \\
&\qquad + 3\exp\left(4\log n - \frac{\delta^2 npq^2(1-q)}{3}\right)
\end{aligned}$$

Under this theorem's assumptions, the right side vanishes, thus concluding the proof. ∎

A. Noisy Fingerprint Algorithm

Based on Lemma 11 there is a variation of the fingerprint assembly algorithm that can be used to assemble a collection of noisy egonets $\{G^*_v\}_{v\in[n]}$. The *noisy-fingerprint algorithm* takes $\{G^*_v\}$ as input and proceeds like the fingerprint assembly algorithm, except for the following modification: For each egonet $G{*}_u$ and each node $j \neq 1$ in $G{*}_v$, we match it to an edge $(1,k)$ on an egonet $G'_{u'}$, but we change the criteria "both fingerprints match exactly" to the criteria "edge mismatch between both fingerprints is minimized".

Just like in the noiseless scenario, this algorithm is able to completely assemble the original graph G.

Theorem 12: Let G be a $G(n,p;q)$ random graph with $\frac{\log n}{s^3} \ll np \ll n^{1/5}$ and q fixed, and let $\mathcal{P} = \{G^*_v\}_{v\in[n]}$ be an anonymized noisy-egonet collection extracted from G. If $\hat{G}$ is the output graph of the noisy-fingerprint algorithm with input $\mathcal{P}$, then $E(\hat{G}) = E(G)$ w.h.p.

Proof: Assume $E(\hat{G}) \neq E(G)$. Then, one of the following cases must hold:

- There is a pair of vertices (u,v) such that $(u,v) \in E(G) \setminus E(\hat{G})$. $(u,v) \in E(G)$ implies v corresponds to a node j in G^*_u, and u corresponds to a node k in G^*_v, for some j,k. By Lemma 16, w.h.p. the fingerprint of $(1,k)$ in G^*_u has minimum edge mismatch to the fingerprint of $(1,j)$ in G^*_v, which implies the algorithm will add an edge (u,v) to $E(\hat{G})$. This contradicts $(u,v) \notin E(\hat{G})$.
- There is a pair of vertices (u,v) such that $(u,v) \in E(\hat{G}) \setminus E(G)$. Since (u,v) has been added to $E(\hat{G})$ by the algorithm, then there are pairs of vertices $(1,j)$ in G^*_u and $(1,k)$ in G^*_v, with $u \neq v$, such that the fingerprint of $(1,k)$ in G^*_v has minimum edge mismatch to the fingerprint of $(1,j)$ in G^*_u. Since j is present at G^*_u, then j must correspond to some node $\hat{v}$ which is a neighbor of u. Similarly, k in G^*_v corresponds to some node $\hat{u}$ which is a neighbor of v. Lemma 16 implies edges $(u,\hat{v})$ and $(v,\hat{u})$ must be the same edge (u,v) w.h.p., which contradicts $(u,v) \notin E(G)$.

∎

However, unlike the fingerprint-assembly algorithm, the noisy-fingerprint algorithm has no trivial efficient implementation. The main reason is that the subgraph isomorphism subroutine must be replaced by the calculation of a minimum inconsistency between the input subgraphs, which is a computationally expensive task to which there is no known efficient approximation, to the best of our knowledge. A practical approximation that warrants some interest is to use the optimized form of the algorithm that has been implemented with the Bliss library, but use a locality-sensitive hash function over labeled graphs to store all subgraphs in our hash map. This way, the task of searching for graphs with similar topology (i.e., similar fingerprints) would be reduced to determining entries that are closely located in this hash map.

VI. Discussion

We stated two results that characterize the regimes where complete graph reconstruction, by using only the structure of very small, ambiguous patches, is feasible. We have shown how the relatively high transitivity of the $G(n,p;q)$ random-graph model leads to the existence of features in egonet patches, these features are used to guide the assembly of the egonet collection by a very intuitive algorithm. To the best of our knowledge, this is the first work to apply a purely structural approach to this problem.

We have also shown that such an assembly is still feasible if the patches in our collection are noisy observations of

egonets. The conditions required on the model's parameters are stronger but only slightly: the lower bounds imposed on the average degree of the intermediate graph G_p differ only by a multiplicative constant and a penalty term of s^{-3} due to the noise parameter s.

It is important to highlight that the focus of this work is not on particular algorithms for solving the graph assembly problem, which can be fine-tuned according to known features of networks from any application domain. Rather, we evaluate the impact of a fundamental network property — its clustering coefficient — on the theoretical feasibility of solving the graph assembly problem. Proposing the $G(n,p;q)$ random graph model, instead of using known models such as the Watts-Strogatz model, enabled us to investigate this matter with a minimal distance from the $G(n,p)$ model. This model can be thought of as a baseline, mean-field random model with little structural correlation, and it leads to an effective study of clustering coefficient in isolation from other network properties. Therefore, although the specific stated theorems are not relevant to scenarios involving real networks, the abstract conclusions are general enough to be relevant in these scenarios. Further work with alternative network models, with the clustering coefficient as a controlled, independent parameter, should strengthen our conclusions.

REFERENCES

[1] Pranay Anchuri, Mohammed J. Zaki, Omer Barkol, Shahar Golan, and Moshe Shamy. Approximate graph mining with label costs. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, IL, USA, August 2013.

[2] V. Arvind, Bireswar Das, and Johannes Köbler. The space complexity of *k* -tree isomorphism. In *Proceedings of the 18th International Symposium on Algorithms and Computation*, Sendai, Japan, December 2007.

[3] László Babai. Graph isomorphism in quasipolynomial time. *ArXiv e-prints*, December 2015.

[4] Indrajit Bhattacharya and Lise Getoor. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[5] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for Development: the D4D Challenge on Mobile Phone Data. *ArXiv e-prints*, September 2012.

[6] Phillip Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5), 1987.

[7] Oliver Brdiczka, Juan Liu, Bob Price, Jianqiang Shen, Abhijit Patil, Richard Chow, Evgeniy Bart, and Nicolas Ducheneaut. Proactive insider threat detection through graph learning and psychological context. In *2012 IEEE Symposium on Security and Privacy Workshops*, San Francisco, CA, USA, May 2012.

[8] Devdatt Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

[9] Dóra Erdős, Rainer Gemulla, and Evimaria Terzi. Reconstructing graphs from neighborhood data. *ACM Transactions on Knowledge Discovery from Data*, 8(4), 2014.

[10] Paul Erdős and Alfred Rényi. On random graphs I. *Publicationes Mathematicae (Debrecen)*, 6, 1959.

[11] Tommi A. Junttila and Petteri Kaski. Engineering an efficient canonical labeling tool for large and sparse graphs. In *Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments and the Fourth Workshop on Analytic Algorithms and Combinatorics*, New Orleans, LA, USA, January 2007.

[12] Frédéric Kaplan. The venice time machine. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, Lausanne, Switzerland, September 2015.

[13] Ehsan Kazemi, Lyudmila Yartseva, and Matthias Grossglauser. When can two unlabeled networks be aligned under partial overlap? In *53rd Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, September 2015.

[14] Paul J. Kelly. A congruence theorem for trees. *Pacific Journal of Mathematics*, 7(1), 1957.

[15] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, NV, USA, August 2008.

[16] Wuqiong Luo and Wee-Peng Tay. Estimating infection sources in a network with incomplete observations. In *2013 IEEE Global Conference on Signal and Information Processing*, Austin, TX, USA, December 2013.

[17] E. Mossel and N. Ross. Shotgun assembly of labeled graphs. *ArXiv e-prints*, April 2015.

[18] Pedram Pedarsani and Matthias Grossglauser. On the privacy of anonymized networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA, August 2011.

[19] C. Seshadhri, Ali Pinar, Nurcan Durak, and Tamara G. Kolda. Directed closure measures for networks with reciprocity. *Journal of Complex Networks*, early access, 2016.

[20] Kumar Sharad and George Danezis. An automated social graph de-anonymization technique. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, Scottsdale, AZ, USA, November 2014.

APPENDIX

VII. USEFUL RESULTS

A. Concentration Lemmas

Lemma 13: [Chernoff-Hoeffding bound [8]] 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 $0 < \varepsilon < 1$,

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

Lemma 14: [Difference of Binomials [18]] Let X_1 and X_2 be two binomial random variables with means λ_1 and λ_2, where $\lambda_2 > \lambda_1$. Then,

$$\mathbb{P}(X_2 - X_1 \leq 0) \leq 2\exp\left(-\frac{1}{8}\frac{(\lambda_2-\lambda_1)^2}{\lambda_2+\lambda_1}\right).$$

B. Graph Alignment [18], [13]

We use the results related to the graph-alignment problem introduced in [18], [13]. The $BiG(G(n,p);s)$ edge-sampling model from [18] generates two similar graphs $G_{1,2}$ from a common vertex set. To elaborate on this, let $G = (V,E)$ be a generator graph with vertex set V and edge set E. For a fixed realization of G that is an Erdös-Rényi random graph $G(n,p)$, we generate two graphs $G_{1,2} = (V, E_{1,2})$ by sampling the vertex set E twice. More precisely, each edge $e \in E$ is in the edge set of $E_{1,2}$ with probability s, independently of everything else. In our work we denote this edge-removal operator by ψ

A more realistic $BiG(n,p;t,s)$ model [13] assumes an additional step of node sampling: Firstly nodes of G are sampled independently w.p. t, and secondly the edges among sampled vertices are sampled w.p. s.

VIII. Lemma 3

Proof: We begin by noting that our hypothesis implies $np^2 \to 0$. In this case, $(1-(1-p^2q)^n) \simeq np^2q$ and, therefore, $\mathbb{E}[Q_{u,v}] \simeq np^2q$ for any u, v.

The first two statements are easily derived by using this fact, after applying the linearity of expectation to $|E_q| = \sum_{u,v} Q_{u,v}$ and $d_u = \sum_x Q_{u,x}$, respectively. For the third statement, note that c_u can be written as $c_u = \frac{\sum_{x,y} Q_{x,u}Q_{u,y}Q_{x,y}}{\sum_{x,y} Q_{x,u}Q_{u,y}} = \frac{N}{D}$ Using first order Tailor exapnsion (assured by the concentration of N and D around their means):

$$\mathbb{E}[c_u] \simeq \frac{\mathbb{E}[N]}{\mathbb{E}[D]} + o_p(\frac{\mathbb{E}[N]}{\mathbb{E}[D]}),$$

where o_p means convergence in probability. Below we show in details that the enumerator $\mathbb{E}[N]$ is asymptotically equal to $(npq)^3/2$. Analogously the denominator $\mathbb{E}[D]$ is asymptotically equal to $\mathbb{E}[D] = (n^2p^2q)^2/2$, but we omit lengthy calculations. This implies our result.

It is enough to determine functions $f_1(n,p,q), f_2(n,p,q) \sim (npq)^3/2$ such that$f_1 \leq \mathbb{E}[N] \leq f_2$. An analogous procedure, which we will omit, can be performed for the denominator as well.

Denote $I_{a,b,c} = I_{a,b,c}(x,y) = P_{a,u}P_{a,x}P_{b,u}P_{b,y}P_{c,x}P_{c,y}T_{u,a,x}T_{u,b,y}T_{x,c,y}$, for $a \neq u, x; b \neq u, y; c \neq x, y$. This enables us to write

$$\begin{aligned}\mathbb{E}[N] &= \sum_{(x,y)} \mathbb{P}(Q_{x,u}Q_{u,y}Q_{x,y} = 1) \\ &= \sum_{x,y\neq u} \mathbb{P}\left(\bigoplus_{a,b,c} P_{a,u}P_{a,x}P_{b,u}P_{b,y}P_{c,x}P_{c,y} \right. \\ &\quad \left. \cdot\, T_{u,a,x}T_{u,b,y}T_{x,c,y} = 1\right) \\ &= \sum_{x,y\neq u} \mathbb{P}\left(\bigoplus_{a,b,c} I_{a,b,c} = 1\right)\end{aligned}$$

Note that, if $x, y \neq u$, then $\mathbb{P}(I_{a,b,c} = 1)$ can take three possible values: p^3q^3 if $a = b = c$, p^5q^5 if $a = b \neq c$ (or the two other symmetric cases), and p^6q^6 if $a \neq b \neq c \neq a$.

For the right inequality, note that $\oplus I_{a,b,c} = 1$ iff $\sum I_{a,b,c} \geq 1$. Union bound yields:

$$\begin{aligned}\mathbb{E}[N] &= \sum_{x,y\neq u} \mathbb{P}\left(\sum_{a,b,c} I_{a,b,c} \geq 1\right) \\ &\leq \sum_{x,y\neq u}\sum_{a,b,c} \mathbb{P}(I_{a,b,c} \geq 1) = \sum_{x,y\neq u}\sum_{a,b,c} \mathbb{P}(I_{a,b,c} = 1) \\ &\leq \sum_{x,y\neq u} \Big(\sum_{a=b=c} \mathbb{P}(I_{a,a,a} = 1) + \sum_{\substack{a=b\neq c \\ \text{+symm cases}}} \mathbb{P}(I_{a,a,c} = 1) \\ &\quad + \sum_{a\neq b\neq c\neq a} \mathbb{P}(I_{a,b,c} = 1)\Big) \\ &\leq \binom{n}{2}(np^3q^3 + 3n(n-1)p^5q^3 \\ &\quad + n(n-1)(n-2)p^6q^3) \\ &\simeq \frac{(npq)^3}{2}(1 + np^2 + n^2p^3) \simeq \frac{(npq)^3}{2}\end{aligned}$$

where, in the last equation, we use the fact that $np^2, n^2p^3 \to 0$.

Now, for the left inequality, be dropping some terms from the binary sum, and manipulating a bit further, we have

$$\begin{aligned}\mathbb{E}[N] &= \sum_{x,y\neq u} \mathbb{P}\left(\bigoplus_{a,b,c} I_{a,b,c} = 1\right) \\ &\geq \sum_{x,y\neq u} \mathbb{P}\left(\bigoplus_{\substack{a\neq u,x,y \\ a=b=c}} I_{a,a,a} = 1\right) \\ &= \sum_{x,y\neq u} 1 - \mathbb{P}\left(\bigoplus_{\substack{a\neq u,x,y \\ a=b=c}} I_{a,a,a} = 0\right)\end{aligned}$$

Now, note that, for $a \neq a'$, $I_{a,a,a}$ and $I_{a',a',a'}$ are independent, as they do not share any random variables. Recall that $I_{a,a,a} = \mathsf{Be}(p^3q^3)$. Then, we have:

$$\begin{aligned}\mathbb{E}[N] &\geq \sum_{x,y\neq u} 1 - \mathbb{P}\left(\bigoplus_{\substack{a\neq u,x,y \\ a=b=c}} I_{a,a,a} = 0\right) \\ &= \sum_{x,y\neq u} 1 - (1-p^3q^3)^{n-3} \simeq \frac{(npq)^3}{2}\end{aligned}$$

Thus $\mathbb{E}[N] = \mathbb{E}(\sum_{x,y} Q_{x,u}Q_{u,y}Q_{x,y}) \simeq \frac{(npq)^3}{2}$

See example of the case where $s = a = b = c$ at Fig 2.

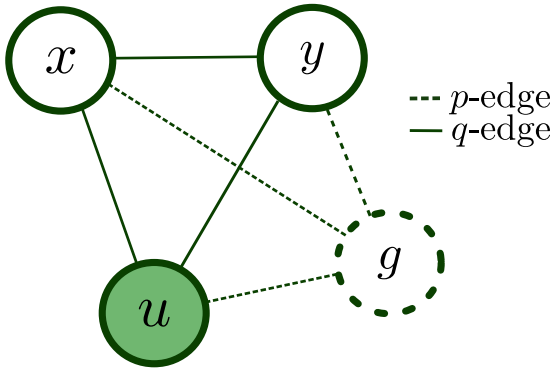

Fig. 2. Edges of neighborhood of u.

■

IX. Proof of Lemma 6

Proof: To prove the first statement, it is enough to show that, under stated assumptions and given u and v are adjacent, the expected size of the set $X = \{x \in V : Q_{u,x}Q_{v,x} = 1, S(u,x) \cap S(v,x) \cap S(u,v) = \emptyset\}$ goes to 0. The first moment method then implies that this set has size 0 a.a.s., which is equivalent to the desired result.

We start by using the tower property:

$$\mathbb{E}[|X| \mid Q_{u,v} = 1]$$
$$= \mathbb{E}[\mathbb{E}[|X| \mid N_u^p, N_v^p, S(u,v), Q_{u,v} = 1] \mid Q_{u,v} = 1]$$

Recall that, for any node g, N_g^p denotes the set of p-neighbors of g that showed up during the construction process of the $G(n,p;q)$ graph. Note that the condition on the inner expectation can be expressed as a function only of random variables of the kinds P_{u*}, P_{v*} and T_{u*v}. By construction, any functions of random variables other than these are independent of this condition.

To bound this inner expectation, fix vertex sets $\mathcal{U}, \mathcal{V}, \mathcal{S}$ with $\mathcal{S} \subseteq \mathcal{U} \cap \mathcal{V}$ and denote current values being fixed event by $\mathcal{T} = \{N_u^p = \mathcal{U}, N_v^p = \mathcal{V}, S(u,v) = \mathcal{S}, Q_{u,v} = 1\}$. Then

$$\mathbb{E}[|X| \,|\, \mathcal{T}] = \sum_{x \notin \{u,v\}} \mathbb{P}(x \in X \,|\, \mathcal{T})$$

Note that $u, v \notin X$, as $S(u,u), S(v,v) = \emptyset$ by construction. For any $x \notin \{u,v\}$, we write

$$\begin{aligned}\mathbb{P}(x \in X \mid \mathcal{T}) &= \mathbb{P}(|S(u,x)| \geq 1, |S(v,x)| \geq 1, \\ &\qquad S(u,x) \cap S(v,x) \cap S(u,v) = \emptyset \mid \mathcal{T}) \\ &= \mathbb{P}(|S(u,x)| \geq 1, |S(v,x)| \geq 1, \\ &\qquad S(u,x) \cap S(v,x) \cap \mathcal{S} = \emptyset \mid \mathcal{T})\end{aligned}$$

Rewriting expressions on $S(u,x)$ and $S(v,x)$ in terms of our basic random variables:

$$\begin{aligned}&\mathbb{P}(x \in X \mid \mathcal{T}) \\ &= \mathbb{P}\Bigg(\bigoplus_{g' \neq u,x} P_{g',u} P_{g',x} T_{u,g',x} = 1, \\ &\qquad \bigoplus_{g'' \neq v,x} P_{g'',v} P_{g'',x} T_{v,g'',x} = 1, \\ &\qquad \bigotimes_{g \in \mathcal{S}} P_{u,g} P_{v,g} P_{x,g} T_{u,g,x} T_{v,g,x} = 0 \;\Bigg|\; \mathcal{T} \Bigg) \\ &= \mathbb{P}\Bigg(\bigoplus_{\substack{g' \neq u,x \\ g'' \neq v,x}} \big(P_{g',u} P_{g',x} T_{u,g',x} P_{g'',v} P_{g'',x} T_{v,g'',x} = 1, \\ &\qquad \bigotimes_{g \in \mathcal{S}} P_{x,g} T_{u,g,x} T_{v,g,x} = 0 \big) \;\Bigg|\; \mathcal{T} \Bigg)\end{aligned}$$

where the last step used the condition that $S(u,v) = \mathcal{S}$ and, therefore, $P_{u,g} P_{v,g} = 1$ for any $g \in \mathcal{S}$. Now, we can apply union bound to the latest expression:

$$\begin{aligned}&\mathbb{P}(x \in X \mid \mathcal{T}) \\ &\leq \sum_{\substack{g' \neq u,x \\ g'' \neq v,x}} \mathbb{P}(P_{g',u} P_{g',x} T_{u,g',x} P_{g'',v} P_{g'',x} T_{v,g'',x} = 1, \\ &\qquad \bigotimes_{g \in \mathcal{S}} P_{x,g} T_{u,g,x} T_{v,g,x} = 0 \;\Bigg|\; \mathcal{T} \Bigg)\end{aligned}$$

The summand has different values depending on g', g''. Let us detail all possible cases:

1) $g' = g'' \in \mathcal{S}$ — since the condition implies $P_{g',u} P_{g',v} T_{u,g',v} = 1$, the event expression reduces to $(P_{g',x} T_{u,g',x} T_{v,g',x} = 1, \otimes_{g \in \mathcal{S}} P_{x,g} T_{u,g,x} T_{v,g,x} = 0)$; the two parts of the expression are mutually exclusive, hence, the event has probability 0;
2) $g' = g'' \in \mathcal{U} \cap \mathcal{V} \setminus \mathcal{S}$ — by the same argument as the previous item, the event expression reduces to $(P_{g',x} T_{u,g',x} T_{v,g',x} = 1, \otimes_{g \in \mathcal{S}} P_{x,g} T_{u,g,x} T_{v,g,x} = 0)$; using independence:

$$\begin{aligned}&\mathbb{P}(P_{g',u} P_{g',x} T_{u,g',x} P_{g'',v} P_{g'',x} T_{v,g'',x} = 1, \\ &\qquad \bigotimes_{g \in \mathcal{S}} P_{x,g} T_{u,g,x} T_{v,g,x} = 0 \;\Bigg|\; \mathcal{T} \Bigg) \\ &= \mathbb{P}(P_{g'x} T_{u,g',x} T_{v,g',x} = 1, \\ &\qquad \bigotimes_{g \in \mathcal{S}} P_{x,g} T_{u,g,x} T_{v,g,x} = 0 \;\Bigg|\; \mathcal{T} \Bigg) \\ &\leq \mathbb{P}(P_{g'x} T_{u,g',x} T_{v,g',x} = 1 \mid \mathcal{T}) \\ &= \mathbb{P}(P_{g'x} T_{u,g',x} T_{v,g',x} = 1) \\ &= pq^2\end{aligned}$$

   This pattern of manipulation also applies to following cases and will be further omitted;
3) $g' = g'' \notin \mathcal{U} \cap \mathcal{V}$ — in this case, either $g' \notin \mathcal{U}$, which implies $P_{g',u} = 0$, or $g'' \notin \mathcal{V}$, which implies $P_{g'',v} = 0$; both facts imply that $P_{g',u} P_{g',x} T_{u,g',x} P_{g'',v} P_{g'',x} T_{v,g'',x} = 0$, so the event has probability 0;
4) $g' \neq g'', g' \in \mathcal{U}, g'' \in \mathcal{V}$ — the expression reduces to $(P_{g',x} T_{u,g',x} P_{g'',x} T_{v,g',x} = 1, \otimes_{g \in \mathcal{S}} P_{x,g} T_{u,g,x} T_{v,g,x} = 0)$, similarly to case 2, and the probability is bounded by $p^2 q^2$;
5) $g' \neq g'', (g' \notin \mathcal{U}$ or $g'' \notin \mathcal{V})$ — as in case 3, the choices of g' and g'' imply that $P_{g',u} P_{g',x} T_{u,g',x} P_{g'',v} P_{g'',x} T_{v,g'',x} = 0$ yielding an event of probability 0;

Case 2 will happen for $|\mathcal{U} \cap \mathcal{V}|$ choices of g' and g'', and case 4 will happen for $|\mathcal{U}| \cdot |\mathcal{V}| - |\mathcal{U} \cap \mathcal{V}|$ such choices. Thus

$$\begin{aligned}&\mathbb{P}(x \in X \mid \mathcal{T}) \\ &\leq |\mathcal{U} \cap \mathcal{V}| pq^2 + (|\mathcal{U}| \cdot |\mathcal{V}| - |\mathcal{U} \cap \mathcal{V}|) p^2 q^2 \\ &= |\mathcal{U} \cap \mathcal{V}| p(1-p) q^2 + |\mathcal{U}| \cdot |\mathcal{V}| p^2 q^2\end{aligned}$$

Since this is valid for any $x \neq u, v$

$$\begin{aligned}&\mathbb{E}[|X| \mid \mathcal{T}, Q_{u,v} = 1] \\ &\qquad \leq \binom{n}{2} (|\mathcal{U} \cap \mathcal{V}| p(1-p) q^2 + |\mathcal{U}| \cdot |\mathcal{V}| p^2 q^2)\end{aligned}$$

and,

$$\begin{aligned}\mathbb{E}[|X| \mid Q_{u,v} = 1] = \mathbb{E}\big[&(n-2)(|N_u^p \cap N_v^p| p(1-p) q^2 \\ &+ |N_u^p| \cdot |N_v^p| p^2 q^2)\big]\end{aligned}$$

By linearity of expectation and independence of N_u^p and N_v^p,

$$\begin{aligned}\mathbb{E}[|X| \mid Q_{u,v}=1] &= (n-2)(\mathbb{E}[|N_u^p \cap N_v^p|]p(1-p)q^2 \\ &\quad + \mathbb{E}[|N_u^p|]\cdot\mathbb{E}[|N_v^p|]p^2q^2) \\ &\simeq n\cdot np^2\cdot pq^2 + (np)^2p^2q^2 \\ &= np^3q^2 + n^2p^4q^2 = o(1)\end{aligned}$$

To show the second and third statements of the lemma, we use the following argument, for any two events A and B, such that $A \subseteq B$ the following holds:

$$\mathbb{P}(A) = \mathbb{P}(B)\frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} = \mathbb{P}(B)\mathbb{P}(A|B).$$

Hence,

$$\begin{aligned}&\mathbb{P}(Q_{x,u}Q_{x,v}=1 \mid Q_{u,v}=1) \\ &= \mathbb{P}\left(x \in \bigcup_{g\in S(u,v)} N_g^p\right) \\ &\quad\cdot \mathbb{P}\left(Q_{x,u}Q_{x,v} \;\middle|\; Q_{u,v}=1 \wedge x \in \bigcup_{g\in S(u,v)} N_g^p\right).\end{aligned}$$

Thus $|N_{u,v}| = \mathsf{Bi}\big(n, |S(u,v)|pq^2\big)$ and $\mathbb{E}[|N_{u,v}| \mid S(u,v)] = n|S(u,v)|pq^2$. ∎

X. Partial Sampling

Let $G(V,E)$ be a realization of an Erdős-Rényi random graph $G(m,p)$, and let $G_1(V_1,E_1)$ and $G_2(V_2,E_2)$ be two samples of G obtained as follows: Each node $u \in V$ is sampled with probability t independently to V_1 and V_2, and E_1 and E_2 are all edges of E whose both endpoints are sampled in V_1 and V_2, respectively.

Lemma 15:

$$\begin{aligned}\mathbb{P}(G_1\sim G_2) &\le \exp\Big(m\log m - cm^2\Big) \\ &\quad + 2\exp -\frac{\delta^2 mt}{2}\end{aligned}$$

where $c(p,t,\delta) = (1-\delta)^2t^2(1-t)\log\big((p^2+(1-p)^2)^{-1}\big)$.

Proof: If $|V_1| \neq |V_2|$, this event has probability 0, so we assume $|V_1| = |V_2| = m'$. Denote by V_0 the set of nodes in G that are sampled in both V_1 and V_2, and let $m_1 = |V_1 \setminus V_0| = |V_2 \setminus V_0|$.

Consider an arbitrary mapping $\pi : V_1 \to V_2$. For any pair of nodes $x \in V_1\setminus V_0, y \in V_1$, if π is an isomorphism, then either $(x,y)\in E_1$ and $(\pi(x),\pi(y))\in E_2$, or $(x,y)\notin E_1$ and $(\pi(x),\pi(y))\notin E_2$. This happens with probability $p^2+(1-p)^2$, since x is not a fixed point of π. In total, we have approx $m'm_1$ such pairs, and the event above happens independently for each pair, hence $\mathbb{P}(G_1 \sim_\pi G_2) \le (p^2+(1-p)^2)^{m'm_1}$.

Denote by $c_1 = (p^2+(1-p)^2)^{-1}$. In total we have at most $m'!$ mappings from G_1 to G_2, thus

$$\begin{aligned}\mathbb{P}(G_1\sim G_2) &\le m'!(c_1)^{-m'm_1} \\ &\le \exp\left(m'\log m' - m'm_1\log c_1\right)\end{aligned}$$

Recall that $m' = \mathsf{Bi}(m,t)$ and $m_1 = \mathsf{Bi}(m,t(1-t))$. Then, $\mathbb{P}(m' \le (1-\delta)mt) \le \exp\Big(-\frac{\delta^2 mt}{2}\Big)$ by Chernoff bound since $m\to\infty$, and similarly $\mathbb{P}(m_1 \le (1-\delta)mt(1-t)) \le \exp\Big(-\frac{\delta^2 mt}{2}\Big)$. Therefore,

$$\begin{aligned}\mathbb{P}(G_1\sim G_2) &\le \exp\left(m\log m - cm^2\right) \\ &\quad + 2\exp\left(-\frac{\delta^2 mt}{2}\right),\end{aligned}$$

where $c = (1-\delta)^2t^2(1-t)\log c_1$. ∎

Consider now the following variation of this graph sampling process. First, graphs G, G_1 and G_2 are generated as previously described. Now, graphs $G_{1,1} = (V_1, E_{1,1})$, $G_{1,2} = (V_1, E_{1,2})$ are obtained by sampling edges from E_1 independently with probability s, this sampling also being independent for $G_{1,1}$ and $G_{1,2}$. Similarly, $G_{2,1} = (V_2, E_{2,1})$ is obtained via this edge-sampling process from G_2. This process is illustrated in Figure 3.

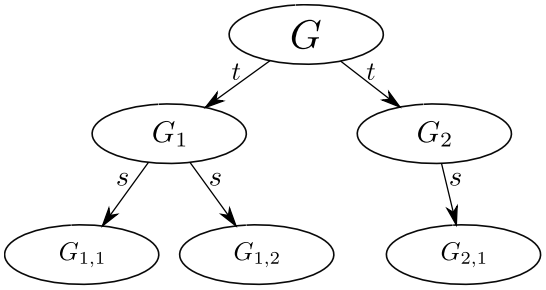

Fig. 3. Node-edge sampling process that generates the edge neighborhoods.

Assume $|V_1| = |V_2| = m$, and denote by π_0 the identity mapping over V_1. Denote by D an event that there exists π such that $\Delta(G_{1,1},G_{1,2},\pi_0) > \Delta(G_{1,1},G_{2,1},\pi)$.

Lemma 16: For $s \gg \left(\frac{\omega(1)\log m}{m}\right)^{\frac{2}{3}}$ and p,t fixed, then

$$\begin{aligned}\mathbb{P}(D) &\le \sum_{k=x+1}^{m} \exp\Big(k\Big(\log m - \frac{mps}{16}\cdot s^2\Big)\Big) \\ &\quad + \exp\left(-\delta^2\frac{m(1-t)}{2}\right).\end{aligned}$$

for $x = \lceil mt(1-t)\rceil$

Proof: Denote by k the number of nodes u such that $\pi(u)\neq u$ and denote by Π_k a subset of all such mappings. Note that always $k \ge |V_2\setminus V_0| = m'$. Then we can write

$$\begin{aligned}\mathbb{P}(D) &\le \mathbb{P}(D|m' \ge mt(1-t)) + \mathbb{P}(m' < mt(1-t)) \\ &\le \sum_{k=x}^{m}\sum_{\pi\in\Pi_k} \mathbb{P}(\Delta(G_{1,1},G_{1,2},\pi_0) > \Delta(G_{1,1},G_{2,1},\pi)) \\ &\quad + \mathbb{P}(m' < x)\end{aligned}$$

First we estimate $\mathbb{P}(\Delta(G_{1,1},G_{1,2},\pi_0) > \Delta(G_{1,1},G_{2,1},\pi))$. We partition V_2 into two sets of nodes $C_\pi \subset V_2$ and $W_\pi \subset V_2$ such that $u \in C_\pi$ iff $\pi^{-1}(u) = u$ and $u \in W_\pi$ otherwise. Also denote by V_0 nodes that are sampled in G_1 and G_2. Note, that $|C_\pi| = m-k$, $|W_\pi| = k$ and $|V_2\setminus V_0| = \mathsf{Bi}(m,1-t)$.

Define mapping $\pi' = \pi \circ g$ where g is a bijection $g : V_2 \to V_1$, which works as follows: If $u \in C_\pi$, then $g(u) = u$; the remaining nodes W_π we map as follows, we arbitrarily split W_π into two equal parts[6] to W_1 and W_2 and we map each $u \in W_1$ s.t. $g(u) = \pi_0(\pi^{-1}(u))$ the rest we map arbitrarily, but not in place. Note that $\pi'|_{W_1 \cup C_\pi} = \pi_0|_{W_1 \cup C_\pi}$

In the following, we show that w.h.p. $\Delta(G_{1,1}, G_{1,2}, \pi') < \Delta(G_{1,1}, G_{2,1}, \pi)$. This follows from Lemma 14. We only need to prove that $\frac{(\lambda_1-\lambda_2)^2}{\lambda_1+\lambda_2} = \omega(1)$. Where $\lambda_1 = \mathbb{E}[\Delta(G_{1,1}, G_{2,1}, \pi)]$ and $\lambda_2 = \mathbb{E}[\Delta(G_{1,1}, G_{1,2}, \pi')]$.

$$\begin{aligned}
\lambda_1 &= \binom{m-k}{2} 2ps(1-s) \\
&\quad + \left((m-k)k + \binom{k}{2}\right) 2ps(1-ps) \\
\lambda_2 &= \binom{m-k+\frac{k}{2}}{2} 2ps(1-s) \\
&\quad + \left(\left(m-k+\frac{k}{2}\right)\frac{k}{2} + \binom{\frac{k}{2}}{2}\right) 2ps(1-ps) \\
\lambda_1 - \lambda_2 &= k\left(m - \frac{3}{4}k\right) ps^2(1-p) \geq k\frac{m}{4} ps^2(1-p) \\
\lambda_1 + \lambda_2 &= \left(2m^2 - 3mk + \frac{5}{4}k^2\right) ps(1-s) \\
&\quad + \left(3mk - \frac{5}{4}k^2\right) ps(1-ps) \\
&\leq 4m^2 ps(2-s-ps)
\end{aligned}$$

Thus, $\frac{(\lambda_1-\lambda_2)^2}{\lambda_1+\lambda_2} \geq \frac{k^2 s^3 p}{64(2-s-ps)}$ which is $\omega(1)$ if for $k > x$. This enables us to bound the first term:

$$\begin{aligned}
&\sum_{k=x+1}^{m} \sum_{\pi \in \Pi_k} \mathbb{P}(\Delta(G_{1,1}, G_{1,2}, \pi_0) > \Delta(G_{1,1}, G_{2,1}, \pi)) \\
&\leq \sum_{k=x+1}^{m} \sum_{\Pi_k} \mathbb{P}(\Delta(G_{1,1}, G_{1,2}, \pi_0) > \Delta(G_{1,1}, G_{1,2}, \pi')) \\
&\leq \sum_{k=x+1}^{m} \exp k\left(\log m - \frac{mps}{16} \cdot s^2\right)
\end{aligned}$$

The last follows from Equation 19 of [18] where conditions of the Theorem 4.1 from [18] are met (except the condition $p \to 0$ that the authors never use).

The second term follows from the fact that

$$\mathbb{P}(m' < x) \leq \mathbb{P}(m' < m(1-t)(1-\delta)) \leq \exp -\frac{m(1-t)\delta^2}{2}$$

due to Chernoff bound. Note that $m' = \mathsf{Bi}(m, 1-t)$.

Then putting all together we get

$$\begin{aligned}
\mathbb{P}(D) \leq &\sum_{k=x+1}^{m} \exp\left(k\left(\log m - \frac{mps}{16} \cdot s^2\right)\right) \\
&+ \exp\left(-\delta^2 \frac{m(1-t)}{2}\right).
\end{aligned}$$

■

[6]Without loss of generality we can assume $|W_\pi|$ is even.