


# Graph Reconstruction from Random Subgraphs

Andrew McGregor 

University of Massachusetts, Amherst, MA, USA

Rik Sengupta 

University of Massachusetts, Amherst, MA, USA

---

## Abstract

---

We consider the problem of reconstructing a graph  $G$  in two natural sampling models: 1) each sample corresponds to a random induced subgraph and 2) for a fixed adjacency matrix  $A_G$  for  $G$ , each sample corresponds to a random principal submatrix (i.e., a submatrix formed by deleting the same set of rows and columns) of  $A_G$ . We refer to these models as the “unordered” and “ordered” models respectively. The two models are motivated by work on the *reconstruction conjecture* in combinatorics and *trace reconstruction* in theoretical computer science. Despite the superficial similarities between the models, we show that the sample complexity of reconstruction can be exponentially different. Our main results are as follows:

- In the unordered model, we show that almost all graphs can be reconstructed with  $\Theta(p^{-2} \log n)$  samples if each node is included in the random subgraph with *any* constant probability  $p$ ; this is optimal. We show our upper bound extends to smaller values of  $p$  as well. In contrast, for arbitrary graphs, we show that  $\exp(\Omega(n))$  samples are required for reconstruction even for 2-regular graphs. One of the key technical steps in the first result is showing that, with high probability, any subgraph isomorphism in a random graph has at most  $O(\log n)$  non-fixed points.
- In the ordered model, we show that any graph with constant arboricity or degeneracy (i.e., every induced subgraph has constant average degree) can be reconstructed with  $\exp(\tilde{O}(n^{1/3}))$  samples and that arbitrary graphs can be reconstructed with  $\exp(\tilde{O}(n^{1/2}))$  samples. The results about almost all graphs in the first model carry over to the second. One of the key technical steps in the first result is showing that reconstruction of low degeneracy graphs can be reduced to learning a small number of moments of sets of the form  $\{i - j : j < i, (i, j) \in E\}$  and  $\{j - i : i < j, (i, j) \in E\}$  where  $G = ([n], E)$  is the unknown graph.

**2012 ACM Subject Classification** Mathematics of computing → Probability and statistics

**Keywords and phrases** graph reconstruction, sample complexity, deletion channel

**Digital Object Identifier** 10.4230/LIPIcs.ICALP.2022.96

**Category** Track A: Algorithms, Complexity and Games

**Funding** *Andrew McGregor*: Work supported in part by NSF CCF-1934846, CCF-1908849, and CCF-1637536.

**Acknowledgements** We thank the reviewers for numerous helpful suggestions.

## 1 Introduction

We consider the problem of reconstructing an undirected graph  $G$  on  $n$  nodes in the following two natural sampling models:

- **Unordered Model:** Each node is sampled independently with probability  $p$ . The returned sample is the induced subgraph  $G[A]$  where  $A$  is the set of sampled nodes. We wish to reconstruct  $G$  up to isomorphism.
- **Ordered Model:** Let  $A_G$  be a fixed adjacency matrix for  $G$ . Each node is sampled independently with probability  $p$ . For each node not sampled, the corresponding row and column of  $A_G$  are deleted and the returned sample corresponds to the resulting submatrix. We wish to reconstruct  $A_G$ .



© Andrew McGregor and Rik Sengupta;

licensed under Creative Commons License CC-BY 4.0

49th International Colloquium on Automata, Languages, and Programming (ICALP 2022).

Editors: Mikołaj Bojańczyk, Emanuela Merelli, and David P. Woodruff;

Article No. 96; pp. 96:1–96:18



Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



We are interested in the sample complexity of reconstruction in each model, i.e., the number of samples required to reconstruct the original graph with high probability. When  $p = 1/2$ , the problem in the unordered model is how many uniformly random induced subgraphs are required to reconstruct a graph. The problem in the ordered model is how many random principal submatrices (i.e., submatrices formed by symmetric row/column deletions) are required to reconstruct a symmetric binary matrix with 0s on the diagonal. We will be interested in reconstructing both arbitrary graphs and random graphs (i.e., *almost all* graphs). In the ordered model, we are also interested in reconstructing low-degeneracy graphs. Degeneracy is perhaps the most natural notion of sparsity in graphs and sparsity has played an important role in other reconstruction problems such as compressed sensing [12] and trace reconstruction [20].

**Unordered Model and Reconstruction from  $k$ -Decks.** Reconstruction in the unordered model is closely connected to the problem of reconstruction from  $k$ -decks. Given an undirected graph  $G$ , its  $k$ -deck is the multiset of all  $\binom{n}{k}$  induced subgraphs on  $k$  of the vertices of  $G$ . The  $(n-1)$ -deck is typically referred to as just the *deck*. The Reconstruction Conjecture, due to Kelly [17] and Ulam [31], asks whether there exist two different graphs with at least three nodes, that have the same deck. Bollobas [3] proved that almost all graphs can be reconstructed by taking three graphs from its deck. Furthermore, almost all sets of three graphs from the deck of  $G$  suffice to reconstruct it. Recently, Spinoza and West [30] generalized this result to the following. Let  $\varepsilon > 0$  be an arbitrarily small constant and let  $\ell \leq (1 - \varepsilon)n/2$ . Then, almost all graphs can be reconstructed from some subset of  $\binom{\ell+2}{2}$  induced subgraphs from the  $(n - \ell)$ -deck. Note that if a  $k$ -deck is sufficient to reconstruct a graph, then one approach to bounding the sample complexity in our problem is to analyze the sample complexity of reconstructing the  $k$ -deck; this could be done by repeatedly sampling subgraphs of the appropriate size and estimating the number of copies of each such subgraph in the graph. However, the results above suggest it might be possible to reconstruct random graphs more efficiently. Many of the above results rely on showing that for almost all graphs  $G$ , any two “large” subgraphs of  $G$  are not isomorphic. In our problem we need to consider subgraphs of  $G$  that are significantly smaller and have to bound the number of isomorphisms rather than ensuring there are none.

**Ordered Model and Trace Reconstruction.** There is also a natural variant of the  $k$ -deck problem for matrices where now the  $k$ -deck corresponds to the multiset of all submatrices or principal submatrices. For example, reconstruction from such  $k$ -decks was studied by Kós et al. [18] and they showed that the  $O(n^{2/3})$ -deck was sufficient for reconstruction. Our problem in the ordered model can be thought of a stochastic variant of this problem.

Our problem is also closely related to the *trace reconstruction* problem. In this problem the goal is to reconstruct an unknown binary string  $x \in \{0, 1\}^n$  from independent random subsequences, or “traces”, where each subsequence is formed by deleting each bit with probability  $q = 1 - p$  and then concatenating the remaining bits. The trace reconstruction problem was first proposed by Batu et al. [2]. Since then, the problem attracted a lot of attention and ended up branching out into several directions and variants [4, 10, 10, 11, 13–16, 20–22, 26, 27, 29, 32]. The best upper and lower bounds known for this problem are  $\exp(\tilde{O}(n^{1/5}))$  and  $\tilde{\Omega}(n^{2/3})$  traces respectively, and were both proved recently by Chase [6, 7]. Our approach for reconstructing arbitrary graphs is very similar to the approach in [20, 27], but the result on reconstructing low-degeneracy graphs requires combining those ideas with a “peeling” approach that iteratively reconstructs the neighborhood of low degree nodes, removes these nodes, and recurses on the remaining graph.

There is a natural variant of the classical string trace reconstruction problem where we have an unknown  $n \times n$  binary matrix, and each trace is a sub-matrix obtained by deleting each row and column independently with probability  $q$ . The best known upper bound for this variant is  $\exp(\tilde{O}(n^{1/2}))$  traces [20]. The matrix variant is different from the string version because there are now dependencies between the bits that are deleted. The matrix variant is very closely related to the ordered model we consider; the only slight difference is that in the ordered model there is the symmetric constraint when deleting rows and columns of the adjacency matrix. There has also recently been work on tree reconstruction [4, 10, 21] that, although related somewhat to our work, primarily deals with different models of deletion channels that are only defined for rooted trees. There have been recent advances in a “smoothed” variant of the problem where each bit of the string is replaced by a uniform random bit with some probability [8]. Another variant, *coded* trace reconstruction, involving efficiently encodable codes that can be recovered despite some constant probability of edit errors, has also been studied extensively [5, 9, 10]. A generalized formulation of the problem where instead of a single unknown string, we draw a string at random from some distribution over strings and pass it through a deletion channel, has also garnered some recent interest [1, 25]. Note that in both the ordered and unordered model, the main challenge to reconstruction is that the nodes are not labelled. However, we note that there are also other interesting reconstruction problems arising in the context of labelled graph, see e.g., Mossel and Ross [23], but these consider very different models from those considered here.

## 1.1 Our Results

1. **Unordered Model:** We show that for almost all graphs,  $\Theta(p^{-2} \log n \log(1/\delta))$  traces (where in this model a trace is a randomly induced subgraph) suffice for reconstruction with probability at least  $1 - \delta$ , as long as the retention probability  $p$  is  $\tilde{\Omega}(1/n^{1/6})$ . Note that this is optimal for the range of  $p$  considered since  $\Theta(p^{-2} \log n)$  traces are required to ensure every edge appears in at least one trace.<sup>1</sup> In contrast, we show that reconstructing arbitrary graphs is hard: even distinguishing between a pair of 2-regular graphs may require  $\exp(\Omega(n))$  traces. We show, however, that if the maximum degree of  $G$  is at most one, then it can be reconstructed with  $\Theta(n)$  traces. One of the key technical steps in the first result is showing that, with high probability, any subgraph isomorphism in a random graph has at most  $O(\log n)$  non-fixed points. This contrasts with a classic result by Müller [31] that shows that there are isomorphic subgraphs (where the isomorphism may contain an unbounded number of non-fixed points) of size  $n/2$  but no isomorphic subgraphs of size  $n(1 + \varepsilon)/2$  for any constant  $\varepsilon > 0$ .
2. **Ordered Model:** Our main result in the ordered model is that  $\exp(\tilde{O}(n^{1/3}))$  samples (i.e., a random principal submatrix of the adjacency matrix) suffice to reconstruct graphs of constant degeneracy, as long as the retention probability  $p$  is a constant. Recall that the *degeneracy* of a graph is the smallest  $k \in \mathbb{N}$  such that every induced subgraph has a vertex of degree at most  $k$ .<sup>2</sup> One of the key technical steps in the first result is showing

<sup>1</sup> This follows by considering a graph that consists of  $n/2$  vertex-disjoint edges. The edges of such a graph appear in a trace independently of one another. Hence, the probability that every edge appears in at least one of  $t$  traces is  $(1 - (1 - p^2)^t)^{n/2}$  and this is at most  $(1 - 2/n)^{n/2} \leq 1/e$  if  $t < \log_{1-p^2}(2/n) = \Omega(\log(n)/p^2)$ .

<sup>2</sup> Note that the degeneracy of a graph is constant factor related to the arboricity of the graph. It is a robust notion of the sparsity of a graph in that it ensures that the induced subgraph on any  $r$  nodes

that reconstruction of low degeneracy graphs can be reduced to learning a small number of moments of sets of the form  $\{i - j : j < i, (i, j) \in E\}$  and  $\{j - i : i < j, (i, j) \in E\}$ . These moments can then be learned via an extension of methods from complex analysis that have been developed for the trace reconstruction problem. Our results represents a strong separation between sample complexity of reconstruction in the ordered and unordered models since 2-regular graphs are a special case of low degeneracy graphs. Finally, we show that any graph can be reconstructed with  $\exp(\tilde{O}(n^{1/2}))$  traces. The upper bound is established via a slight modification of a result by Krishnamurthy et al. [20]; they considered independent row/column deletions and the modification is required to deal with the fact that in our setting a row is deleted iff the corresponding column is deleted.

## 2 Reconstruction of Almost All Graphs in the Unordered Model

For reconstructing almost all graphs in the unordered model, the high level approach is:

1. Determine a *consistent* labeling of all nodes in the traces such that two nodes receive the same label iff they correspond to the same node in the unknown graph  $G$ . Determining this labelling will be the main technical challenge in our approach and it is especially challenging in the unordered model (compared to the ordered model) because there is no apparent ordering of the nodes that are observed in a trace.
2. If each pair of nodes of  $G$  appears together in some trace, we know whether or not there exists an edge between these two nodes. Note that

$$T := 3p^{-2} \log n$$

traces are sufficient to ensure this second condition with high probability.<sup>3</sup>

The natural question is how to determine a consistent labeling. We do this by considering all pairs of traces and for each node in the first trace of the pair we determine which node, if any, it corresponds to in the second trace. If we do this for all pairs of traces and every node of  $G$  appears in at least one trace, then for each node of  $G$  we identify all of its occurrences amongst the traces. Hence, the problem of finding a consistent labeling reduces to the problem of finding corresponding nodes in two traces.

How do we find and label corresponding nodes between two traces? Suppose one trace is the induced subgraph on a set of nodes  $A$  and the second trace is the induced subgraph on a set of nodes  $B$ . To find the corresponding nodes, a possible approach is to find the largest subgraph in  $G[A]$  that is isomorphic to a subgraph in  $G[B]$ . If this subgraph were  $G[A \cap B]$  and  $G[A \cap B]$  were asymmetric, i.e., had no non-trivial automorphisms, then this would allow us to identify corresponding nodes. If  $G$  is random, there is indeed reason to hope that  $G[A \cap B]$  is the largest common subgraph. In fact, if the probability  $p$  used in the generation of the traces is strictly greater than  $1/\sqrt{2}$  we can show that this approach works exactly as stated, via a classic graph theory result by Müller [24]. His result establishes that for any constant  $\varepsilon > 0$ , for almost every graph  $G$ , the induced subgraphs with at least  $(1 + \varepsilon)n/2$  vertices have no nontrivial automorphisms and are pairwise non-isomorphic. We omit the details as we will instead prove a more general result that applies even when  $p < 1/\sqrt{2}$ . To

---

has  $O(r)$  edges.

<sup>3</sup> Throughout this paper, we will mean a “high probability” bound to be one that holds with probability  $1 - 1/\text{poly}(n)$ . In this case, the high probability bound follows because the probability that there exists a pairs of nodes that does not appear up in the same trace is at most  $\binom{n}{2}(1 - p^2)^T \leq n^2 e^{-p^2 T}$ .

prove the more general result, in the next section we will define a family of graphs called *distinctive graphs* and prove that a random graph is distinctive with high probability. In the following section, we show that it is possible to find corresponding nodes between two traces with high probability if the graph  $G$  is distinctive. This will require more than just finding the largest pair of isomorphic subgraphs, as described below.

## 2.1 Active Isomorphisms and Distinctive Graphs

Strengthening Müller's result to apply to subgraphs of size less than  $n/2$  would give a possible approach to finding corresponding nodes when  $p < 1/\sqrt{2}$ . Unfortunately this is not possible: a random graph contains two isomorphic subgraphs on  $n/2$  nodes with probability at least  $1/2$ . For example, for any pair of nodes  $u$  and  $v$ ,  $G[C \cup \{u\}] \cong G[C \cup \{v\}]$  where  $C$  consists of nodes that are neighbors of both  $u$  and  $v$  or neighbors of neither, and the expected size of  $C \cup \{u\}$  is  $n/2$ . However, note that the isomorphism in this example will consist almost entirely of fixed points. And indeed, this turns out to be a constraint that we can leverage to our benefit: if we disallow isomorphisms with many fixed points, we can break the  $n/2$  barrier in Müller's result.

► **Definition 1 (Active Isomorphisms).** *Given a graph  $G$  and two subsets  $A, B$  of vertices, say  $A$  and  $B$  have an active isomorphism if there exists an isomorphism between  $G[A]$  and  $G[B]$  with no fixed points. Say  $G$  has an active subgraph isomorphism of size  $M$  if there exist vertex subsets  $A$  and  $B$  with  $|A| = |B| = M$  with an active isomorphism.*

We next show that a random graph drawn from  $\mathcal{G}(n, 1/2)$  is unlikely to have large active subgraph isomorphisms.

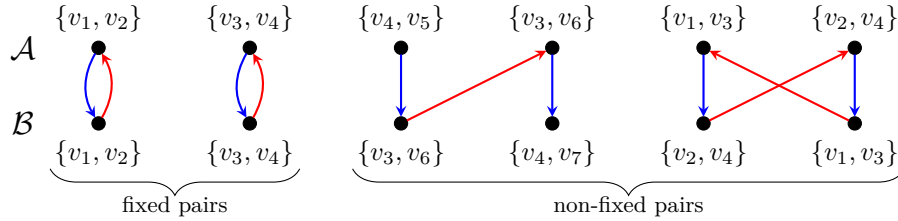
► **Theorem 2.** *For all sufficiently large  $n$ , with high probability, a random graph on  $n$  vertices has no active subgraph isomorphisms of size  $M = 20 \log n$ .*

**Proof.** Let  $G$  be a random graph drawn from  $\mathcal{G}(n, 1/2)$ , and let  $A, B \subseteq V(G)$  be subsets of size  $M$  and let  $\varphi : A \rightarrow B$  be a bijection that has no fixed points. Consider the pairs  $\mathcal{A} := \{\{a_1, a_2\} : a_1, a_2 \in A, a_1 \neq a_2\}$  and  $\mathcal{B} := \{\{b_1, b_2\} : b_1, b_2 \in B, b_1 \neq b_2\}$ . Note that  $\varphi$  naturally induces a bijection  $\varphi'$  from  $\mathcal{A}$  to  $\mathcal{B}$ . Each vertex pair  $\mathcal{A}$  is either an edge or a non-edge; we think of these as the two *types* of pairs. The map  $\varphi$  is an isomorphism precisely when  $\varphi'$  preserves types. Note that while  $\varphi$  has no fixed points,  $\varphi'$  may have fixed points. In particular,  $\{u, v\}$  is a fixed point of  $\varphi'$  iff  $\varphi(u) = v$  and  $\varphi(v) = u$ . Hence, if  $t$  is the number of non-fixed points of  $\varphi'$ , it follows that  $t \geq \binom{M}{2} - M/2$ . Define a directed graph  $\mathcal{G}$  on the vertex set  $\mathcal{A} \sqcup \mathcal{B}$  with:

- An arc from  $\{u, u'\} \in \mathcal{A}$  to  $\{v, v'\} \in \mathcal{B}$  if  $\varphi'(\{u, u'\}) = \{v, v'\}$ .
- An arc from  $\{v, v'\} \in \mathcal{B}$  to  $\{w, w'\} \in \mathcal{A}$  if  $\{v, v'\} = \{w, w'\}$ .

See Figure 1 for an example. Since each node has out-degree and in-degree at most 1 (where every node in  $\mathcal{A}$  has out-degree exactly 1),  $\mathcal{G}$  is a disjoint union of even cycles and odd paths that start in  $\mathcal{A}$  and end in  $\mathcal{B}$ . For  $\varphi$  to be an isomorphism, all nodes in the same connected component of  $\mathcal{G}$  must have the same type. If the component is a path with  $k$  nodes in  $\mathcal{B}$ , the probability all nodes of the component have the same type is  $1/2^k$ . If the component is a cycle with  $k$  nodes in  $\mathcal{B}$ , the probability is  $1/2^{k-1}$ . Hence, the probability of  $\varphi$  being an isomorphism is  $1/2^{|\mathcal{B}| - c}$  where  $c$  is the number of cycles. Note that there are at most  $|\mathcal{B}| - t$  cycles with exactly one node in  $\mathcal{B}$  and the rest have at least two nodes in  $\mathcal{B}$ . Hence,  $(|\mathcal{B}| - t) + 2(c - |\mathcal{B}| + t) \leq |\mathcal{B}|$  and this implies  $|\mathcal{B}| - c \geq t/2$ . Hence, the probability  $\varphi$  is an isomorphism is at most  $1/2^{t/2}$ .

The probability of  $A$  and  $B$  having an active isomorphism is bounded above by the total number of bijections from  $A$  to  $B$  with all points non-fixed, times the probability of such a bijection being an isomorphism, which is bounded above by  $M! \cdot 2^{-t/2} \leq 2^{M \log M - t/2} \leq 2^{M \log M - \frac{1}{2} \binom{M}{2} + M/4}$ . Taking the union bound over all subgraphs of size  $M$  still gives  $\binom{n}{M}^2 \cdot 2^{M \log M - \frac{1}{2} \binom{M}{2} + M/4} \leq 2^{3M \log n + M/4 - \frac{1}{2} \binom{M}{2}} = 2^{-40 \log^2 n + 10 \log n} \leq 2^{-30 \log^2 n} = n^{-\Omega(\log n)}$ . ◀



■ **Figure 1** A construction from the proof of Theorem 2. Here, the blue arcs represent the isomorphism  $\varphi'$  between pairs in  $\mathcal{A}$  and pairs in  $\mathcal{B}$ , while red arcs represent the *same* pair of vertices.

► **Corollary 3** (Extension to Müller). *For all sufficiently large  $n$ , with high probability, there are no two isomorphic subgraphs of a random  $n$ -vertex graph  $G$  for which the isomorphism has  $20 \log n$  or more non-fixed points.*

**Proof.** Theorem 2 immediately implies there does not exist a subgraph isomorphism with more than  $M$  non-fixed points because a subgraph isomorphism with  $M$  non-fixed points implies the existence of a size  $M$  active subgraph isomorphism. ◀

Suppose a graph  $G$  has a subgraph  $H$  that has no nontrivial automorphisms. We can now fix a *canonical* ordering  $\tau$  of  $V(H)$ , and define the *signature* of a vertex  $v \in V(G)$  with respect to  $H$  as the length- $|V(H)|$  binary vector whose  $i$ th entry is 1 if and only if  $v$  is adjacent to the  $i$ th vertex of  $H$  in the ordering  $\tau$ . For a fixed asymmetric subgraph  $H$ , we say vertices  $u, v \in V(G)$  are *distinguishable with respect to  $H$*  if and only if they have distinct signatures with respect to  $H$ .

► **Definition 4** (Distinctive Graphs). *We say that a graph  $G$  on  $n$  vertices is distinctive if the following conditions are met:*

1. *Any subgraph of  $G$  on  $200 \log n$  or more vertices is asymmetric.*
2. *For any two subsets  $A, B \subseteq V(G)$  satisfying  $G[A] \cong G[B]$ , there are at most  $200 \log n$  non-fixed points in the isomorphism between them.*
3. *For all but a  $1/n$  fraction of subgraphs  $H$  of  $G$  with  $|V(H)| \geq 200 \log n$ , the vertices in  $V(G)$  are pairwise distinguishable with respect to  $H$ .*

Note that the isomorphism in the second condition in the definition is well-defined and unique, because  $G[A]$  and  $G[B]$  of size more than  $200 \log n$  are asymmetric by the first condition; similarly, the third condition is well-defined, because  $H$  is asymmetric, also by the first condition.

► **Theorem 5.** *Almost all graphs are distinctive.*

**Proof.** Consider a random graph  $G$ . The probability that any particular  $k$ -node random subgraph has a non-identity isomorphism is at most  $2^{k \log k - k^2/100}$  (see e.g. Theorem 3.1 in [28]). Taking the union bound over all  $\binom{n}{k}$  subgraphs of  $G$  and all possible sizes of the subgraph from  $k$  to  $n$ , the probability of  $G$  having a non-asymmetric subgraph on  $k$  or more vertices is at most  $2^{\log n(1+2k)+k-k^2/100}$ , which is  $1/\text{poly}(n)$  for  $k \geq 200 \log n$ . The second follows directly from Corollary 3 above. The third is due to the following counting argument. Fix a random subgraph  $H$  of  $G$  of size at least  $200 \log n$ . For a fixed  $v \in V(G)$ , the probability of there being an edge to any given vertex of  $H$  is  $1/2$ , and these are independent for different vertices  $v$ . Therefore, the signature of a vertex  $v$  with respect to a fixed  $H$  corresponds to a uniformly random binary string if  $v \notin H$ . If  $v \in H$ , the string is uniformly random aside from the index corresponding to  $v$ , which is deterministically 0. The probability that two of these signatures match is at most  $2^{-|V(H)|+1} \leq 2^{-200 \log n+1} \leq n^{-199}$ . Therefore, by the union bound, the probability that two vertices in  $G$  have the same signature with respect to  $H$  is at most  $\binom{n}{2}/n^{199} \leq 1/n^{197}$ . Therefore, the expected fraction of random subgraphs  $H$  of size at least  $200 \log n$  with the property that there are two distinct vertices that are indistinguishable with respect to  $H$  is at most  $1/n^{197}$ . The probability this fraction exceeds  $1/n$  is at most  $1/n^{196}$  by an application of the Markov bound.  $\blacktriangleleft$

## 2.2 Reconstruction of Distinctive Graphs

Let  $G$  be a distinctive graph on  $n$  vertices. In this section, we will show an upper bound on the sample complexity of reconstructing  $G$  from uniformly random induced subgraphs, obtained by retaining each vertex independently with probability  $p \geq 12n^{-1/6} \log^{2/3} n$ . Recall that if we take  $T := 3p^{-2} \log n$  samples uniformly at random, then we see each pair of vertices appear together in *some* trace.

Consider first just two such random induced subgraphs  $G[A]$  and  $G[B]$  where  $A$  and  $B$  are subsets of the nodes formed independently by sampling each node in  $G$  with probability  $p$ . Let  $H$  be the largest graph that appears as an induced subgraph of both  $G[A]$  and  $G[B]$ . Let  $A' \subseteq A$  and  $B' \subseteq B$  be the nodes in  $A$  and  $B$  that induce  $H$ . Note that it could be that  $A' = B' = A \cap B$ , but while  $G[A \cap B]$  is a subgraph of both  $G[A]$  and  $G[B]$ , we do not know if it is the largest subgraph that appears in both. However, since  $|A'| = |B'| \geq |A \cap B|$  and  $\mathbb{E}[|A \cap B|] = p^2 n \gg 200 \log n$  by our choice of  $p$  we know  $|A'| = |B'| \geq 200 \log n$  with high probability. Then, by Distinctive Property 1, we can assume that  $G[A']$  and  $G[B']$  are asymmetric and that there is a unique isomorphism  $\varphi$  between the copy of  $H$  in  $G[A]$  and the copy of  $H$  in  $G[B]$ .

Let us now subsample the vertices in  $G[A']$  with probability

$$\alpha := \frac{1}{1200T^2 \log n},$$

and call the resulting set of vertices  $C' \subset A'$ . Observe that the number of vertices not fixed by  $\varphi$  is at most  $200 \log n$  by Distinctive Property 2. So by Markov's inequality, the probability that we subsample such a non-fixed point is at most  $200\alpha \log n$ . So with probability at least  $1 - 200\alpha \log n$ , the set  $C'$  consists entirely of fixed points in  $\varphi$  and is therefore entirely in the intersection  $A \cap B$ .

Let  $\mathcal{D}_{C'}$  be the distribution of  $C'$ , i.e., we sample  $A$ , find  $H$ , and then subsample the nodes of  $H$ . For the sake of analysis, suppose we can identify the nodes in  $A \cap B$  and let  $C$  be the set formed by sampling from  $A \cap B$  with probability  $\alpha$ . Let  $\mathcal{D}_C$  be the distribution of  $C$ . Both the definition of  $\mathcal{D}_{C'}$  and  $\mathcal{D}_C$  is with respect to some fixed  $A$  and  $B$ .

► **Theorem 6.** *The variational distance between  $\mathcal{D}_C$  and  $\mathcal{D}_{C'}$  is at most  $200\alpha \log n$ .*

**Proof.** Let  $x, y \in \{0, 1\}^{|(A \cap B) \cup A'|}$  be the characteristic vectors of  $C$  and  $C'$  respectively, where we have padded each of their domains up with zeros if necessary, for notational convenience. Let  $\gamma_i(b) := \mathbb{P}(x_i = b)$  and  $\beta_i(b) := \mathbb{P}(y_i = b)$  where  $b \in \{0, 1\}$ . Define  $S_1 := A \cap B$ , and  $S_2 := A'$ . Then,  $\ell_1(\mathcal{D}_C, \mathcal{D}_{C'})$  is

$$\begin{aligned} & \sum_{z \in \{0, 1\}^{|(A \cap B) \cup A'|}} \left| \prod_i \gamma_i(z_i) - \prod_i \beta_i(z_i) \right| \\ & \leq \sum_i \sum_{b \in \{0, 1\}} \left| \gamma_i(b) - \beta_i(b) \right| \\ & = \sum_{i \in S_1 \cap S_2} \sum_{b \in \{0, 1\}} \left| \gamma_i(b) - \beta_i(b) \right| + \sum_{i \in S_1 \setminus S_2} \sum_{b \in \{0, 1\}} \left| \gamma_i(b) - \beta_i(b) \right| \\ & \quad + \sum_{i \in S_2 \setminus S_1} \sum_{b \in \{0, 1\}} \left| \gamma_i(b) - \beta_i(b) \right|. \end{aligned}$$

where the first inequality follows from the fact the  $\ell_1$ -distance between two product distributions is at most the sum of the  $\ell_1$  distance between the marginals.<sup>4</sup>

Of these, the first term vanishes, as the inner difference is zero, whereas the two other terms are bounded by  $2\alpha$  for each term, giving us  $\ell_1(X, Y) \leq 2\alpha(|S_1 \setminus S_2| + |S_2 \setminus S_1|) \leq 2\alpha \cdot 200 \log n$ . Since the variational distance is half the  $\ell_1$ -distance, the stated result follows. ◀

► **Corollary 7.** *If  $p \geq 12n^{-1/6} \log^{2/3} n$ , then with probability at least  $1 - 400\alpha \log n$  (where the probability is taken over the choice of  $A, B$  and the subsampling of  $A$ ),  $C'$  satisfies the condition in distinctive property 3.*

**Proof.** Let  $S$  be the event that the graph we draw satisfies Distinctive Property 3. Then, by Theorem 6, we know  $\mathbb{P}_{\mathcal{D}_C}(S) - \mathbb{P}_{\mathcal{D}_{C'}}(S) \leq \|\mathcal{D}_C - \mathcal{D}_{C'}\|_{TV} \leq 200\alpha \log n$ . Recall that  $T = 3p^{-2} \log n$  and  $\alpha = 1/(1200T^2 \log n)$ . Therefore,

$$\mathbb{E}[|C|] = p^2 n \alpha = \frac{p^2 n}{1200T^2 \log n} = \frac{p^2 n \cdot p^4}{1200 \cdot 9 \log^2 n \cdot \log n} > 250 \log n$$

for  $p \geq 12n^{-1/6} \log^{2/3} n$ . By an application of the Chernoff bound,  $|C| \geq 200 \log n$  with high probability. Because  $G$  is distinctive, this implies that  $C$  satisfies the condition in Distinctive Property 3. Note that for  $p = \tilde{\Omega}(n^{-1/6})$ , we have  $200\alpha \log n = \tilde{\Omega}(n^{-2/3}) \gg 1/n$ , and so it follows that  $\mathbb{P}_{\mathcal{D}_{C'}}(S)$  is at least  $1 - 1/n - 200\alpha \log n \gg 1 - 400\alpha \log n$ . So, with the stated probability,  $C'$  satisfies the property of condition 3. ◀

This gives rise to our main result, the following algorithm for reconstructing  $G$ .

► **Theorem 8.** *Let  $G$  be a distinctive graph on  $n$  vertices and  $\delta > 0$ . We can reconstruct  $G$  with probability at least  $1 - \delta$  from  $\Theta(p^{-2} \cdot \log n \cdot \log(1/\delta))$  traces, when the retention probability  $p$  satisfies  $p = \tilde{\Omega}(n^{-1/6})$ .*

<sup>4</sup> This can be verified via induction of the number of marginals since by the triangle inequality:

$$\left| \prod_{i \geq 1} \gamma_i(z_i) - \prod_{i \geq 1} \beta_i(z_i) \right| \leq \left| \prod_{i \geq 1} \gamma_i(z_i) - \alpha_1(z_1) \prod_{i \geq 2} \beta_i(z_i) \right| + \left| \alpha_1(z_1) \prod_{i \geq 2} \beta_i(z_i) - \prod_{i \geq 1} \beta_i(z_i) \right|$$

where the second term when summed over  $z$  equals the  $\ell_1$  distance between the first marginal and we apply the induction hypothesis to the first term.



**Proof.** The proof relies on the following observation. Suppose that for *any* two traces  $G_1$  and  $G_2$ , and any vertices  $x \in G_1$  and  $y \in G_2$ , we can identify whether or not  $x$  and  $y$  are the “same” in the sense of corresponding to the same original vertex in  $G$ . Now, if we have enough traces so that each vertex in the original graph  $G$  appears in at least one of them, then we can *consistently* cluster all the vertices in all the traces into  $n$  clusters, with the vertices in each cluster corresponding to the same vertex in  $G$ . These clusters give rise to a “labeling” of the nodes in the traces with labels 1 through  $n$ . If now, in addition, each *pair* of vertices in  $G$  appear together in some trace, we have a way of identifying whether there is an edge between the pair of nodes in  $G$ , and therefore we can recover the isomorphism class of  $G$ , which is equivalent to reconstructing  $G$  in the unordered setting. This corresponds to the approach highlighted in the high level description at the start of this section.

Formally, we have the following algorithm for reconstructing  $G$  given  $T = 3p^{-2} \log n$  random induced subgraphs of  $G$ . For any two of these subgraphs, we can find the largest common subgraph to both of them, and then subsample from it with probability  $\alpha$  as defined above. By Corollary 7, this subsampled subgraph satisfies the third distinctive property with high probability, and so we can use it to obtain the signatures of all vertices in  $A \cup B$ , enabling us to label the two sampled subgraphs consistently with respect to each other. With high probability, every pair of vertices will appear together in one of the sampled subgraphs. Therefore, as long as we have a consistent labeling of all vertices in these subgraphs with respect to each other, we will have a consistent labeling of the entire graph  $G$ , and therefore be able to reconstruct it.

The probability of this happening, by union bounding over the at most  $T^2$  pairs of random subgraphs we generated, is at least  $1 - 400T^2\alpha \log n = 1 - 400T^2 \log n \cdot \frac{1}{1200T^2 \log n} = 1 - \frac{1}{3} = 2/3$ . We can now reduce the failure probability to  $\delta$  by repeating the process  $O(\log 1/\delta)$  times and taking the most commonly reconstructed graph. This requires an additional factor of  $O(\log 1/\delta)$  in the number of traces. ◀

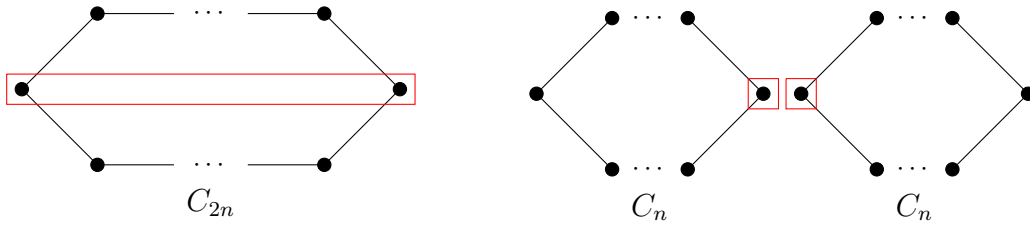
► **Corollary 9.** *Let  $n$  be a sufficiently large integer and  $p = \tilde{\Omega}(n^{-1/6})$ . For almost all  $n$ -node graphs,  $\Theta(p^{-2} \cdot \log n \cdot \log(1/\delta))$  traces suffice for reconstruction with probability at least  $1 - \delta$ .*

### 3 Reconstruction of Arbitrary Graphs in the Unordered Model

We next consider reconstructing arbitrary graphs and show that even distinguishing two fixed graphs may require  $2^{\Omega(n)}$  random induced subgraphs, highlighting the vast gap between random graphs and arbitrary graphs. In fact the lower bound even applies to distinguishing between two graphs with maximum degree 2. This immediately implies a lower bound for reconstruction. Note that for any constant  $p$ , the entire graph is selected as a random subgraph with probability  $p^n$  and therefore  $O(1/p^n) = 2^{O(n)}$  is a trivial upper bound on the sample complexity for full reconstruction. So the following lower bound establishes that this trivial upper bound is optimal.

► **Theorem 10.** *Distinguishing the cycle  $C_{2n}$  with high probability from two disjoint copies of the cycle  $C_n$  requires  $2^{\Omega(n)}$  traces in the unordered model.*

**Proof.** Let  $D_1$  be the distribution over subgraphs generated when the original graph is  $C_{2n}$ . Let  $D'_1$  be the distribution conditioned on the event  $A$  that we now define. Partition the vertices of  $C_{2n}$  into  $n$  pairs of “opposite” nodes (i.e. pairs of nodes at a distance exactly  $n$  from each other). Let  $A$  be the event that there exists an opposite pair in which both nodes are deleted. Note that  $\Pr(A) = 1 - (1 - q^2)^n$  where  $q$  is the deletion probability.



■ **Figure 2** The construction for the proof of Theorem 10. The event  $A$  is the deletion of a pair of “opposite” nodes, such as the pair shown in red on the left; the even  $B$  is the deletion of a pair of nodes, one from each of the cycles, such as the pair shown in red on the right. Observe that these two events leave us with the same graph.

Now suppose the input graph is two copies of  $C_n$ . Let  $D_2$  be the distribution over subgraphs generated. Let  $D'_2$  be the distribution conditioned on the following event  $B$ . Partition the vertices into pairs of nodes where each pair consists of exactly one node from each  $C_n$ . Let  $B$  be the event that there exists a pair in which both nodes are deleted. Note that  $\Pr(B) = 1 - (1 - q^2)^n$ . Therefore, by the triangle inequality, we can bound the total variational distance between  $D_1$  and  $D_2$  as follows:

$$\|D_1 - D_2\|_{TV} \leq \|D_1 - D'_1\|_{TV} + \|D'_1 - D'_2\|_{TV} + \|D'_2 - D_2\|_{TV} = O(\Pr(\bar{A}) + \Pr(\bar{B})) = 2^{-\Omega(n)}$$

where we used the fact that  $D'_1 = D'_2$  and substituted  $q = 1/2$ . Hence, we need at least  $2^{\Omega(n)}$  samples to distinguish between  $D_1$  and  $D_2$ . ◀

It is worth remarking here that we needed degree-2 graphs for the example in Theorem 10, as evidenced by the following observation, which we state as a theorem.

► **Theorem 11.** *For any  $\delta > 0$ , a graph with maximum degree one can be reconstructed with probability at least  $1 - \delta$  in  $\Theta(n \log(1/\delta))$  samples.*

**Proof.** A graph with maximum degree one is a matching and some isolated vertices. It suffices, therefore, to learn the size  $k \leq n/2$  of the matching. This size in the random subgraph is distributed as  $\text{Bin}(k, 1/4)$ . But the unknown value  $k$  can be determined by taking the average matching observed over  $t = O(n)$  traces. Specifically, let  $X$  be defined to be the average matching size. Then, because  $k$  is an integer, if we have  $|X - k/4| < 1/8$ , then  $4X$  rounded to the nearest integer is exactly  $k$ . We have  $\mathbb{E}[X] = k/4$  and  $\mathbb{V}[X] \leq k/(4t)$ . Hence, by Chebyshev’s inequality, we have  $\Pr[|X - k/4| \geq 1/8] \leq \frac{k/(4t)}{1/8^2} \leq 1/10$ , where  $t = cn$  for some sufficiently large constant  $c$ . We can boost the probability up as before for an additional  $\log(1/\delta)$  factor. ◀

#### 4 Reconstruction of Low Degeneracy Graphs in the Ordered Model

In this section, we turn our attention to the *ordered* model. We show that the sample complexity of reconstructing graphs with constant degeneracy is  $\exp(\tilde{O}(n^{1/3}))$ , as long as the retention probability  $p$  is a constant. Recall that the *degeneracy* of an undirected graph is the smallest value  $d$  such that every induced subgraph has a node of degree at most  $d$ . Note that the degeneracy is within a factor 2 of the *arboricity*, i.e., the minimum number of forests into which its edges can be partitioned. Hence, the result applies to a natural and large class of graphs, that includes all trees, planar graphs, and indeed, all graphs of bounded treewidth.

## 4.1 Discussion of Challenges

Given a graph  $G$  and its adjacency matrix  $A$ , let us first consider what happens structurally to the matrix  $A$  when we pass it through a deletion channel. For instance, consider the degree sequence, viewed as the sequence of row weights of the matrix  $A$ . Of course, the number of terms in the degree sequence in general decreases (and we would expect about  $pn$  terms to survive in expectation), but observe that the terms that do survive may also change in *value*. In fact, each surviving vertex  $v$  now contributes a term to the degree sequence that is drawn from the distribution  $\text{Bin}(\deg(v), p)$ . However, the *value* of a particular element of the degree sequence in the trace is not in general independent of its *position* within that trace. This is because the decrease in its value is not independent of the total number of neighbors that are deleted in the channel. The *shift* in position, on the other hand, is only dependent on the number of neighbors that appear *before* the corresponding row that are deleted. This is the main difficulty that our approach circumvents.

## 4.2 The Offset Method

We assume that we have a fixed  $n$ -vertex graph  $G$  with a fixed adjacency matrix  $A$ . For any vertex  $i \in [n]$ , we define the *backward* and *forward offsets*.

$$A_{\leftarrow}(i) = \{i - j : j < i, (i, j) \in E\} \quad A_{\rightarrow}(i) = \{j - i : i < j, (i, j) \in E\}$$

For example, if

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

then we have backward offsets  $A_{\leftarrow}(1) = \emptyset$ ,  $A_{\leftarrow}(2) = \{1\}$ ,  $A_{\leftarrow}(3) = \{1\}$ , and  $A_{\leftarrow}(4) = \{1, 2\}$ , and forward offsets  $A_{\rightarrow}(1) = \{1\}$ ,  $A_{\rightarrow}(2) = \{1, 2\}$ ,  $A_{\rightarrow}(3) = \{1\}$ , and  $A_{\rightarrow}(4) = \emptyset$ . Of course,  $A_{\leftarrow}(1) = A_{\rightarrow}(n) = \emptyset$  for any adjacency matrix.

Let  $a_{i,k} = \sum_{x \in A_{\leftarrow}(i)} x^k$  and  $b_{i,k} = \sum_{x \in A_{\rightarrow}(i)} x^k$ , where by convention we set  $0^0 = 0$  and  $1^0 = 1$ . These are the *offset moments of order  $k$*  for vertex  $i$ . Note that  $a_{i,0} + b_{i,0}$  is the degree of vertex  $i$ .

We need the following result due to Krasikov and Roditty, whose proof follows from Corollaries 2.4 and 2.5 in [19].

► **Theorem 12** (Krasikov-Roditty, 1997). *Let  $S = \{u_1, \dots, u_d\}$  be any subset of  $\{0, 1, \dots, n-1\}$  of size  $d$ . Then,  $S$  is uniquely determined by the system*

$$u_1^r + \dots + u_d^r = n_r, \quad r = 1, \dots, d.$$

Now we state the core idea of our proof, which relies on the following observation about the quantities  $\{a_{i,k}\}$  and  $\{b_{i,k}\}$  that we just defined.

► **Theorem 13.** *Let  $G$  have degeneracy  $d$ . Then, the values  $\{a_{i,k}\}_{i \in [n]}$  and  $\{b_{i,k}\}_{i \in [n]}$  for  $k = 0, \dots, d$  uniquely determine  $G$ . In other words, reconstructing  $\{a_{i,k}\}$  and  $\{b_{i,k}\}$  with high probability suffices to reconstruct  $G$  with high probability.*

**Proof.** We use induction on the number of non-isolated nodes. The base case when the number of non-isolated nodes is 0 is trivial. Suppose the Theorem is true when there are up to  $t$  non-isolated nodes. Let  $G$  be a  $d$ -degenerate graph with  $t + 1$  non-isolated nodes. There

exists a node  $i$  with degree at most  $d$ , with backward and forward offsets  $A_{\leftarrow}(i)$  and  $A_{\rightarrow}(i)$  respectively, and offset moments  $a_{i,k}$  and  $b_{i,k}$  for  $k = 0, \dots, d$ . By Theorem 12, given  $a_{i,k}$  and  $b_{i,k}$  we can reconstruct  $A_{\leftarrow}(i)$  and  $A_{\rightarrow}(i)$ .

The idea is to identify this vertex  $i$  and reconstruct its neighbors, and induct on the remaining graph with these edges between node  $i$  and its neighbors removed. Clearly,  $i$  is identifiable from the zeroth moment. Let  $G'$  be the graph formed by removing all edges incident to  $i$ . For all remaining vertices  $j$ , let  $a'_{j,k}$  and  $b'_{j,k}$  be the corresponding offset moments. Observe that  $a'_{j,k}$  and  $b'_{j,k}$  can be computed from  $i$ ,  $a_{j,k}$ ,  $b_{j,k}$ ,  $A_{\leftarrow}(i)$ , and  $A_{\rightarrow}(i)$ ; furthermore,  $G'$  itself is a  $d$ -degenerate graph with at most  $t - 1$  non-isolated nodes, and so by induction, there is no other such graph with  $t - 1$  non-isolated nodes with the same offset moments. We can now reconstruct  $G'$  precisely by induction, and add in the missing edges from vertex  $i$  using  $A_{\leftarrow}(i)$  and  $A_{\rightarrow}(i)$ . ◀

► **Remark 14.** Note that the quantities  $a_{i,k}$  and  $b_{i,k}$  for  $k = 0, 1, \dots, d$  would suffice to learn the neighborhood of node  $i$  directly (given Theorem 12) if the degree of node  $i$  were at most  $d$ , and we could bypass induction altogether. But if the degree is strictly bigger than  $d$ , we would have to first reconstruct other parts of the graph so that, after doing so, there are at most  $d$  unknown edges incident to  $i$ . This is where the inductive argument becomes necessary.

We now state the main result of this section. We relegate the somewhat technical proof to the next subsection. This proof uses similar complex analytic techniques as in [27], which are now standard in the literature, which involve bounding the values of Littlewood polynomials on the unit circle in the complex plane. In our case, crucially, we need to understand how the moments behave, which requires additional work.

► **Theorem 15.**  $\{a_{i,k}\}_{i \in [n]}$  and  $\{b_{i,k}\}_{i \in [n]}$  can be reconstructed with high probability using  $\exp(\tilde{O}(d^{2/3}n^{1/3}))$  traces.

### 4.3 Computing Offset Moments: Proof of Theorem 15

Recall that  $p$  and  $q$  are the retention and deletion probability respectively, so that  $p + q = 1$ . In this subsection, we make a simplification: if  $p \in (1/m, 1/(m - 1)]$  for some integer  $m$ , we assume, without loss of generality, that  $p = 1/m$ . This makes the analysis easier. Of course, given a deletion channel corresponding to retention probability  $p$ , we can always manually simulate one with any lower retention probability, so this is a valid assumption.

Recall that the object of interest in this subsection is the  $k$ th moment, where  $k$  can go up to the degeneracy  $d$ , which suffices by Theorem 13. We denote by  $\tilde{a}_{j,k}$  and  $\tilde{b}_{j,k}$  the *observed* (i.e., sampled) values of  $a_{j,k}$  and  $b_{j,k}$  respectively from our traces. Consider the polynomial  $\sum_{i \geq 1} b_{i,k} w^{i-1}$ , and consider its expected value when we look at the trace from the deletion channel. We have, by linearity of expectation,

$$\mathbb{E} \left[ \sum_{i \geq 1} \tilde{b}_{i,k} w^{i-1} \right] = \sum_{i \geq 1} w^{i-1} \mathbb{E}[\tilde{b}_{i,k}]. \tag{1}$$

Consider the event that the  $i$ th row of the trace comes from the  $j$ th row of the original adjacency matrix  $A$ , for some  $i \leq j \leq n$ . This happens precisely when exactly  $i - 1$  of the first  $j - 1$  rows are retained, and the  $j$ th row is also retained, which happens with probability  $\binom{j-1}{i-1} p^i q^{j-i}$ . However, the *value* of  $b_{j,k}$  changes as well, which we need to account for.

To analyze this change, the first key thing to observe is that the shift in the eventual position of  $b_{j,k}$  is independent of the change in its value; the former is a function of the rows *before*  $j$  that are deleted, while the latter is a function of the rows *after*  $j$  that are deleted, and we assert independence in the deletion probability of each row.

It remains to analyze the expected value of  $b_{j,k}$  after  $A$  is passed through the deletion channel. Recall that  $b_{j,k} = \sum_{x \in A \rightarrow (j)} x^k$ , the sum of  $k$ th powers of the forward offsets in row  $j$ . Now, for a given offset  $x$  in the row under consideration, it survives with probability  $p$ , and if so, ends up as the offset  $1 + y$ , where  $y$  is a random variable that follows a  $\text{Bin}(x - 1, p)$  distribution, since each of the  $x - 1$  columns between the diagonal and the original offset can be deleted with probability  $p$ . Therefore, the value of  $b_{j,k}$  is

$$\sum_{x \in A \rightarrow (j)} p(1 + y)^k, \quad (2)$$

where  $y \sim \text{Bin}(x - 1, p)$ . Since each offset is bounded above by  $n$ , and there are at most  $n$  of them, this expression is bounded above by  $O(n^{k+1})$ .

We also need a lower bound: ignoring the extra factor of  $p$ , each of the  $k + 1$  terms in the expansion of (2) is an integer multiple of a power of  $y$ , and so the expected value of each such term is an integer multiple of  $m^{-k}$  (since the terms less than 1 are all products of the form  $p^r q^s$ , where  $r + s \leq k$ ). Therefore, each nonzero term in the expectation is bounded away in absolute value from 0 by  $m^{-k-1}$ .

An exactly symmetric argument holds for the expected value of  $a_{j,k}$  as well, by ‘‘indexing’’ in the opposite direction to  $b_{j,k}$ . Denote by  $\Phi_a(j, k)$  and  $\Phi_b(j, k)$  these expected values of  $a_{j,k}$  and  $b_{j,k}$  respectively, where  $\Phi_a(j, k)$  and  $\Phi_b(j, k)$  are between  $\Omega(m^{-k-1})$  and  $O(n^{k+1})$ . The lower bound is necessary, as we do not want these expected values to be (exponentially) close to zero.

It follows from the arguments above that (1) reduces to (using backward offsets instead of forward ones, by a symmetric argument):

$$\begin{aligned} \mathbb{E} \left[ \sum_{i \geq 1} \tilde{a}_{i,k} w^{i-1} \right] &= \sum_{i \geq 1} w^{i-1} \sum_{j=i}^n \binom{j-1}{i-1} p^i q^{j-i} \Phi_a(j, k) \\ &= \sum_{j \geq 1} \Phi_a(j, k) \sum_{i=1}^j \binom{j-1}{i-1} p^i w^{i-1} q^{j-i}. \end{aligned}$$

With a change of variables, this becomes

$$\sum_{j \geq 1} \Phi_a(j, k) p \sum_{i'=0}^{j-1} \binom{j-1}{i'} (pw)^{i'} q^{j-1-i'} = \sum_{j \geq 1} \Phi_a(j, k) p(pw + q)^{j-1}$$

We write  $pw + q = z$ . To obtain a lower bound on the variational distance between two distributions coming from two different matrices, say  $A$  and  $A'$  with  $\tilde{a}_{i,k}$  and  $\tilde{a}'_{i,k}$  denoting the distribution of backward offsets from them respectively, we obtain

$$\mathbb{E} \left[ \sum_{i \geq 1} (\tilde{a}_{i,k} - \tilde{a}'_{i,k}) w^{i-1} \right] = \sum_{j=1}^n (\Phi_a(j, k) - \Phi_{a'}(j, k)) p z^{j-1}. \quad (3)$$

The right side of equation (3) is a polynomial in  $z$  of degree less than  $n$ , where by the triangle inequality, each nonzero coefficient is upper bounded in absolute value by  $O(n^{k+1})$ , and lower bounded by  $\Omega(m^{-k-1})$ .

## 96:14 Graph Reconstruction from Random Subgraphs

We now need a technical and non-trivial lemma. This is a generalization of Theorem 7 in [20], which in turns adapts a lemma in [27]. The arguments in those papers can be extended in a straightforward way to apply the same results to powers higher than 2, which is the core idea of this lemma. We use a localized lower bound on a generalized version of Littlewood polynomials, where the coefficients can be (up to) polynomially large instead of being in  $\{-1, 0, 1\}$ . We omit the proof here.

► **Lemma 16** (Generalized Littlewood polynomial bound). *Let  $G(z)$  be a nonzero complex polynomial in  $z$  with its degree bounded by  $n$ , integer coefficients, and each coefficient bounded in absolute value by  $O(n^r)$ . Then, for any fixed positive  $L$ , there is some  $z^* \in \{e^{i\theta} : -\pi/L \leq \theta \leq \pi/L\}$  such that  $|G(z^*)| \geq n^{(1-L)(r+1)}$ .*

This lemma enables us to conclude the following.

► **Lemma 17.** *We have,*

$$\sum_{i \geq 1} |\mathbb{E} [\tilde{a}_{i,k} - \tilde{a}'_{i,k}]| \geq \exp\left(-\tilde{\Omega}(k + k^{2/3}n^{1/3})\right).$$

**Proof.** Multiplying both sides of (3) by  $m^{k+1}$  yields a polynomial precisely of the kind described in Lemma 16 on the right hand side. Therefore, applying Lemma 16, we conclude that for some  $z^* \in \{e^{i\theta} : -\pi/L \leq \theta \leq \pi/L\}$ ,

$$\begin{aligned} m^{k+1} \sum_{i \geq 1} |\mathbb{E} [\tilde{a}_{i,k} - \tilde{a}'_{i,k}]| \cdot |(z^*)^{i-1}| &\geq \left| \mathbb{E} \left[ \sum_{i \geq 1} m^{k+1} (\tilde{a}_{i,k} - \tilde{a}'_{i,k}) (z^*)^{i-1} \right] \right| \\ &\geq n^{(1-L)(k+2 + \frac{\log m}{\log n}(k+1))}. \end{aligned}$$

Noting that  $|z^*| < \exp(C_1/L^2)$  for some finite constant  $C_1$ , we now obtain

$$\begin{aligned} m^{k+1} \sum_{i \geq 1} |\mathbb{E} [\tilde{a}_{i,k} - \tilde{a}'_{i,k}]| &\geq n^{(1-L)(k+2 + \frac{\log m}{\log n}(k+1))} \exp(-C_1 n/L^2) \\ &= \exp(-C_2 k L \log n - C_1 n/L^2) \end{aligned}$$

for some constant  $C_2$ . Here,  $C_2$  is dependent on  $\log m$ , which is constant when the retention probability  $p$  is a constant. The right hand side of this equation is maximized when  $L$  is  $\tilde{O}(n^{1/3}/k^{1/3})$ . We then conclude

$$\sum_{i \geq 1} |\mathbb{E} [\tilde{a}_{i,k} - \tilde{a}'_{i,k}]| \geq m^{-k-1} \cdot \exp\left(-\tilde{\Omega}(k^{2/3}n^{1/3})\right) = \exp\left(-\tilde{\Omega}(k + k^{2/3}n^{1/3})\right),$$

where  $k = O(n)$  can be absorbed into the second term. ◀

To finish the proof of Theorem 15, we just need a standard union bound argument.

► **Theorem 18** (Folklore). *Let  $\mathcal{F}$  be a family of distributions where any two distributions  $A, B \in \mathcal{F}$  have variational distance at least  $\varepsilon$ , for some  $\varepsilon > 0$ . Then, we can distinguish any member of  $\mathcal{F}$  using  $O(\log(|\mathcal{F}|)/\varepsilon^2)$  samples.*

Using Theorem 18 directly on the distributions defined in the proof of Theorem 15 proves that we can recover  $\{a_{i,k}\}_{i \in [n]}$  in  $\exp(\tilde{O}(k^{2/3}n^{1/3}))$  traces. This holds for  $\{b_{i,k}\}_{i \in [n]}$  as well, proving Theorem 15.

## 5 Reconstructing Arbitrary Graphs in the Ordered Model

In this section, we prove that arbitrary adjacency matrices can be reconstructed with high probability using  $\exp(\tilde{O}(n^{1/2}))$  samples in the ordered model. This is in contrast to the unordered model where we showed, in Theorem 10, that  $\exp(\Omega(n))$  samples were necessary. The proof is a small modification of an existing result by Krishnamurthy et al. [20] on the problem on reconstructing arbitrary binary matrices when rows and columns are deleted independently.

► **Theorem 19.** *For graph reconstruction,  $\exp(O(n^{1/2}\sqrt{q\log n}/p))$  traces suffice with high probability to recover an arbitrary adjacency matrix  $A \in \{0, 1\}^{n \times n}$ , where  $p$  is the retention probability and  $q = 1 - p$ .*

**Proof Sketch.** In our problem, the  $i$ th row is deleted iff the  $i$ th column is deleted, while the proof in [20] breaks when there are such dependencies. However, it is possible to make a small modification in the existing proof to handle this. The idea is to re-index the entries of the matrices such that the probability the entry in position  $(i', j')$  of the original matrix ends up in position  $(i, j)$  can be expressed conveniently in terms of two independent random variables. Let us formalize the change and sketch the rest of the approach.

For a matrix  $A \in \{0, 1\}^{n \times n}$ , let  $\tilde{A}$  denote a matrix trace. Let us denote the  $(i, j)$ th entry of the matrix as  $A_{i,j}$ , for  $i, j = 0, 1, \dots, n-1$ , an indexing protocol we adhere to for every matrix. We restrict our attention to the entries above the diagonal, which suffices for reconstruction. For complex numbers  $w_1, w_2 \in \mathbb{C}$ , similar to the proof of Theorem 15, observe that

$$\begin{aligned} \mathbb{E} \left[ \sum_{i,j=0}^{n-1} \tilde{A}_{i,i+j} w_1^i w_2^j \right] &= p^2 \sum_{i,j} w_1^i w_2^j \sum_{k_i \geq i, k_j \geq j} A_{k_i, k_i+k_j} \binom{k_i}{i} \binom{k_j-1}{j-1} p^i q^{k_i-i} p^{j-1} q^{k_j-j} \\ &= p^2 \sum_{k_1=0, k_2=1}^{n-1} A_{k_1, k_1+k_2} (pw_1 + q)^{k_1} (pw_2 + q)^{k_2} \end{aligned}$$

Thus, for two adjacency matrices  $A, B$ , we have

$$\begin{aligned} \frac{1}{p^2} \mathbb{E} \left[ \sum_{i,j=0}^{n-1} (\tilde{A}_{i,i+j} - \tilde{B}_{i,i+j}) w_1^i w_2^j \right] &= \sum_{\substack{k_1=0 \\ k_2=1}}^{n-1} (A_{k_1, k_1+k_2} - B_{k_1, k_1+k_2}) (pw_1 + q)^{k_1} (pw_2 + q)^{k_2} \\ &\triangleq f(z_1, z_2), \end{aligned}$$

where  $z_1 = pw_1 + q$  and  $z_2 = pw_2 + q$ . The rest of the argument is identical to the proof of Krishnamurthy et al. [20]. Specifically, since all the coefficients of  $f(z_1, z_2)$  are in  $\{-1, 0, 1\}$ , and the degree is  $n-1$  in each variable it can be shown that for any  $L > 0$  there exist  $z_1^*, z_2^* \in \{e^{i\theta} : |\theta| \leq \pi/L\}$  such that  $|f(z_1^*, z_2^*)| \geq \exp(-C_1 L^2 \log n)$  for some constant  $C_1$ . If  $z_1^* = pw_1^* + q$  and  $z_2^* = pw_2^* + q$  then  $|w_1^*|, |w_2^*| \leq \exp(C_2 q / (Lp)^2)$  for some constant  $C_2$ .

Substituting these bounds and applying the triangle inequality gives,

$$\begin{aligned} \frac{1}{p^2} \sum_{i,j} \left| \mathbb{E}[\tilde{A}_{i,i+j} - \tilde{B}_{i,i+j}] \right| &\geq \frac{f(z_1^*, z_2^*)}{|w_1^*|^n |w_2^*|^n} \\ &\geq \exp \left( -C_1 L^2 \log n - \frac{2C_2 q n}{L^2 p^2} \right) \\ &\geq \exp \left( -C \frac{\sqrt{nq \log n}}{p} \right) \triangleq \varepsilon \end{aligned}$$

where the second inequality follows by optimizing for  $L$ , similar to our approach in the previous section. So if we estimate each  $\mathbb{E}[\tilde{A}_{i,i+j}]$  and  $\mathbb{E}[\tilde{B}_{i,i+j}]$  up to additive error bounded above by  $p^2\varepsilon/(2n^2)$ , we can distinguish between  $A$  and  $B$ . This immediately leads to the claimed bound via a union bound over all possible pairs and adjacency matrices. ◀

It seems difficult to construct lower bounds the sample complexity of reconstruction in the ordered model, beyond an obvious reduction from string trace reconstruction as follows. Given a binary string  $\sigma = (x_1, \dots, x_n)$ , we can create the ordered graph  $G_\sigma$  on  $n + 1$  vertices  $v_1, \dots, v_{n+1}$  by adding edges of the form  $(v_i, v_{n+1})$  if and only if  $x_i = 1$ . Clearly, there is an injective map from length- $n$  binary strings to  $(n + 1)$ -vertex ordered graphs. Furthermore, taking a trace of a binary string by passing it through a deletion channel with probability  $q$  of deletion corresponds exactly to taking a trace from the ordered subgraph conditioned on the vertex  $n + 1$  being preserved. This automatically implies a lower bound of  $\tilde{\Omega}(n^{3/2})$  samples for graph reconstruction in the ordered model due to [6]. On the surface, the ordered graph reconstruction problem seems to be fundamentally harder than the string trace reconstruction problem as well, but it is significantly harder to improve the lower bound.

## 6 Conclusion and Open Problems

We considered two natural graph reconstruction problems: reconstructing a graph from random induced subgraphs (the unordered model) and reconstructing an graph adjacency matrix via random symmetric submatrices (the ordered model). We showed that for almost all graphs  $G$  on  $n$  nodes,  $\Theta(p^{-2} \log n)$  random induced subgraphs are necessary and sufficient to reconstruct  $G$  with high probability if each subgraph is formed by deleting each node with probability  $1 - p$ . In contrast, we showed that there exist pairs of graphs that require  $2^{\Omega(n)}$  random induced subgraphs to distinguish even when  $p = 1/2$ . We showed that  $\exp(\tilde{O}(n^{1/3}))$  random symmetric submatrices are sufficient to construct sparse graphs (specifically, graphs with constant degeneracy or arboricity) and observed that  $\exp(\tilde{O}(n^{1/2}))$  random symmetric submatrices are sufficient to reconstruct arbitrary graphs.

**Some Open Questions.** In a fairly general sense, our results resolve the sample complexity of graph reconstruction in the unordered model. However, it may interesting to also consider time complexity. For example, the current approach requires isomorphism testing for an exponential number of pairs of subgraphs and it seems plausible that a more efficient approach could exist. For the ordered model, it is natural to ask whether the sample complexity of our upper bounds can be improved. Proving lower bounds for trace construction type problems is notoriously difficult and there is currently an exponential gap between the best lower and upper bounds. However, more tractable open questions include whether our results are optimal for mean-based algorithms [27], i.e., algorithms that only use the expected value of each bit in the trace. Another potential direction is whether it is possible to adapt ideas from Chase's recent work [7] to improve the exponential dependence on  $n$  or the degeneracy  $d$ . This may require substantial work in finding the two-dimensional extensions to several results. Of course, while  $d$ -degenerate graphs are a robust class of structures in themselves, it would be a natural next step to try to relax that condition altogether. Removing this constraint means denser rows of the adjacency matrix, which seems to require many more samples to effectively reconstruct, as well as novel ideas that go beyond the Krasikov-Roditty methods of set reconstruction from [19]. However, there may well be other classes of graphs that our current approach is suitable for and easily generalizable to. For instance, graphs with high girth are a natural candidate for trying to reconstruct using the techniques of this paper, as



small-diameter neighborhoods in an arbitrary high-girth graph are acyclic, and therefore the high-girth condition is akin to a “local” bounded-degeneracy condition. It seems natural to try and extend our approach to this class of graphs as well. Yet another direction to explore would be to consider lower values of  $p$  in the unordered model. Our current bounds are good enough for  $p = \tilde{\Omega}(1/n^{1/6})$ , but we suspect this bound is an artifact of our approach rather than being inherent to the problem. While the majority of the related literature on similar problems typically concern themselves with constant  $p$ , which are covered by our work, exploring e.g.,  $p = O(\text{poly log } n)/n$  may require new techniques.

---

## References

- 1 Frank Ban, Xi Chen, Adam Freilich, Rocco A. Servedio, and Sandip Sinha. Beyond trace reconstruction: Population recovery from the deletion channel. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 745–768, 2019. doi:10.1109/FOCS.2019.00050.
- 2 Tugkan Batu, Sampath Kannan, Sanjeev Khanna, and Andrew McGregor. Reconstructing strings from random traces. In *Symposium on Discrete Algorithms*, 2004.
- 3 Bela Bollobas. Almost every graph has reconstruction number three. *Journal of Graph Theory*, 14(1):1–4, 1990.
- 4 Tatiana Brailovskaya and Miklós Z. Rácz. Tree trace reconstruction using subtraces. *CoRR*, abs/2102.01541, 2021. arXiv:2102.01541.
- 5 Joshua Brakensiek, Ray Li, and Bruce Spang. Coded trace reconstruction in a constant number of traces. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 482–493, 2020. doi:10.1109/FOCS46700.2020.00052.
- 6 Zachary Chase. New lower bounds for trace reconstruction. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 57(2):627–643, 2021. doi:10.1214/20-AIHP1089.
- 7 Zachary Chase. Separating words and trace reconstruction. In *STOC 2021: Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, June 2021*, pages 21–31, 2021.
- 8 Xi Chen, Anindya De, Chin Ho Lee, Rocco A. Servedio, and Sandip Sinha. Polynomial-time trace reconstruction in the smoothed complexity model. In Dániel Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pages 54–73. SIAM, 2021. doi:10.1137/1.9781611976465.5.
- 9 Mahdi Cheraghchi, Ryan Gabrys, Olgica Milenkovic, and Joao Ribeiro. Coded trace reconstruction. *IEEE Transactions on Information Theory*, PP:1–1, May 2020. doi:10.1109/TIT.2020.2996377.
- 10 Sami Davies, Miklos Z. Racz, and Cyrus Rashtchian. Reconstructing trees from traces. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 961–978, Phoenix, USA, 25–28 June 2019. PMLR. URL: <http://proceedings.mlr.press/v99/davies19a.html>.
- 11 Anindya De, Ryan O’Donnell, and Rocco A. Servedio. Optimal mean-based algorithms for trace reconstruction. In *Symposium on Theory of Computing*, 2017.
- 12 Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, 2013. doi:10.1007/978-0-8176-4948-7.
- 13 Lisa Hartung, Nina Holden, and Yuval Peres. Trace reconstruction with varying deletion probabilities. In *Workshop on Analytic Algorithmics and Combinatorics*, 2018.
- 14 Nina Holden and Russell Lyons. Lower bounds for trace reconstruction. *The Annals of Applied Probability*, 30(2):503–525, 2020. doi:10.1214/19-AAP1506.
- 15 Nina Holden, Robin Pemantle, and Yuval Peres. Subpolynomial trace reconstruction for random strings and arbitrary deletion probability. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 1799–1840, 2018.

- 16 Thomas Holenstein, Michael Mitzenmacher, Rina Panigrahy, and Udi Wieder. Trace reconstruction with constant deletion probability and related results. In *Symposium on Discrete Algorithms*, 2008.
- 17 Paul J. Kelly. A congruence theorem for trees. *Pacific Journal of Mathematics*, 7(1):961–968, 1957. doi:pjm/1103043674.
- 18 Géza Kós, Péter Ligeti, and Péter Sziklai. Reconstruction of matrices from submatrices. *Mathematics of Computation*, 2009. doi:10.1090/S0025-5718-09-02210-8.
- 19 I. Krasikov and Y. Roditty. On a reconstruction problem for sequences. *Journal of Combinatorial Theory, Series A*, 1997.
- 20 Akshay Krishnamurthy, Arya Mazumdar, Andrew McGregor, and Soumyabrata Pal. Trace reconstruction: Generalized and parameterized. *IEEE Trans. Inf. Theory*, 67(6):3233–3250, 2021. doi:10.1109/TIT.2021.3066010.
- 21 Thomas Maranzatto and Lev Reyzin. Reconstructing arbitrary trees from traces in the tree edit distance model. *CoRR*, abs/2102.03173, 2021. arXiv:2102.03173.
- 22 Andrew McGregor, Eric Price, and Sofya Vorotnikova. Trace reconstruction revisited. In *European Symposium on Algorithms*, 2014.
- 23 Elchanan Mossel and Nathan Ross. Shotgun assembly of labeled graphs. *IEEE Transactions on Network Science and Engineering*, 6(2):145–157, 2019. doi:10.1109/TNSE.2017.2776913.
- 24 Vladimír Müller. Probabilistic reconstruction from subgraphs. *Commentationes Mathematicae Universitatis Carolinae*, 017(4):709–719, 1976. URL: <http://eudml.org/doc/16787>.
- 25 Shyam Narayanan. Improved algorithms for population recovery from the deletion channel. In Dániel Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pages 1259–1278. SIAM, 2021. doi:10.1137/1.9781611976465.77.
- 26 Shyam Narayanan and Michael Ren. Circular Trace Reconstruction. In James R. Lee, editor, *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*, volume 185 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 18:1–18:18, Dagstuhl, Germany, 2021. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.ITCS.2021.18.
- 27 Fedor Nazarov and Yuval Peres. Trace reconstruction with  $\exp(O(n^{1/3}))$  samples. In *Symposium on Theory of Computing*, 2017.
- 28 Jaroslav Nešetřil. Graph theory and combinatorics. *Fields Institute Summer Thematic Program on the Mathematics of Constraint Satisfaction*, 2011.
- 29 Yuval Peres and Alex Zhai. Average-case reconstruction for the deletion channel: Subpolynomially many traces suffice. In *Symposium on Foundations of Computer Science*, 2017.
- 30 Hannah Spinoza and Douglas West. Reconstruction from the deck of  $k$ -vertex induced subgraphs. *Journal of Graph Theory*, 90(4):497–522, 2019.
- 31 S. M. Ulam. *A collection of mathematical problems*. Interscience Tracts in Pure and Applied Mathematics, no. 8. Interscience Publishers, New York-London, 1960.
- 32 Krishnamurthy Viswanathan and Ram Swaminathan. Improved string reconstruction over insertion-deletion channels. In *Symposium on Discrete Algorithms*, 2008.