Graph Reconstruction from Noisy Random Subgraphs

Andrew McGregor
Manning College of Information and Computer Sciences
University of Massachusetts Amherst
Amherst, MA, USA
Email: mcgregor@cs.umass.edu

Rik Sengupta
MIT-IBM Watson AI Lab
IBM Research
Cambridge, MA, USA
Email: rik@ibm.com

2024 IEEE International Symposium on Information Theory (ISIT) | 979-8-3503-8284-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISIT57864.2024.10619491

***Abstract*—We consider the problem of reconstructing an undirected graph G on n vertices given multiple random noisy subgraphs or "traces". Specifically, a trace is generated by sampling each vertex with probability p_v, then taking the resulting induced subgraph on the sampled vertices, and then adding noise in the form of either a) deleting each edge in the subgraph with probability $1-p_e$, or b) deleting each edge with probability f_e and transforming a non-edge into an edge with probability f_e. We show that, under mild assumptions on p_v, p_e and f_e, if G is selected uniformly at random, then $O(p_e^{-1}p_v^{-2}\log n)$ or $O((f_e-1/2)^{-2}p_v^{-2}\log n)$ traces suffice to reconstruct G with high probability. In contrast, if G is arbitrary, then $\exp(\Omega(n))$ traces are necessary even when $p_v=1, p_e=1/2$.**

I. Introduction

We consider the problem of reconstructing a graph G given noisy observations of random subgraphs of G. We call these observations *traces* and consider two different noise models: edge deletions or edge flips. Formally, we have the following.

Definition 1.1 (Traces): Given a graph $G=(V,E)$, a *trace* $G'=(V',E')$ is a random graph generated as follows: first, each vertex of G is sampled independently with probability p_v, to form $V'\subseteq V$. Then G' is formed from the induced subgraph on V', denoted $G[V']$, by either:

1) *Edge Deletion Trace*: Deleting each edge in $G[V']$ independently with probability $1-p_e$.
2) *Edge Flip Trace*: Deleting each edge in $G[V']$ independently with probability f_e and adding an edge between each non-adjacent pair in $G[V']$ with probability f_e.

Note that the vertices are not labeled; given two traces G'_1 and G'_2, it is impossible in general to determine whether a vertex $v\in G'_1$ and $v'\in G'_2$ correspond to the same vertex in G.

We are interested in the number of independently generated traces that are necessary to reconstruct the graph G (with high probability). We refer to this number as the *sample complexity* of reconstruction. The problem was studied by McGregor and Sengupta [1], who considered the noiseless setting where $p_e=1$ (or equivalently $f_e=0$). They showed that $O(p_v^{-2}\log n)$ traces are sufficient for random graphs drawn from $\mathcal{G}(n,1/2)$ (i.e., n-vertex graphs where edges are present independently with probability $1/2$), assuming $p_v=\Omega(n^{-1/6}\log^{2/3}n)$. They also showed that $2^{\Omega(n)}$ traces are necessary to distinguish arbitrary graphs, even when $p_v=1/2$.

The graph reconstruction problem outlined above is partially inspired by the analogous problem for binary strings, initially proposed by Batu et al. [2] and subsequently studied extensively [3]–[16]. In the case of strings, the traces correspond to random subsequences (potentially subject to further noise). Despite extensive research, there is still a considerable gap between the best known upper and lower bounds on the sample complexity, whether the unknown string is arbitrary or random. The other motivation for our problem is the *graph reconstruction problem* from classical structural graph theory. There, the objective is to reconstruct an undirected n-vertex graph from the multiset of its induced subgraphs on $(n-1)$ vertices. Determining whether this is possible for arbitrary graphs is a famous unsolved problem [17], [18].

A. Our Results

In this paper, we show the following upper bound on the sample complexity of reconstructing random graphs.

Theorem 1.1 (Upper Bound for Random Graphs): Let $G\sim\mathcal{G}(n,1/2)$ and:

$$\begin{aligned} p_v &= \omega(\log n/\sqrt{n})\\ p_e &= \omega(p_v^{-1/3}n^{-1/6}\sqrt{\log n})\\ f_e &= 1/2-\omega(p_v^{-1/4}n^{-1/8}(\log n)^{3/8})\ . \end{aligned}$$

Then, in the edge deletion model, $4p_v^{-2}p_e^{-1}\log n$ traces are sufficient to reconstruct G with probability at least $1-1/n$, where the probability is also taken over the random choice of G. In the edge flip model, the corresponding bound is $12p_v^{-2}(1/2-f_e)^{-2}\log n$.

This theorem generalizes the result by McGregor and Sengupta [1], which only applied when $p_e=1$ or $f_e=0$, i.e., when the the traces are noise-free. However, even in that setting, our approach improves upon the previous result: our algorithm is simpler, and holds for a larger range of p_v values. The improvement is based on a new approach for recognizing when two vertices in difference traces correspond to the same vertex in the original graph.

In Section V, we discuss lower bounds for reconstructing arbitrary graphs. The proof technique used in [1] to establish an $\exp(\Omega(n))$ lower bound when $p_v=1/2$ and $p_e=1$ can be modified to show that $\exp(\Omega(n))$ traces are necessary in the

noisy setting where $p_e = 1/2$, even when $p_v = 1$. However, we conjecture that this bound can be strengthened to show that $\exp(\Omega(n^2))$ traces are necessary. We briefly discuss the challenges in proving such a result.

II. Preliminaries

1) Notation and Conventions: Let $[k]$ denote the set $\{1, \ldots, k\}$ and, for a set S, let $\binom{S}{k}$ denote all subsets with cardinality k. We only consider undirected graphs $G = (V, E)$. We use (u, v) to denote an edge, and $\{u, v\}$ to denote a pair in $\binom{V}{2}$, regardless of whether $(u, v) \in E$ or not. For $v \in V$, $\Gamma_G(v)$ denotes the *neighborhood* $\{v' \in V : (v, v') \in E\}$.

Given two graphs, $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ where $|V_1| = |V_2|$, and a bijection $\pi : V_1 \to V_2$, define the *induced bijection on vertex pairs* to be $\sigma_\pi : \binom{V_1}{2} \to \binom{V_2}{2}$, where $\sigma_\pi(\{u, v\}) = \{\pi(u), \pi(v)\}$. Also, for $\mathcal{S} \subseteq \binom{V_1}{2}$ we define:

$$\Delta_\pi^{\mathcal{S}} := |\{\{u, v\} \in \mathcal{S} : (u, v) \in E_1 \text{ iff } (\pi(u), \pi(v)) \notin E_2\}|$$

and $\Delta_\pi := \Delta_\pi^{\binom{V_1}{2}}$. The quantity Δ_π measures how "far" π is from being an isomorphism (by mapping edges to non-edges, and vice versa). For instance, if G_1 and G_2 are isomorphic, there exists a bijection π such that $\Delta_\pi = 0$. If the mapping is clear from the context, we will suppress the subscript on Δ_π.

Now suppose G_1 and G_2 are subgraphs of traces, and hence V_1 and V_2 are subsets of V. For the sake of analysis, suppose the vertices in V have distinct labels, which V_1 and V_2 inherit (we reiterate that these labels are not available to our reconstruction algorithm).

In this situation, we say $v \in V_1$ is *fixed* by π if $v \in V_1$ and $\pi(v) \in V_2$ have the same label. Otherwise, v is *non-fixed*. Similarly, we say a pair of vertices $\{u, v\} \in \binom{V_1}{2}$ is *fixed* by σ_π if $\{\text{label}(u), \text{label}(v)\} = \{\text{label}(\pi(u)), \text{label}(\pi(v))\}$. The following lemma[1] establishes a lower bound on the number of non-fixed pairs in σ_π.

Lemma 2.1: Suppose the bijection π has b non-fixed vertices and that $|V_1| = |V_2| = n' \geq 6$. Let m_b be the the number of non-fixed pairs in σ_π. Then $m_b \geq b(n' - 1 - b/2) \geq n'b/3$.

2) Correlated Bits and Concentration Bounds: For a random variable X and a probability distribution $\mathcal{D}$, we say $X \sim \mathcal{D}$ to denote that X is distributed according to $\mathcal{D}$. We denote by $\text{Bin}(N, \gamma)$ the binomial distribution with parameters N and γ. We use the following lemma throughout to quantify the probability that, given two traces containing vertices u and v, the edge (u, v) is present in *exactly* one of them.

Lemma 2.2: Let $X_1, X_2, Y_1, Y_2 \in \{0, 1\}, Z_1, Z_2, W_1, W_2 \in \{-1, 1\}$ be independent random variables, where:

$$\Pr[X_i = 1] = 1/2 \qquad \Pr[Z_i = 1] = 1/2$$
$$\Pr[Y_i = 1] = p_e \qquad \Pr[W_i = 1] = 1 - f_e \ .$$

[1]In the interest of space and readability, our technical proofs are omitted and can be found in the full version [19].

for $i \in \{1, 2\}$. Then, we have:

$$\Pr[X_1Y_1 \neq X_2Y_2] = p_e(1 - p_e/2)$$
$$\Pr[X_1Y_1 \neq X_2Y_2 | X_1 = X_2] = p_e(1 - p_e)$$
$$\Pr[Z_1W_1 \neq Z_2W_2] = 1/2$$
$$\Pr[Z_1W_1 \neq Z_2W_2 | Z_1 = Z_2] = 2f_e(1 - f_e) \ .$$

The next lemma establishes concentration bounds that we will need at multiple steps of our analysis.

Lemma 2.3: Let $p_e \leq 1/2$ and $1/4 \leq f_e \leq 1/2$. Define:

$$\gamma_1 = p_e(1 - p_e) \qquad \gamma_2 = 2\gamma_1/3 + \gamma_4/3$$
$$\gamma_3 = \gamma_1/3 + 2\gamma_4/3 \qquad \gamma_4 = p_e(1 - p_e/2)$$
$$\rho_1 = 2f_e(1 - f_e) \qquad \rho_2 = 2\rho_1/3 + \rho_4/3$$
$$\rho_3 = \rho_1/3 + 2\rho_4/3 \qquad \rho_4 = 1/2$$

Then, we have:

$$\Pr[\text{Bin}(N, \gamma_1) \geq \gamma_2 N] \leq \exp(-p_e^3 N/108)$$
$$\Pr[\text{Bin}(N, \gamma_4) \leq \gamma_3 N] \leq \exp(-p_e^3 N/108)$$
$$\Pr[\text{Bin}(N, \rho_1) \geq \rho_2 N] \leq \exp(-(1/2 - f_e)^4 N/4)$$
$$\Pr[\text{Bin}(N, \rho_4) \leq \rho_3 N] \leq \exp(-(1/2 - f_e)^4 N/4) \ .$$

3) Parameter Ranges: In the rest of this paper, we will assume $p_e \leq 1/2$ and $f_e \geq 1/4$ to make the analysis simpler. However, our results immediately hold for larger p_e and smaller f_e values. This follows because, in the edge deletion model, if $p_e > 1/2$, then deleting every edge in the observed traces with probability $(p_e - 1/2)/p_e$ ensures that every edge is ultimately deleted with probability $(1 - p_e) + p_e \cdot (p_e - 1/2)/p_e = 1/2$. In the edge flip model, if $f_e < 1/4$ then flipping the state of every pair in a trace with probability $(1/4 - f_e)/(1 - 2f_e)$ ensures the final flip probability is

$$(1 - f_e) \cdot \frac{1/4 - f_e}{1 - 2f_e} + f_e \cdot \left(1 - \frac{1/4 - f_e}{1 - 2f_e}\right) = 1/4 \ .$$

We may also assume $f_e \leq 1/2$ because otherwise, we can flip the state of each pair in the traces.

III. Reconstructing Random Graphs: Edge Deletion Model

To understand our approach, first suppose the vertices of the unknown graph G have n unique labels, and that these labels are preserved when the traces are generated. If this were the case, in the edge deletion model we would just need to ensure that we take enough traces so that every edge in the original graph was present in at least one trace. We will shortly argue that $\Theta(p_v^{-2} p_e^{-1} \log n)$ traces are sufficient for this to hold with high probability. Unfortunately, in our setting, the vertices of the graph do not *a priori* come equipped with these labels. Our main technical contribution is a systematic way to label the vertices in each trace *consistently*, i.e., two vertices in different traces would receive the same label iff they correspond to the same vertex in G. Our approach will be to construct bijections in order to "pair" common vertices in each pair of traces $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ where $V_1, V_2 \subset V$, i.e.,

we will be able to identify the vertices common to V_1 and V_2. Of course, if we can do this for all pairs of traces without any errors, then we can extend these bijections to equivalence classes; two vertices in different traces will be in the same equivalence class iff they correspond to the same vertex in G. If every vertex appears in at least one trace, then there will be exactly n equivalence classes, which would give consistent labels to the vertices. Once this is done, reconstruction would be easy.

The following key lemma establishes the number of traces required to ensure that every edge in the original graph appears at least once, and shows that if we can pair vertices between each pair of traces with sufficiently high probability, then we can reconstruct the graph.

Lemma 3.1 (Reconstruction via Pairing Traces): Let:

$$p_v = \omega(\log n/\sqrt{n}) \qquad p_e = \omega(p_v^{-1/3} n^{-1/6} \sqrt{\log n}) \ .$$

Given two traces $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ of $G \sim \mathcal{G}(n, 1/2)$, suppose that it is possible to identify the vertices in $V_1 \cap V_2$ and find the correct correspondence between those vertices with probability at least $1 - 1/n^{10}$, where the probability is taken over the generation of G_1, G_2 and G. Then, $t := 4p_v^{-2}p_e^{-1}\log n$ traces are sufficient for reconstruction with probability at least $1 - 2/n^2$.

Proof: First note that

$$t = o((\sqrt{n}/\log n)^{2-1/3} n^{1/6} \sqrt{\log n}) \leq n \ ,$$

for sufficiently large n, given the conditions on p_v and p_e. By the union bound, with probability at least $1 - t^2/n^{10} \geq 1 - 1/n^8$, we can pair up the vertices between every pair of traces. For any $(u, v) \in E$, the probability that this edge is preserved in a given trace is $p_v^2 p_e$ (since both vertices as well as the edge itself need to be preserved). So with t traces, at least one of them preserves this edge with probability $1 - (1 - p_v^2 p_e)^t \geq 1 - \exp(-p_v^2 p_e t)$. Union bounding over n^2 pairs gives us a probability of $1 - n^2 \exp(-p_v^2 p_e t) = 1 - n^{-2}$, since $t = 4p_v^{-2}p_e^{-1}\log n$. So the overall success probability is $1 - 1/n^2 - 1/n^8 \geq 1 - 2/n^2$. ∎

Algorithm 1 describes our procedure for pairing two traces by matching their common vertices. Informally, given two traces $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, we find two induced subgraphs $G_1[S]$ and $G_2[T]$ with $|S| = |T| = k$, that are as close to being isomorphic as possible; specifically, we match their vertices in a way that minimizes the number of vertex pairs that induce an edge in one but not in the other. We set k sufficiently large such that $k \approx |V_1 \cap V_2|$. Our analysis shows that this process is guaranteed to find a large subset of the intersection $V_1 \cap V_2$. We then augment the bijection to also match the remaining vertices in $V_1 \cap V_2$. To do this, for each $v \in V_1$ and $v' \in V_2$, we generate a *signature* based on S and T respectively, and match v and v' iff their signatures are sufficiently similar. The signature is a binary vector that encodes the neighbors and non-neighbors of v (resp. v') amongst S (resp. T). The intuition is that these vectors are sufficiently similar iff v and v' correspond to the same vertex in G.

Algorithm 1 Pairing Traces in the Edge Deletion Model

1: Initialize $r \leftarrow \sqrt{33p_v^2 n \log n}$. If $p_v = 1$, $k \leftarrow n$ and $k \leftarrow p_v^2 n - r$ otherwise.
2: Given traces $G_1 = (V_1, E_1), G_2 = (V_2, E_2)$, find $S^* \subset V_1$ of size k, $T^* \subset V_2$ of size k, and bijection $\pi^* : S^* \to T^*$ that minimizes Δ_{π^*}.
3: Pick an ordering of the elements in $S^* = \{s_1, s_2, \ldots, s_k\}$ arbitrarily. Let $t_i = \pi^*(s_i)$ for all $1 \leq i \leq k$.
4: For $v \in V_1$ and $v' \in V_2$, define binary strings:

$$\mathrm{sig}_1(v) = (I(s_1 \in \Gamma_{G_1}(v)), \ldots, I(s_k \in \Gamma_{G_1}(v)))$$
$$\mathrm{sig}_2(v') = (I(t_1 \in \Gamma_{G_2}(v')), \ldots, I(t_k \in \Gamma_{G_2}(v'))) \ ,$$

where $I(\mathcal{E})$ denotes the indicator function of event $\mathcal{E}$.
5: Pair $v \in V_1$ and $v' \in V_2$ iff

$$\mathrm{H}(\mathrm{sig}_1(v), \mathrm{sig}_2(v')) \leq k(\gamma_1 + \gamma_4)/2 - 1 \ ,$$

where $\mathrm{H}(x, y)$ is the Hamming distance between x and y.

A. Analysis

Let $G = (V, E) \sim \mathcal{G}(n, 1/2)$, $G_1 = (V_1, E_1)$, and $G_2 = (V_2, E_2)$ be defined as above.

Lemma 3.2: Let A_1 be the event that:

$$p_v^2 n - r \leq |V_1 \cap V_2| \leq p_v^2 n + r \ ,$$

where $r = \sqrt{33p_v^2 n \log n}$. Then, $\Pr[A_1] \geq 1 - 2/n^{11}$.

Note that if A_1 occurs, then $|V_1|$ and $|V_2|$ both have size at least $|V_1 \cap V_2| \geq k$, and so there exists at least one triple (S, T, π) where $S \subset V_1, T \subset V_2$, and $|S| = |T| = k$; in other words, step 2 of the algorithm returns some triple. Let $\mathcal{T}$ denote the set of such triples. We next argue that with high probability the triple (S^*, T^*, π^*) that minimizes Δ_{π^*} has mostly fixed vertices. To do this, we define a mapping on triples $f : \mathcal{T} \to \mathcal{T}$ with $f(S, T, \pi) := (S', T', \pi')$, where S' is an arbitrary set of vertices satisfying:

$$|S'| = k \qquad \text{and} \qquad S \cap V_2 \subseteq S' \subseteq V_1 \cap V_2 \ .$$

Let $T' = S'$ and let π' be the identity map. Note that f is well-defined, as $(S', T', \pi') \in \mathcal{T}$. We now show that it is very likely that $\Delta_{\pi'}$ is less than Δ_π if π has many non-fixed vertices.

Lemma 3.3: For any $(S, T, \pi) \in \mathcal{T}$ and $(S', T', \pi') = f(S, T, \pi)$, we have $\Pr[\Delta_\pi > \Delta_{\pi'}] \geq 1 - 4\exp(-kb \cdot p_e^3/1296)$, where b is the number of non-fixed vertices in π.

Proof: Let $\mathcal{N}$ be the set of non-fixed pairs of the induced bijection σ_π, and let $\mathcal{F} = \binom{V_1}{2} - \mathcal{N}$ be the fixed pairs. Then Δ_π can be written as $\Delta_\pi^{\mathcal{N}} + \Delta_\pi^{\mathcal{F}}$. Let $N = |\mathcal{N}|$ and $F = |\mathcal{F}|$.

Claim 3.4: $\mathcal{N}$ can be partitioned as $\mathcal{N}_1 \cup \mathcal{N}_2 \cup \mathcal{N}_3$, s.t. for all i, $|\mathcal{N}_i| \geq N/4$ and $\Delta_\pi^{\mathcal{N}_i} \sim \mathrm{Bin}(N_i, \gamma_4)$.

It follows that $\Delta_\pi = \Delta_\pi^{\mathcal{N}_1} + \Delta_\pi^{\mathcal{N}_2} + \Delta_\pi^{\mathcal{N}_3} + \Delta_\pi^{\mathcal{F}}$. The crucial observation is that all fixed pairs in σ_π are fixed pairs in $\sigma_{\pi'}$, and so $\Delta_{\pi'} = \Delta_{\pi'}^{\mathcal{F}} + \Delta_{\pi'}^{\mathcal{N}_1 \cup \mathcal{N}_2 \cup \mathcal{N}_3}$ where $\Delta_{\pi'}^{\mathcal{F}} = \Delta_\pi^{\mathcal{F}}$ and

$\Delta_{\pi'}^{\mathcal{N}_1\cup\mathcal{N}_2\cup\mathcal{N}_3} \sim \mathrm{Bin}(N,\gamma_4)$. Therefore, $\Pr[\Delta_\pi < \Delta_{\pi'}]$ can be bounded above as:

$$\begin{aligned}
&\Pr\left[\sum_{i\in[3]} \Delta_\pi^{\mathcal{N}_i} < \Delta_{\pi'}^{\mathcal{N}_1\cup\mathcal{N}_2\cup\mathcal{N}_3}\right] \\
&\le \Pr\left[\sum_{i\in[3]} \Delta_\pi^{\mathcal{N}_i} < N\gamma_3\right] + \Pr\left[\Delta_{\pi'}^{\mathcal{N}_1\cup\mathcal{N}_2\cup\mathcal{N}_3} > N\gamma_3\right] \\
&\le \sum_{i\in[3]} \Pr[\mathrm{Bin}(N_i,\gamma_4) < N_i\gamma_3] + \Pr[\mathrm{Bin}(N,\gamma_1) > N\gamma_2] \\
&\le \sum_{i\in[3]} \exp(-p_e^3 N_i/108) + \exp(-p_e^3 N/108)\ ,
\end{aligned}$$

using Lemmas 2.2 and 2.3. This is upper bounded by:

$$\begin{aligned}
\Pr[\Delta_\pi < \Delta_{\pi'}] &\le 3\cdot\exp(-p_e^3 kb/1296) + \exp(-p_e^3 kb/108) \\
&< 4\cdot\exp(-p_e^3 kb/1296)\ ,
\end{aligned}$$

using $N_i \ge N/4$ and $N \ge kb/3$ (Lemma 2.1). ■

Theorem 3.1: Let A_2 be the event that the triple in $\mathcal{T}$ minimizing Δ has no non-fixed vertices. Then $\Pr[A_2|A_1] \ge 1 - 4n\exp(-kp_e^3/2592+2r\log n)$ assuming $4\log n \le k\cdot p_e^3/2592$.

Proof: Let m_b be the number of triples (S,T,π) in $\mathcal{T}$ where π has b non-fixed vertices. Let $m = |V_1 \cap V_2|$. Note that there are at most $\binom{m}{k-b}n^b$ ways to pick S, and then given S, $\binom{k}{k-b}n^b$ choices for π because we can first choose $k-b$ fixed elements of S, and then choose the images of the other b vertices. This also fixes T. Hence, assuming A_1, we have:

$$\begin{aligned}
m_b &\le \binom{m}{k-b}n^b\binom{k}{k-b}n^b \\
&\le \exp(2b\log n + b\log k + (2r+b)\log m) \\
&\le \exp(4b\log n + 2r\log n)\ .
\end{aligned}$$

By Lemma 3.3, for any triple in $\mathcal{T}$ with at least b non-fixed vertices, there exists another triple with all fixed vertices that has a smaller value of Δ with probability at least $1-4\exp(-kbp_e^3/1296)$. So the probability there are any non-fixed vertices is at most $\sum_{b=1}^{n} 4\cdot\exp(-kbp_e^3/1296+4b\log n+2r\log n)$, by the union bound. If $4\log n \le kp_e^3/2592$, then this is at most $4n\exp(-kp_e^3/2592+2r\log n)$. ■

Theorem 3.2: Let $U = V_1\cap V_2$, $m = |U|$, and let π_U be the identity map between vertices $U\subset V_1$ and $U\subset V_2$. Pick an arbitrary ordering of $U = \{u_1,\ldots,u_m\}$. Finally, for all $v\in V_1$ and $v'\in V_2$ define:

$$\begin{aligned}
\mathrm{psig}_1(v) &= (I(u_1\in\Gamma_{G_1}(v)),\ldots,I(u_m\in\Gamma_{G_1}(v)))\ , \\
\mathrm{psig}_2(v') &= (I(u_1\in\Gamma_{G_2}(v')),\ldots,I(u_m\in\Gamma_{G_2}(v')))\ .
\end{aligned}$$

Let A_3 be the event that for all $v\in V_1$ and $v'\in V_2$:

$$\begin{aligned}
v = v' &\Rightarrow \mathrm{H}(\mathrm{psig}_1(v),\mathrm{psig}_2(v')) \le \gamma_2 m\ , \\
v \ne v' &\Rightarrow \mathrm{H}(\mathrm{psig}_1(v),\mathrm{psig}_2(v')) \ge \gamma_3(m-2)\ .
\end{aligned}$$

Then $\Pr[A_3] \ge 1 - 2n^2\exp(-p_e^3 m/216)$.

Proof: If v and v' correspond to the same vertex in G, then $\mathrm{H}(\mathrm{psig}_1(v),\mathrm{psig}_2(v'))$ is distributed as $\mathrm{Bin}(m,\gamma_1)$ or $\mathrm{Bin}(m-1,\gamma_1)$ depending on whether or not $v\in U$. On the other hand, if v and v' are different vertices in G, then the Hamming distance is distributed as $\mathrm{Bin}(m,\gamma_4)$ (if they are both outside U), $\mathrm{Bin}(m-2,\gamma_4)+X+Y$ (if they are both inside U), or $\mathrm{Bin}(m-1,\gamma_4)+X$ (if one is inside U and the other is outside) where $X\sim\mathrm{Bin}(1,p_e/2)$ and $Y\sim\mathrm{Bin}(1,p_e/2)$.

So, if $v = v'$, then using Lemmas 2.2 and 2.3, we get:

$$\begin{aligned}
\Pr[\mathrm{H}(\mathrm{psig}_1(v),\mathrm{psig}_2(v')) > \gamma_2 m] &\le \Pr[\mathrm{Bin}(m,\gamma_1) > \gamma_2 m] \\
&\le \exp(-p_e^3 m/108)\ .
\end{aligned}$$

On the other hand, if $v\ne v'$, we get:

$$\begin{aligned}
&\Pr[\mathrm{H}(\mathrm{psig}_1(v),\mathrm{psig}_2(v')) < \gamma_3(m-2)] \\
&\le \Pr[\mathrm{Bin}(m-2,\gamma_4) < \gamma_3(m-2)] \\
&\le \exp(-p_e^3(m-2)/108) \le \exp(-p_e^3 m/216)\ .
\end{aligned}$$

Applying the union bound over v and v' yields the result. ■

Recall $p_v = \omega(\log n/\sqrt{n})$ and $p_e = \omega(p_v^{-1/3}n^{-1/6}\sqrt{\log n})$. Then, we have:

$$\begin{aligned}
r &= \sqrt{33p_v^2 n\log n} = o(np_e^3p_v^2/\log n) \\
kp_e^3 &= p_e^3(p_v^2 n - r) = p_e^3p_v^2 n(1-o(1)) \\
np_v^2p_e^3 &= \omega(\log n)\ .
\end{aligned}$$

Note that the last two of these imply that $kp_e^3 = \omega(\log n)$, and so for large enough n, $4\log n \le kp_e^3/2592$, so the conditional in Theorem 3.1 applies. Therefore, using Lemma 3.2 and Theorems 3.1 and 3.2, we have:

$$\begin{aligned}
&\Pr[A_1\cap A_2\cap A_3] \\
&\ge 1 - 2/n^{11} - 4n\exp(-kp_e^3/2592 + 2r\log n) \\
&\quad - 2n^2\exp(-kp_e^3/216) \\
&\ge 1 - 2/n^{11} - 4n\exp(-p_v^2np_e^3/2592 + o(p_v^2np_e^3)) \\
&\quad - 2n^2\cdot\exp(-p_v^2np_e^3/216 + o(p_v^2np_e^3)) \\
&\ge 1 - 2/n^{11} - 4n\exp(-\omega(\log n)) - 2n^2\cdot\exp(-\omega(\log n)) \\
&\ge 1 - 1/n^{10}\ .
\end{aligned}$$

Assuming $A_1\cap A_2\cap A_3$, for any v, v', we have:

$$\begin{aligned}
v = v' &\Rightarrow \mathrm{H}(\mathrm{sig}_1(v),\mathrm{sig}_2(v')) \le \mathrm{H}(\mathrm{psig}_1(v),\mathrm{psig}_2(v')) \\
&\qquad \le \gamma_2 m \le \gamma_2 k + 2r \\
v \ne v' &\Rightarrow \mathrm{H}(\mathrm{sig}_1(v),\mathrm{sig}_2(v')) \ge \mathrm{H}(\mathrm{psig}_1(v),\mathrm{psig}_2(v')) - 2r \\
&\qquad \ge \gamma_3(m-2) - 2r \ge \gamma_3 k - 2 - 2r\ .
\end{aligned}$$

Finally, note that:

$$(\gamma_3-\gamma_2)k = kp_e^2/8 = \omega(\sqrt{n}p_v\log n) = \omega(r)\ ,$$

and so $\gamma_2 k + 2r < \gamma_3 k - 2r - 2$ for sufficiently large n. Hence:

$$\frac{k(\gamma_1+\gamma_4)}{2} - 1 = \frac{(k\gamma_2+2r)+(k\gamma_3-2r-2)}{2}\ ,$$

and so the threshold in Algorithm 1 always lies between $\gamma_2 k + 2r$ and $\gamma_3 k - 2r - 2$.

IV. Reconstructing Random Graphs: Edge Flip Model

The algorithm and analysis for the edge flip model follows along almost identical lines to those for the edge deletion model. In fact, almost all of the necessary changes are achieved by replacing every occurrence of γ_i by ρ_i and appealing to Lemmas 2.2 and 2.3. Specifically,

1) The only change in the algorithm is to replace the pairing condition to $\mathrm{H}(\mathrm{sig}_1(v), \mathrm{sig}_2(v')) \le k(\rho_1+\rho_4)/2 - 1$.
2) The lower bound on the probability of A_2 in Theorem 3.1 becomes $1 - 4n\exp(-kf_e(1/2-f_e)^4/96 + 2r\log n)$ assuming $4\log n \le kf_e(1/2-f_e)^4/96$.
3) The event A_3 is defined in terms of ρ_2 and ρ_3, and the lower bound for the probability of A_3 in Theorem 3.2 becomes $1 - 2n^2\exp(-mf_e(1/2-f_e)^4/8)$.

To quickly verify this, note that changing each γ_i to ρ_i and appealing to the second part of Lemma 2.3 results in every occurrence of $p_e^3/108$ getting replaced by $(1/2-f_e)^4/4$. With this substitution, the valid ranges for p_v and f_e become:

$$p_v = \omega(\log n/\sqrt{n}) \tag{1}$$

$$f_e \in [1/4, 1/2 - \omega(p_v^{-1/4} n^{-1/8} (\log n)^{3/8})] \ . \tag{2}$$

Once these ranges are set, it is easy to verify that $k(1/2 - f_e)^4 = np_v^2(1/2 - f_e)^4(1 - o(1)) = \omega(\log n)$, so the conditional in the edge flip equivalent to Theorem 3.1 applies.

Modifying the pairing proceduring (Lemma 3.1) is slightly more involved. We now need enough traces so that for each pair of vertices $\{u, v\} \in \binom{V}{2}$, the majority of traces containing both nodes contain the edge (u, v) iff $(u, v) \in E$.

Lemma 4.1 (Reconstruction via Pairing Traces): Let p_v and f_e satisfy Eqs. 1 and 2. Given two traces $G_1 = (V_1, E_1), G_2 = (V_2, E_2)$, suppose that it is possible to pair the vertices in $V_1 \cap V_2$ with probability at least $1 - 1/n^{10}$. Then, $t := 12p_v^{-2}(1/2-f_e)^{-2}\log n$ traces are sufficient for reconstruction with probability at least $1 - 2/n^2$.

V. Lower Bounds for Arbitrary Graphs

In this section, we consider lower bounds for reconstructing arbitrary graphs in the edge deletion model. For the rest of this section, assume $p_e = 1/2$ and $p_v = 1$.

We first observe that the lower bound technique used in McGregor and Sengupta [1] can be modified to prove a result in this setting, thereby providing an exponential separation between the cases of random graphs and arbitrary graphs.

Theorem 5.1 (Lower Bound for Arbitrary Graphs): Consider the graphs C_n, an n-cycle, and $C_{n/2} + C_{n/2}$, the disjoint union of two $(n/2)$-cycles. Then, $\exp(\Omega(n))$ traces are necessary to distinguish them with constant probability, in the edge deletion model with $p_v = 1, p_e = 1/2$.

We conjecture that reconstructing arbitrary graphs actually requires $\exp(\Omega(n^2))$ traces. Note that this would match the trivial upper bound of $\exp(O(n^2))$, which is a consequence of the fact that with this many traces, one of the traces is likely to be the entire graph!

However, it seems difficult to construct two non-isomorphic graphs such that distinguishing them requires $\exp(\Omega(n^2))$ traces. For instance, consider the following plausible approach.

Let $n = 16r - 8$ for a large integer r, and let P_i denote a path graph with i vertices. Let G_1' be the vertex disjoint union of r copies of P_2, $r-1$ copies of P_3, $r-1$ copies of P_5, and r copies of P_6. Let u_i be a leaf of the ith copy of P_2 and let v_j be either of the middle vertices of the jth copy of P_6. Similarly, let G_2' be the vertex disjoint union of $r-1$ copies of P_2, r copies of P_3, r copies of P_5, and $r-1$ copies of P_6. Let w_k be a leaf of the kth copy of P_3 and let x_ℓ be the middle vertex of the ℓth copy of P_5. Let G_1 (resp. G_2) be the complement of G_1' (resp. G_2'). Note that G_1 and G_2 both have n vertices and are not isomorphic.

Let E_1 be the r^2 edges in G_1 of the form (u_i, v_j), and E_2 be the r^2 edges in G_2 of the form (w_k, x_ℓ). Note that $G_1 - e_1$ is isomorphic to $G_2 - e_2$ for any $e_1 \in E_1$ and $e_2 \in E_2$ (Fig. 1).

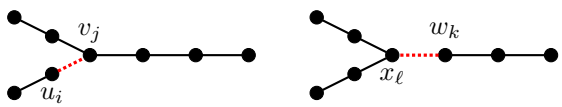

Fig. 1. The subgraph in G_1' formed by adding (u_i, v_j) is isomorphic to the subgraph formed in G_2' by adding (w_k, x_ℓ). Therefore, in the complements G_1 and G_2, removing those edges create isomorphic graphs.

Note that with probability at least $1 - 1/2^{r^2}$, some edge e_1 from E_1 is deleted if the original graph is G_1 (or e_2 from E_2 if the original graph is G_2). Since $G_1 - e_1$ is isomorphic to $G_2 - e_2$, it might then seem reasonable that the variational distance between the distributions of traces generated from G_1 and G_2 is bounded above by $2^{-O(r^2)}$. Since $r = \Theta(n)$, it would then follow that we need $2^{\Omega(n^2)}$ traces to distinguish them. However, this turns out to not be the case.

Proposition 5.1: We can distinguish between G_1 and G_2 with high probability using only $\exp(O(n^{1/3}\log^{2/3} n))$ traces in the edge deletion model with $p_v = 1$, $p_e = 1/2$.

We leave the problem of closing the gap between the upper and lower bounds as an open problem.

VI. Conclusion

Our main result was to establish an upper bound on the number of traces required to reconstruct a random graph with high probability. We note that our result is optimal in the edge deletion setting, since we require $\Theta(p_v^{-2}p_e^{-1}\log n)$ traces to ensure every edge shows up at least once. It is conceivable that a similar analysis can show that the theorem is also optimal in the edge flip setting.

As in many variants of the trace reconstruction problem, an important direction for future research is in the realm of time complexity. While we optimized significantly for the sample complexity, we still require a subroutine that goes over super-exponentially many triples (S, T, π). This process might be sped up, but this is outside the scope of this present work.

Acknowledgment

This work was partially supported by NSF 1934846.

REFERENCES

[1] A. McGregor and R. Sengupta, "Graph Reconstruction from Random Subgraphs," in *49th International Colloquium on Automata, Languages, and Programming (ICALP 2022)*, ser. Leibniz International Proceedings in Informatics (LIPIcs), M. Bojańczyk, E. Merelli, and D. P. Woodruff, Eds., vol. 229. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022, pp. 96:1–96:18. [Online]. Available: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICALP.2022.96

[2] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," in *Symposium on Discrete Algorithms*, 2004.

[3] K. Viswanathan and R. Swaminathan, "Improved string reconstruction over insertion-deletion channels," in *Symposium on Discrete Algorithms*, 2008.

[4] A. McGregor, E. Price, and S. Vorotnikova, "Trace reconstruction revisited," in *European Symposium on Algorithms*, 2014.

[5] N. Holden, R. Pemantle, and Y. Peres, "Subpolynomial trace reconstruction for random strings and arbitrary deletion probability," in *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, 2018, pp. 1799–1840.

[6] F. Nazarov and Y. Peres, "Trace reconstruction with $\exp(O(n^{1/3})$ samples," in *Symposium on Theory of Computing*, 2017.

[7] Y. Peres and A. Zhai, "Average-case reconstruction for the deletion channel: Subpolynomially many traces suffice," in *Symposium on Foundations of Computer Science*, 2017.

[8] N. Holden and R. Lyons, "Lower bounds for trace reconstruction," *The Annals of Applied Probability*, vol. 30, no. 2, pp. 503 – 525, 2020. [Online]. Available: https://doi.org/10.1214/19-AAP1506

[9] T. Holenstein, M. Mitzenmacher, R. Panigrahy, and U. Wieder, "Trace reconstruction with constant deletion probability and related results," in *Symposium on Discrete Algorithms*, 2008.

[10] A. Krishnamurthy, A. Mazumdar, A. McGregor, and S. Pal, "Trace reconstruction: Generalized and parameterized," *IEEE Trans. Inf. Theory*, vol. 67, no. 6, pp. 3233–3250, 2021. [Online]. Available: https://doi.org/10.1109/TIT.2021.3066010

[11] A. De, R. O'Donnell, and R. A. Servedio, "Optimal mean-based algorithms for trace reconstruction," in *Symposium on Theory of Computing*, 2017.

[12] S. Narayanan and M. Ren, "Circular Trace Reconstruction," in *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*, ser. Leibniz International Proceedings in Informatics (LIPIcs), J. R. Lee, Ed., vol. 185. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2021, pp. 18:1–18:18. [Online]. Available: https://drops.dagstuhl.de/opus/volltexte/2021/13557

[13] L. Hartung, N. Holden, and Y. Peres, "Trace reconstruction with varying deletion probabilities," in *Workshop on Analytic Algorithmics and Combinatorics*, 2018.

[14] S. Davies, M. Z. Racz, and C. Rashtchian, "Reconstructing trees from traces," in *Proceedings of the Thirty-Second Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, A. Beygelzimer and D. Hsu, Eds., vol. 99. Phoenix, USA: PMLR, 25–28 Jun 2019, pp. 961–978. [Online]. Available: http://proceedings.mlr.press/v99/davies19a.html

[15] T. Brailovskaya and M. Z. Rácz, "Tree trace reconstruction using subtraces," *CoRR*, vol. abs/2102.01541, 2021. [Online]. Available: https://arxiv.org/abs/2102.01541

[16] T. Maranzatto and L. Reyzin, "Reconstructing arbitrary trees from traces in the tree edit distance model," *CoRR*, vol. abs/2102.03173, 2021. [Online]. Available: https://arxiv.org/abs/2102.03173

[17] P. J. Kelly, "A congruence theorem for trees." *Pacific Journal of Mathematics*, vol. 7, pp. 961–968, 1957. [Online]. Available: https://api.semanticscholar.org/CorpusID:55091877

[18] B. Bollobás, "Almost every graph has reconstruction number three," *J. Graph Theory*, vol. 14, pp. 1–4, 1990. [Online]. Available: https://api.semanticscholar.org/CorpusID:43506446

[19] A. McGregor and R. Sengupta, "Graph reconstruction from noisy random subgraphs," *CoRR*, vol. abs/2405.04261, 2024. [Online]. Available: https://arxiv.org/abs/2405.04261