An Algorithm to Recover Shredded Random Matrices

Caelan Atamanchuk* **Luc Devroye†** **Massimo Vicenzo‡**

July 2, 2024

Abstract

Given some binary matrix M, suppose we are presented with the collection of its rows and columns in independent arbitrary orderings. From this information, can we recover the unique original orderings and matrix? We present an algorithm that identifies whether there is a unique ordering associated with a set of rows and columns, and outputs either the unique correct orderings for the rows and columns or the full collection of all valid orderings and valid matrices. We show that there is a constant $c > 0$ such that the algorithm terminates in $O(n^2)$ time with high probability and in expectation for random $n \times n$ binary matrices with i.i.d. entries $(m_{ij})_{ij=1}^{n}$ such that $\mathbb{P}(m_{ij} = 1) = p$ and $\frac{c \log^2(n)}{n(\log\log(n))^2} \leq p \leq \frac{1}{2}$.

McGill University: caelan.atamanchuk@gmail.com
†*McGill University*: lucdevroye@gmail.com
‡*University of Waterloo*: mvicenzo@uwaterloo.ca

1 Introduction

In this work, we study the problem of reconstructing a binary matrix after being "shredded". That is, we aim to explain when and how a matrix (in our case, drawn from a random model) can be uniquely reconstructed from just the information contained in the rows and columns without knowing how they are ordered. To give the setup more precisely, let $M = (m_{ij})_{i,j=1}^{n}$ be a $n \times n$ binary matrix with the rows and columns given labels in $[n] = \{1, ..., n\}$ and let $\mathcal{C}(M) = \{\gamma_1, ..., \gamma_n\}, \mathcal{R}(M) = \{\rho_1, ..., \rho_n\}$ be the multisets of all the columns and rows of M with some arbitrary ordering that is not necessarily the ones which they belong in. We call these two collections the shredded columns and shredded rows respectively. We say that M is uniquely reconstructible (or just reconstructible) if there are two unique permutations $\sigma = (\sigma_1, ..., \sigma_n)$ and $\tau = (\tau_1, ..., \tau_n)$ of $[n]$ such that

$$\begin{bmatrix} \rho_{\sigma_1} \\ \vdots \\ \rho_{\sigma_n} \end{bmatrix} = \begin{bmatrix} \gamma_{\tau_1} & \cdots & \gamma_{\tau_n} \end{bmatrix}. \tag{1}$$

In particular, if a unique solution exists then both of the resulting matrices are equal to M (there is always at least one solution equal to M, that being the correct, original ordering). If there are at least two pairs of permutations that satisfy (1), then the matrix M is not reconstructible and the collection of all pairs of permutations that satisfy the identities are the potential reconstructions of the original matrix. For example, the matrix that has every entry set to 0 is not uniquely reconstructible and has $(n!)^2$ solutions to (1), even though each of the solutions corresponds to the same matrix. Every matrix that has two equal rows (or columns) is not reconstructible. Suppose

that rows ρ_i and ρ_j are equal, (σ, τ) are a pair of permutations that satisfy (1), and λ is the transposition (ij). Then, the pair $(\lambda \circ \sigma, \tau)$ also satisfies (1) and so the matrix is not reconstructible.

Analogously, we can view M as a square binary picture. The problem is that of rebuilding the picture given the strips that come out after we send one copy of M through a paper shredder upright and one copy sideways. If two strips are identical, we do not know which spots to place the two strips, and we conclude that the picture is not reconstructible. However, for the algorithm, this notion of reconstructibility is not of much importance as all potential reconstructions are outputted.

In this work, we consider a matrix M that has i.i.d. entries m_{ij} with $\mathbb{P}(m_{i,j} = 1) = p$ and $\mathbb{P}(m_{i,j} = 0) = 1 - p$ for some p that we view as a function of n. Since the 1's and 0's are essentially just labels in our model, there is a natural symmetry around $p = \frac{1}{2}$, and thus we assume throughout that $p \leq \frac{1}{2}$ for our analysis.

When $p \geq \frac{(1+\epsilon)\log(n)}{n}$ for some $\epsilon > 0$ the matrix M has pairwise distinct rows and columns with high probability (see Lemma 3). Hence, for p above that threshold, our definition of re-constructibility where matrices that have equal rows or columns are not reconstructible is with high probability equivalent to the simplified version where a matrix M is reconstructible if

$$\forall M', \left[\mathcal{C}(M) = \mathcal{C}(M') \text{ and } \mathcal{R}(M) = \mathcal{R}(M') \implies M = M'\right]. \tag{2}$$

In this paper we present two main results concerning reconstruction in this model. Their full formal presentations can be found in section 4 and section 5. The first result concerns the algorithmic problem of the explicit reconstruction of M (the original matrix) from the shredded rows and columns. It

states that there is a constant $c > 0$ such that, for any $p \geq \frac{c \log^2(n)}{n(\log\log(n))^2}$, there is an algorithm that successfully reconstructs M in $O(n^2)$ time with high probability. Furthermore, the expected running time is also $O(n^2)$ for the same values of p. The exact reconstruction algorithm we use is defined in section 3 and analyzed in section 4. Loosely speaking, the algorithm groups together the columns according to the number of 1's in them, and then analyzes the "subvectors" of the rows that consist of only the entries of the row that are also in columns with k 1's. This is possible because one can line up the rows and verify which indices correspond to columns with k 1's. The number of 1's in these sub-vectors is invariant to permutations of the columns, and so information about row positions can be extracted. The second result is purely theoretical and states that, when $p \geq \frac{2(1+\epsilon)\log(n)}{n}$ for any arbitrary $\epsilon > 0$, the probability that M is reconstructible tends to 1 as $n \to \infty$.

The paper is organized as follows: In section 2 we discuss related work and some of the motivations behind work in the area of reconstruction problems. In Section 3 we present the reconstruction algorithm along with our main result, and in Section 4 we prove the result. In section 5 we prove a result concerning when matrices can be reconstructed. Section 6 houses proofs for the lemmas used in the preceding sections. Finally, section 7 contains a couple of open directions for further research.

2 Related work and motivation

Many existing works on reconstruction problems deal with structures on graphs. Before discussing these problems it is helpful to make explicit how our matrix reconstruction problem can also be seen as a graph reconstruction

one.

Let G be a directed graph. In G an edge (u, v) seen as being directed from u to v, we call v the out-neighbour of u and u the in-neighbour of v. Furthermore, we define the 1-in(out)-neighbourhood of a vertex v is all vertices that are in(out)-neighbours of v. v is called the central vertex of the neighbourhood. If M is viewed as the adjacency matrix for some random directed graph on n vertices (one with loops allowed), the columns $\gamma_1, ..., \gamma_n$ represent the collection of all 1-in-neighbourhoods with only the central vertex's label removed and $\rho_1, ..., \rho_n$ represent the collection of all 1-out-neighbourhoods with only the central vertex's label removed (removing the labels is the same as permuting them into some arbitrary labelling). In this new view, we can reword our problem as follows: given a collection of random 1-in-neighbourhoods $\gamma_1, ..., \gamma_n$ and 1-out-neighbourhoods $\rho_1, ..., \rho_n$, determine if there is a unique directed (with loops) graph containing exactly these 1-in(out)-neighbourhoods and if one exists, reconstruct the original graph efficiently. Problems closely matching this form have received attention from combinatorists and probabilists. In this section, we briefly discuss some of these graph reconstruction models and other related works. There is also a way of looking at our model that turns it into a reconstruction problem on bipartite graphs. We mention this idea at the end of this section.

Combinatorial reconstruction problems arise naturally in several pure and applied settings. The largest inspiration for such exploration comes from the reconstruction conjecture in combinatorics (see Harary (1974), Harary and Plantholt (1985), Kelly (1957), and Ulam (1960)): any graph G on at least three vertices is reconstructible from the multiset of isomorphism classes of all the vertex-deleted subgraphs of G, often called the deck or G and labelled $D(G)$ (the vertex deleted subgraphs of G are all the induced

subgraph obtained through deleting exactly one of the vertices of G). To be more exact, the conjecture states that for all graphs G and H on at least three vertices, G is isomorphic to H if and only if $D(G) = D(H)$. The use of random models has been vital in the study of this conjecture, with one important result coming from Bollobás (1990) who proved that as $n \to \infty$, an Erdős-Rényi random graph with $\frac{c\log(n)}{n} \leq p \leq 1 - \frac{c\log(n)}{n}$ is uniquely reconstructible from a collection of only three of the vertex deleted subgraphs for any $c > \frac{5}{2}$. In particular, this means that with high probability, for appropriate choices of p, there is a subset $\{G_1, G_2, G_3\} \subseteq D(G)$ of three subgraphs such that for any other graph H, if $\{G_1, G_2, G_3\} \subseteq D(H)$, then H is isomorphic to G. Before Bollobás' result, Müller (1976) had previously explored the reconstruction of random graphs from the whole deck.

One interesting abstraction of the reconstruction conjecture related to the random pictures model is the new digraph reconstruction conjecture. Let G and H be two directed graphs and suppose that there is a bijection $f : V(G) \to V(H)$ such that $G \setminus v$ is isomorphic to $H \setminus f(v)$ for all $v \in V(G)$. Further, suppose that the in-degrees and out-degrees of v and $f(v)$ match for all $v \in V(G)$. Then, G and H must be isomorphic. The answer to this problem remains open. See Ramachandran (1981) and Ramachandran and Arumugam (2004) for a discussion of the problem and the families of graphs for which the conjecture has been proven to be true.

Recently, extensive work has gone into studying the shotgun assembly problem for graphs. Introduced by Mossel and Ross (2019), the problem asks how large must r be so that a graph, commonly drawn from some random model, is uniquely determined by its collection of distance r-neighbourhoods around each vertex $v \in V(G)$ (by a distance r-neighbourhood of v we mean the subgraph $N_r(v)$ that is induced by all vertices of graph distance at most

r from v). They consider both labelled and unlabelled versions of the problem. This topic has been studied for a variety of random models including Erdős-Rényi graphs, random regular graphs, and simplicial complexes (for examples, see Adhikari and Chakraborty (2022), Ding, Jiang, and Ma (2022), Gaudio and Mossel (2022), Huang and Tikhomirov (2022), Johnston et al. (2023), and Mossel and Sun (2015)). There has also been work put towards shotgun assembly problems in different contexts such as reconstructing random vertex colourings from r-neighbourhoods as seen in Ding and Liu (2022), Mossel and Ross (2019), and Przykucki, Roberts, and Scott (2022).

In a similar vein, there is the problem of canonically labelling graphs and random graphs, and its main application in checking graph isomorphisms (early work in the topic can be seen in Babai (1980), Babai, Erdös, and Selkow (1980), and Babai and Luks (1983)). An algorithm which canonically labels a graph G, assigns the labels $1, 2, \ldots, n$ to the n vertices of G such that if G is isomorphic to some graph H, then both should be given the same labelling by the algorithm. Of particular note are the results on canonically labelling the Erdős-Rényi graph using only the r-neighbourhoods of each vertex. Mossel and Ross (2019) showed it is possible to canonically label a graph $G \sim G(n, p_n)$ when $np = \omega(\log^2(n))$ with using only 2 neighbourhoods. On the other hand, Gaudio, Rácz, and Sridhar (2022) showed for $np = o(\log^2(n)/(\log\log(n))^3)$ there are multiple isomorphic 2-neighbourhoods with high probability, which inhibits us from creating a canonical labelling.

Some papers deal with the reconstruction of random jigsaw puzzles (Mossel and Ross (2019)). Here, we are given the collection of vertices in a lattice with coloured half-edges drawn from some collection of q colours. The prob-

lem is to determine how large q must be so that with high probability the puzzle can be constructed into a complete picture from the collection of vertices and their coloured half-edges. Some work concerning this problem can be found in Balister, Bollobás, and Narayanan (2019), Martinsson (2016, 2019), and Nenadov, Pfister, and Steger (2017).

There is no lack of motivation from other sciences for studying reconstruction problems, such as the problem of DNA shotgun sequencing. In shotgun assembly, the long DNA strands are "shotgunned" into smaller pieces that are sequenced. From here, a reconstruction algorithm is used to infer what the original long strand was. For a probabilistic analysis of the unique re-constructibility of DNA sequences from shotgunned strands see Arratia and Reinert (1996), Dyer, Frieze, and Suen (1994), and Motahari, Bresler, and Tse (2013). Note that the models here are what one of the shotgun assembly problems from Mossel and Sun (2015) is based on, with the special case of the path on n vertices being studied. The shotgun assembly has also begun to appear in neural network theory. Soudry et al. (2015) consider the problem of reconstructing large neural networks from smaller sub-networks.

The topic of this paper, reconstructing random matrices, has been studied before from another point of view. In Narayanan and Yap (2023), the complete multiset of all $(n-k)^2$ $k \times k$ sub-matrices of an $n \times n$ matrix is given as the information to reconstruct with. This multiset is called the k-deck of the matrix. They proved what they call "two-point concentration", which loosely states that the probability that the matrix is reconstructible from the k-deck converges to 1 as $n \to \infty$ for $k > (2\log_2(n))^{1/2} + \frac{3}{4}$ and converges to 0 as $n \to \infty$ for $k < (2\log_2(n))^{1/2} + \frac{1}{4}$. This model has also been referred to as the reconstruction of random vertex colourings on the $n \times n$ grid graph. Further work has been done under this second name in Demi-

dovich, Panichkin, and Zhukovskii (2023), where the authors provided similar "two-point" concentration theorems for the case of general d-dimensional grids with r colours. They find the critical threshold for reconstructibility in general to be $\sim (d \log_r(n))^{1/d}$. Furthermore, in this second paper, they also consider the reconstruction of r colourings in more general families of random graphs from k-decks (for a general graph, the k-deck is the multiset of all induced k-vertex subgraphs), and even make connections between the reconstruction of random colourings on graphs and reconstruction of random graphs.

Since the release of the original pre-print copy of this paper work has been done on the model. Balister, Kronenberg, et al. (2024) have furthered our theoretical result, proving that $\frac{1}{n}\log(n)$ is the sharp threshold for our definition of shredded matrix reconstructibility, i.e., for $p \geq \frac{(1+\epsilon)\log(n)}{n}$ M is reconstructible with probability tending to 1 as $n \to \infty$ and for $p \leq \frac{(1-\epsilon)\log(n)}{n}$ the probability of reconstructibility is tending to 0 as $n \to \infty$. They also introduce the notion of weak reconstructibility and find that its sharp threshold is at $p = \frac{\log(n)}{2n}$. A matrix M is said to be weakly reconstructible if the condition in (2) is satisfied. Note that this definition allows matrices with identical rows and columns to be reconstructible. Algorithms for reconstruction are also provided. In this work, the authors establish these results through a key connection between the reconstruction of matrices and the reconstruction of random bipartite graphs on $2n$ vertices. In particular, the graph they consider is a random subgraph of the complete bipartite graph $K_{n,n}$ where each edge is deleted with probability $(1-p)$. If one labels the vertices on each side of the $K_{n,n}$ $1, ..., n$, the matrices considered in this paper are equivalent to adjacency matrices for these bipartite graphs.

3 The reconstruction algorithm

For a vector $x = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$, we call $|x| = \sum_{i=1}^{n} x_i$ the weight or Hamming weight of x. If $S \subset [n]$ is a set of indices, then $\sum_{i \in S} x_i$ is the sub-weight of x on S. Alternatively, the weight of x can be seen as the number of 1's which appear in the entire vector, and the sub-weight in S is the number of 1's in the vector x restricted to the positions indicated by S. We have two algorithmic problems to solve:

(i) Find any permutation pair (σ, τ) that satisfies (1).

(ii) Find all permutation pairs (σ, τ) that satisfy (1).

Our algorithm solves (ii) and hence also (i). It can be broken down into two main parts: First, we partition each row ρ_i into sub-strings and compute the vector of the associated sub-weights for all $i \in [n]$. Then, using a trie (see below for the definition of a trie), we can efficiently identify each ρ_i with a position by matching these sub-weight vectors. If we can identify each ρ_i with a unique position, then the algorithm is complete. We show this happens with high probability.

In the case where this does not occur, we move on to part two of the algorithm, where we iterate through all possible permutations of the rows and check if the matrix is correct by checking if it contains all of the columns in $\mathcal{C}(M)$ with the correct multiplicities. Using the information gained from part one, we can reduce our search space from all $n!$ permutations of the rows to a collection that has expected size $O(1)$.

A prefix trie (or simply just a trie) is a k-ary search tree data structure used to store vectors or strings from some finite alphabet (Fredkin (1960), Briandais (1959)). A sequence $x_1, \ldots, x_\ell$ of symbols drawn from $\{1, \ldots, k\}$

defines a path of length ℓ from the root down, where the x_i's indicate which child is taken in the i-step. The trie for n strings of length ℓ is the data structure that stores the union of the n paths. In time $O(\ell)$, one can for example determine whether a given string of length ℓ matches one of the n strings that are stored in the trie. For a more detailed explanation of tries, one may consult Knuth (1997), Knuth (1998), or Morin (2013).

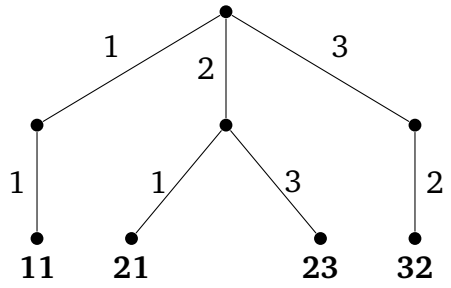


Figure 1: An example of a trie with alphabet $\{1, 2, 3\}$.

The algorithm described in the rest of this section only utilizes information from columns with weight in the set $[\lfloor np \rfloor, \lfloor np \rfloor + \lfloor \sqrt{np} \rfloor]$. Restricting ourselves only to weights in this set does not result in too much information being lost. Since each entry is independently 1 with probability p, $|\gamma_j|$ is a $\text{binomial}(n, p)$ random variable, implying that it is concentrated around its mean. In particular, using a standard Chernoff bound for binomials we get that

$$\mathbb{P}\left(|\gamma_j| \geq np + (np)^{1/2+\epsilon}\right) \leq e^{-\frac{1}{3}(np)^{\epsilon}} \to 0$$

as $n \to \infty$ when $np \to \infty$ (in the cases we consider $np \to \infty$ always). Furthermore, the median of binomial random variables is in $\{\lfloor np \rfloor, \lceil np \rceil\}$. Combining these two facts, we can see that approximately $\frac{1}{2}$ of the columns weight the range $[\lfloor np \rfloor, \lfloor np \rfloor + \lfloor \sqrt{np} \rfloor]$ in the limit.

3.1 Part One

Given the collection of unordered columns $\gamma_1, ..., \gamma_n$, we create a Hamming weight partition of the columns $\mathcal{P} = (\mathcal{P}_0, \ldots, \mathcal{P}_n)$, where $\mathcal{P}_i = \{1 \leq j \leq n : |\gamma_j| = i\}$. Now for each $j \in [n]$, and for each integer $k \in [\lfloor np \rfloor, \lfloor np \rfloor + \lfloor \sqrt{np} \rfloor]$, we compute

$$s_{j,k} = \sum_{i \in \mathcal{P}_k} \gamma_{ij}, \quad \text{where} \quad \gamma_i = \begin{bmatrix} \gamma_{i,1} \\ \vdots \\ \gamma_{i,n} \end{bmatrix}.$$

For a row to be able to be put in position j, its sub-weight on $\mathcal{P}_k$ must be equal to $s_{j,k}$ for all $k \in [\lfloor np \rfloor, \lfloor np \rfloor + \lfloor \sqrt{np} \rfloor]$. Using the values $s_{j,k}$ we store every potential position $j \in [n]$ in the leaves of a trie using the vectors $S_j = (s_{j,\lfloor np \rfloor} \ldots, s_{j,\lfloor np \rfloor + \lfloor \sqrt{np} \rfloor})$ as input, which we call the sub-weight vectors associated with position j. In our trie, we associate each input with a path. Therefore, it is possible that several paths coincide and that S_j is not unique, i.e., $|\{S_j : 1 \leq j \leq n\}| < n$.

From the collection of rows $\rho_1, ..., \rho_n$, we can compute the weight of each column in the original matrix M even without knowing the order, since the weight of a column is invariant under permutation of the rows. This allows us to determine which column positions have which weights. Let $\mathcal{I} = (I_0, I_1, \ldots, I_n)$, where

$$I_j = \{i \in [n] : \text{The column in position } i \text{ has weight } j\}.$$

Now, for all $j \in [n]$, and for each integer $k \in [\lfloor np \rfloor, \lfloor np \rfloor + \lfloor \sqrt{np} \rfloor]$, we compute $t_{j,k}$, which is the sub-weight of the row ρ_j on the indices I_k. We

collect all of them into a vector

$$T_j = \left(t_{j,\lfloor np \rfloor} \ldots, t_{j,\lfloor np \rfloor + \lfloor \sqrt{np} \rfloor}\right),$$

which we call the signature of ρ_j. Since entries S_j (the sub-weight vector of position j) and T_j (the signature of ρ_j) are generated from the same information with only potentially incorrect labels on T_j, we know that

$$\{S_j : 1 \leq j \leq n\} = \{T_j : 1 \leq j \leq n\}.$$

It follows that if $|\{S_j : 1 \leq j \leq n\}| = n$, then are able to identify a unique permutation for each row: For each $j \in [n]$, we define σ_j to be the unique $\ell \in [n]$ such that $S_j = T_\ell$. Once the rows have been placed we have reconstructed the matrix and the permutation τ on the unordered columns can be determined. We do this by first constructing a trie based on all of the columns $C_1, ..., C_n$ in the reconstructed matrix M (these are the columns in their original, pre-shredded positions). If the trie has n distinct leaves, then we can define a permutation τ for $\gamma_1, ..., \gamma_n$ in the following way: For each $j \in [n]$, define τ_j to be the unique $\ell \in [n]$ such that $\gamma_j = C_\ell$. If either of the two tries do not have distinct leaves we move on to part 2.

3.2 Part Two

There are two possible cases where we end up requiring part two to complete the algorithm. First, we require part two when there is at least one leaf in the trie containing row sub-weight vectors which coincide with multiple rows, i.e. $|\{S_j : 1 \leq j \leq n\}| = L < n$. The second case where we require part two is when at least two columns coincide with a single leaf in the trie containing

the column vectors, i.e. $C_j = C_k$ for $j \neq k$.

For each vector $S_i \in \{S_j : 1 \leq j \leq n\}$, let x_i be the multiplicity of that vector, i.e. the number of rows ρ_j where $S_j = S_i$. Then, since ρ_j can only be assigned to a position k such that $S_j = T_k$, there are $x_1!x_2!\ldots x_L!$ possible permutations of $\rho_1, ..., \rho_n$ that must be checked. For each possible permutation σ, we construct a matrix,

$$M' = \begin{bmatrix} \rho_{\sigma_1} \\ \vdots \\ \rho_{\sigma_n} \end{bmatrix}.$$

Using the column trie, we determine whether $\mathcal{C}(M) = \mathcal{C}(M')$, i.e., we determine whether both matrices contain the same set of columns with the same multiplicities. If this is the case, then M' is a valid reconstruction. Let τ_j be an $\ell \in [n]$ such that the column in position j in M' is equal to γ_ℓ for all $j \in [n]$ (in particular choose the τ_j such that $\tau = (\tau_1, ..., \tau_n)$ is a permutation). Note that at this point, ℓ need not be unique so this could yield many valid matrices. The pair (σ, τ) permutes the rows and columns to create a valid reconstruction M'. Let $I_1, \ldots, I_m$ be the sets of column indices ($|I_k| > 1$) such that for every two indices $i, j \in I_k$, the columns C_i, C_j in M' are equal. The columns within each I_k can be permuted and still give a valid τ for reconstructing.

Therefore, for every valid σ we compute one of the corresponding column permutations τ and the sets of indices $I_1, \ldots, I_m$, then output

$$(\sigma, \tau, S_{I_1} \times \cdots \times S_{I_m}),$$

where S_{I_k} is the group of permutations on the elements in the set I_k. The

set of these triples can generate all of the pairs (σ, τ) and defines a valid reconstruction. If we wish to retrieve every pair from the triple, we need only iterate over $\pi \in S_{I_1} \times \cdots \times S_{I_k}$ and compute $(\sigma, \pi\tau)$.

3.3 An Example

Consider the following collection of its rows and columns (assume that $\mathcal{C}(M) = \{\gamma_1, \gamma_2, \gamma_3, \gamma_4\}$ and $\mathcal{R}(M) = \{\rho_1, \rho_2, \rho_3, \rho_4\}$ are ordered left to right and top to bottom respectively):

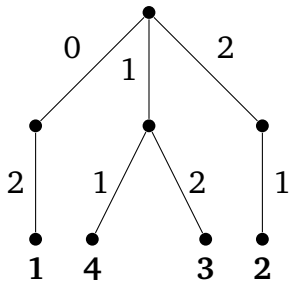
$$\mathcal{C}(M) = \left(\begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \right), \quad \mathcal{R}(M) = \left(\begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0 & 1 \end{bmatrix} \right).$$

We first construct the partition $\mathcal{P}$ from the column collection $\mathcal{C}(M)$. From this, for each position j, we compute the sub-weight vectors $S_j = (s_{j,2}, s_{j,3})$,

$$\mathcal{C}(M) = \left\{ \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \right\} \quad \mathcal{P} = \left(\emptyset, \emptyset, \left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \right\}, \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \right\}, \emptyset \right) \rightarrow \left\{ \begin{matrix} (0,2), \\ (2,1), \\ (1,2), \\ (1,1) \end{matrix} \right\}.$$

In this case, each of the S_j is distinct, so the trie we construct with them has exactly n leaves. Each of the leaves contains the indices of the positions with

sub-weight vectors which take them to said leaf (in bold):



Next, we compute the signatures for each of the row vectors. We do this by first computing $\mathcal{I}$, and using the indices to determine the values of each entry,

$$\mathcal{R}(M) = \left\{ \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0 & 1 \end{bmatrix} \right\} \quad \mathcal{I} = (\emptyset, \emptyset, \{1,4\}, \{2,3\}, \emptyset) \rightarrow \left\{ \begin{matrix} (1,1), \\ (0,2), \\ (1,2), \\ (2,1) \end{matrix} \right\}.$$

Now we use the set of signatures and search through the trie generated by the sub-weight vectors. Each signature reaches a leaf, which then tells us which positions that row is allowed to be in. In this example, they are each mapped to a unique position, telling us that $\sigma = (142)(3)$ is the permutation to apply on $\mathcal{R}(M)$ to obtain M. Doing so gives us our unique matrix

$$M = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

Since we have no duplicate columns, there is also a unique $\tau = (13)(24)$. The final output would be $((142)(3), (13)(24), \{\mathrm{Id}\})$ as there is only one way to permute the columns to reconstruct the matrix.

For a second example, let us consider a case where we have duplicate sub-weight vectors and duplicate columns. Below is the result of doing part one to some matrix M, we can see that the first row in $\mathcal{R}(M)$ belongs in the second position, but the remaining rows' positions are unknown,

$$\mathcal{R}(M) = \left\{ \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix} \right\} \to \begin{Bmatrix} (1,0), \\ (1,2), \\ (1,2), \\ (1,2) \end{Bmatrix} \quad \mathcal{C}(M) = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \right\} \to \begin{Bmatrix} (1,2), \\ (1,0), \\ (1,2), \\ (1,2) \end{Bmatrix}.$$

As three rows have the same signature, we have 6 permutations of the rows to check,

$$\{(12), (12)(34), (123), (124), (1234), (1243)\},$$

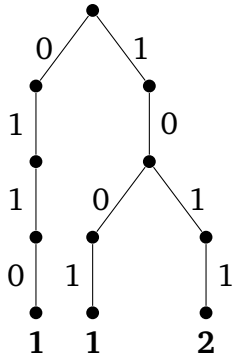
which results in matrices,

$$M_{(12)} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad M_{(12)(34)} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

$$M_{(123)} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \qquad M_{(124)} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

$$M_{(1234)} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \qquad M_{(1243)} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}.$$

Since there are duplicate rows, some of these permutations result in the same matrix. Regardless, using the column trie below, we can iterate through each M_σ and see if $\mathcal{C}(M) = \mathcal{C}(M_\sigma)$:



From this we can see that the only σ that give us valid matrices are from (123) and (1234), and since there are two identical columns in positions 2 and 3, the corresponding permutation groups are both $S_{\{2,3\}}$. The solution set for this example is

$$\{((123), (142), S_{\{2,3\}}), ((1234), (142), S_{\{2,3\}})\}.$$

3.4 Time Complexity

The time complexity achieved by our algorithm assumes the RAM model of computation. Computing the weights of the vectors and computing all the sub-weights takes time $O(n^2)$, since we can upper bound both of these by computing the sum of all entries in the matrix. Creating the trie with the sub-weight vectors would take time $O(n^{3/2})$ since the size of the strings used in the trie is bounded above by $\sqrt{np} \leq \sqrt{n}$. Since the height of the trie is $O(\sqrt{n})$, matching each R_i to a set of positions at a leaf, takes total time $O(n^{3/2})$. Next, we create the column trie, which takes $O(n^2)$ as we have n length n vectors to insert. It is interesting to note that determining which rows belong in which positions is not the most time-intensive step; in fact, simply determining the weights of the vectors is what gives us our time complexity.

In part two, for each valid permutation, we first check that $\mathcal{C}(M)$=$\mathcal{C}(M')$ by searching for each column in M' in the column trie, keeping track of multiplicities. This takes time $O(n^2)$. Once a valid σ is found, we must compute a single τ, which we can get from reading the columns of M' generated by σ applied on the rows, in $O(n)$ time. Using the column trie, we can create the sets $I_1, \ldots, I_m$ in $O(n^2)$ time.

Let $P = x_1!x_2! \ldots x_L!$ be the number of permutations σ that we have to check. Then part two takes expected time $O(n^2\mathbb{E}[P])$. In section 4, we show that $\mathbb{E}[P] \to 1$ as $n \to \infty$ for p in some range, implying that the expected time for the algorithm is $O(n^2)$. We also show that the probability we require step two to complete the algorithm tends to 0 as $n \to \infty$ for p in another range, implying that the completion time is also $O(n^2)$ with high probability.

4 Main result

The time complexity discussion from the previous section culminates in our main result.

Theorem 1. *If* $p \geq \frac{16(1+\epsilon)\log^2(n)}{n(\log\log(n))^2}$ *for* $\epsilon > 0$*, then,*

$$\mathbb{P}(\textit{Algorithm terminates at first step}) \to 1 \textit{ as } n \to \infty.$$

Hence, with high probability, the algorithm produces a unique reconstruction in $O(n^2)$ *time. Furthermore, if* $p \geq \frac{36(1+\epsilon)\log^2(n)}{n(\log\log(n))^2}$ *for* $\epsilon > 0$*, the expected running time of the algorithm is also* $O(n^2)$*, with the expected number of permutations that require checking in step two converging to 1 as* $n \to \infty$

To complete the proof of Theorem 1, we need to bound the probability that two rows ρ_i and ρ_j share the same signature vectors T_i and T_j. To analyze this we need to obtain some bounds on the size of each group in the partition $|\mathcal{P}| = (|\mathcal{P}_1|, ..., |\mathcal{P}_n|)$. In particular, we want the groups near the average np to be sufficiently large as these columns are the ones that the algorithm uses to generate sub-weight vectors and larger sub-strings produce sub-weights with larger variance. Since each column sum is a $\text{binomial}(n, p)$ random variable, and we have n distinct columns, $|\mathcal{P}|$ has a multinomial distribution with parameters n and $b = (b_{n,p,1}, ..., b_{n,p,n})$, where

$$b_{n,p,k} = \mathbb{P}(\text{binomial}(n, p) = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

The bounds we desire for $|\mathcal{P}|$ are shown in the following lemma.

Lemma 2. *Suppose that* $p = p(n)$ *is some sequence such that* $np \geq 16$*. There exists a positive constant* $\gamma > 0$ *such that* $b_{n,p,\lfloor np \rfloor + i} \geq 2\gamma \frac{1}{\sqrt{np}}$ *for all*

$i \in [0, \lfloor\sqrt{np}\rfloor]$. *Furthermore,*

$$\mathbb{P}\left(|\mathcal{P}_{\lfloor np\rfloor+i}| \leq \gamma\sqrt{\frac{n}{p}}\right) \leq e^{-\frac{1}{6}\gamma\sqrt{\frac{n}{p}}}.$$

Since the algorithm also requires passing to part two when two columns are equal, we also need the next lemma.

Lemma 3. *Let M be an $n \times n$ random binary matrix with i.i.d. entries m_{ij} such that $\mathbb{P}(m_{ij} = 1) = p$ and $\mathbb{P}(m_{ij} = 0) = 1 - p$. Then, for any $\epsilon > 0$, $\mathbb{P}(M$ has at least two equal rows or columns$) \to 0$ as $n \to \infty$ if $p \geq \frac{(1+\epsilon)\log(n)}{n}$.*

Proof of Theorem 1. There are two cases in which we proceed to the second step of the algorithm: first, when there are at least two identical sub-weight vectors, or second when at least two columns are identical. The probability of the second criterion is shown by Lemma 3 to converge to 0 as $n \to \infty$ for p of the form described, so it suffices to show that the probability of the first criteria occurring also converges to 0 as $n \to \infty$. We call this event $A(n, p)$. Recall from Section 3 that we execute step one of the algorithms by partitioning the columns according to their weight into collections $\mathcal{P} = (\mathcal{P}_1, ..., \mathcal{P}_n)$, and that I_k denotes the indices corresponding to columns in $\mathcal{P}_k$.

For a particular k, let the sub-strings of ρ_1 and ρ_2 that only contain entries with indices in I_k be denoted by $X = (X_1, ..., X_{|\mathcal{P}_k|})$ and $Y = (Y_1, ..., Y_{|\mathcal{P}_k|})$. In order to have $t_{1,k} = t_{2,k}$, we require that $\sum_{i=1}^{|\mathcal{P}_k|} X_i = \sum_{i=1}^{|\mathcal{P}_k|} Y_i$. The sums are equal if and only if

$$|\{i : 1 \leq i \leq n,\ (X_i, Y_i) = (0, 1)\}| = |\{i : 1 \leq i \leq n,\ (X_i, Y_i) = (1, 0)\}|,$$

as an outcome of $(0, 0)$ or $(1, 1)$ does not change the gap between the sum (for shorthand we write $\#(0, 1)$ and $\#(1, 0)$ to denote the two cardinalities).

Since each of the (X_i, Y_i) are pairs of row entries that both lie within columns of weight k, and the 1's are equally likely to be anywhere in each of the columns, we can see that for any $i \in 1, ..., |\mathcal{P}_k|$,

$$\mathbb{P}((X_i, Y_i) = (0,1)) = \mathbb{P}((X_i, Y_i) = (1,0)) = \frac{k(n-k)}{n(n-1)}.$$

Since we assume that $k \in [\lfloor np \rfloor, \lfloor np \rfloor + \lfloor \sqrt{np} \rfloor]$ it holds that there is some $\alpha \in (0,1)$ such that

$$\begin{aligned}\frac{k(n-k)}{n(n-1)} &\sim \frac{(np + \alpha\sqrt{np})(n(1-p) - \alpha\sqrt{np})}{n(n-1)} \\ &= p\left(\frac{n}{n-1}\right)\left(1 + \frac{\alpha}{\sqrt{np}}\right)\left(1 - p - \frac{\alpha p}{\sqrt{np}}\right),\end{aligned}$$

and so $\mathbb{P}((X_i, Y_i) = (0,1)) = \mathbb{P}((X_i, Y_i) = (1,0)) = \Theta(p)$ (note that $np \to \infty$ by the assumptions on p). For each $m \in \{1, ..., n\}$, the conditional probability $\mathbb{P}(t_{1,k} = t_{2,k} | \{|\mathcal{P}_k| = m\})$ is equal to

$$\sum_{i=0}^{\lfloor m/2 \rfloor} \mathbb{P}(\#(0,1) + \#(1,0) = 2i)\mathbb{P}\big(\{\#(0,1) = \#(1,0) = i\} \big| \{\#(0,1) + \#(1,0) = 2i\}\big).$$

Since $(0,1)$ and $(1,0)$ occur with equal probability, when we condition on there being $2i$ of the total, the values $\#(0,1)$ and $\#(0,1)$ follow a $\mathrm{binomial}(2i, 1/2)$ distribution. Define $\tilde{p} := \frac{2k(n-k)}{n(n-1)} = \mathbb{P}((X_i, Y_i) = (0,1) \text{ or } (1,0))$. From here, applying Stirling's approximation we obtain some $\beta > 0$ such that

$$\mathbb{P}(t_{1,k} = t_{2,k} | \{|\mathcal{P}_k| = m\}) = \sum_{i=0}^{\lfloor m/2 \rfloor} \mathbb{P}(\mathrm{binomial}(m, \tilde{p}) = 2i)\mathbb{P}(\mathrm{binomial}(2i, 1/2) = i)$$

$$
\begin{aligned}
&\leq \beta \left(\sum_{i=0}^{\lfloor m/2 \rfloor} \frac{1}{\sqrt{2i \vee 1}} \mathbb{P}(\text{binomial}(m, \tilde{p}) = 2i) \right) \\
&\leq \beta \mathbb{E} \left[\frac{1}{\sqrt{\text{binomial}(m, \tilde{p}) \vee 1}} \right] \\
&\leq \frac{3\beta}{\sqrt{m\tilde{p}}}.
\end{aligned}
$$

See Lemma 6 for a proof of the final inequality. Since we care about the case where $m \geq \gamma\sqrt{\frac{n}{p}}$ and take $n \to \infty$ we can safely assume the inequality holds. Let

$$
S = \left\{ \{(x_1, ..., x_n) : x_i \geq \gamma\sqrt{\frac{n}{p}} \text{ for all } i \in [\lfloor np \rfloor, \lfloor np \rfloor + \lfloor \sqrt{np} \rfloor] \right\},
$$

where $\gamma > 0$ is the one from Lemma 2. When we condition on the sizes of $|\mathcal{P}_k|$ for all $k \in \{1, ..., n\}$, the events $\{t_{1,k} = t_{2,k}\}$ and $\{t_{1,j} = t_{2,j}\}$ are independent for all k and j in $\{1, ..., n\}$. This is because $t_{i,k}$ $(i = 1, 2)$ is simply a $\text{binomial}(|\mathcal{P}_k|, p)$, where each Bernoulli trial is an entry in row i corresponding a column in $\mathcal{P}_k$, and so, each trial is independent of the other rows. This implies that all dependencies between $t_{i,k}$ and $t_{i',j}$ are linked to the values of $|\mathcal{P}_k|$ and $|\mathcal{P}_j|$ for all $i, i' \in \{1, 2\}$ and $j, k \in \{1, ..., n\}$, so conditioning on the sizes results in independence among the before mentioned events because. Since increasing m only decreases the upper bound for $\mathbb{P}(t_{1,k} = t_{2,k} | \{|\mathcal{P}_k| = m\})$,

$$
\sum_{(x_1,...,x_n) \in S} \mathbb{P}\left(T_1 = T_2 \middle| \bigcap_{k=1}^{n} \{|\mathcal{P}_k| = x_k\} \right) \mathbb{P}\left(\bigcap_{k=1}^{n} \{|\mathcal{P}_k| = x_k\} \right)
$$

$$
\begin{aligned}
&\leq \sum_{(x_1,\ldots,x_n)\in S} \left(\frac{3\beta}{\sqrt{\gamma\tilde{p}\sqrt{\frac{n}{p}}}}\right)^{\sqrt{np}} \mathbb{P}\left(\bigcap_{k=1}^{n}\{|\mathcal{P}_k| = x_k\}\right) \\
&= \left(\frac{3\beta}{\sqrt{\gamma\tilde{p}\sqrt{\frac{n}{p}}}}\right)^{\sqrt{np}} \mathbb{P}\Big((|\mathcal{P}_1|,\ldots,|\mathcal{P}_n|) \in S\Big) \\
&\leq \left(\frac{3\beta}{\sqrt{\gamma\tilde{p}\sqrt{\frac{n}{p}}}}\right)^{\sqrt{np}},
\end{aligned}
$$

where T_i is the signature of ρ_i as defined in section 3. On the other hand for S^c, we have that

$$
\begin{aligned}
\sum_{(x_1,\ldots,x_n)\in S^c} \mathbb{P}\left(T_1 = T_2 \middle| \bigcap_{k=1}^{n}\{|\mathcal{P}_k| = x_k\}\right) \mathbb{P}\left(\bigcap_{k=1}^{n}\{|\mathcal{P}_k| = x_k\}\right) \\
\leq \mathbb{P}\Big((|\mathcal{P}_1|,\ldots,|\mathcal{P}_n|) \in S^c\Big),
\end{aligned}
$$

which is a good enough bound because Lemma 2 combined with the union bound ensures that the right side of the inequality is upper bounded by $(\sqrt{np})e^{-\frac{1}{6}\gamma\sqrt{\frac{n}{p}}}$. Putting these two pieces together we get that

$$
\mathbb{P}(A(n,p)) \leq n^2\mathbb{P}(T_1 = T_2) \leq n^2\left(\frac{3\beta}{\sqrt{\gamma\tilde{p}\sqrt{\frac{n}{p}}}}\right)^{\sqrt{np}} + n^2(\sqrt{np})e^{-\frac{1}{6}\gamma\sqrt{\frac{n}{p}}}. \quad (3)
$$

The right term tends to 0 as $n \to \infty$. For the left term, we note that since

$\tilde{p} = \Theta(p)$, we can group up all the constants into some $C > 0$ such that

$$n^2 \left(\frac{3\beta}{\sqrt{\gamma \tilde{p} \sqrt{\frac{n}{p}}}} \right)^{\sqrt{np}} \leq n^2 \left(\frac{C}{(np)^{1/4}} \right)^{\sqrt{np}}$$
$$= \exp\left\{ 2\log(n) + 2\log(C)\sqrt{np} - \frac{1}{4}\sqrt{np}\log(np) \right\},$$

which tends to 0 as $n \to \infty$ whenever $p \geq \frac{16(1+\epsilon)\log^2(n)}{n(\log\log(n))^2}$ for some $\epsilon > 0$.

Now we discuss the time complexity of part two. As mentioned in Section 3, the time complexity of part two is $O(n^2 P)$, where P is the number of valid permutations to check. Hence, it is sufficient to show that $\mathbb{E}[P] = O(1)$ as part one always takes $O(n^2)$ time. The number of permutations we need to check only depends on the sizes of the sets of rows with the same sub-weight vectors and not their positions. Thus, we sum over j representing the number of non-unique sub-weight vectors and then sum over $n_1, n_2, \ldots, n_j$ such that $n_1 + \cdots +, n_j \leq n$, which represent the number of rows that share the same sub-weight vector. We also have the conditions $n_i > 1$ as otherwise this would imply that it is a unique sub-weight vector and $n_i \geq n_{i+1}$ as this avoids double counting. We get the following upper bounds for $\mathbb{E}[P]$:

$$\sum_{j=1}^{n} \sum_{\substack{n_1+n_2+\cdots+n_j \leq n \\ \forall i, n_i > 1 \quad n_i \geq n_{i+1}}} n_1! n_2! \ldots n_j! \binom{n}{n_1, n_2, \ldots, n_j} \prod_{i=1}^{j} \pi_i$$

(where $\pi_i = \mathbb{P}(n_i \text{ rows have same sub-weight vector})$)

$$\leq \sum_{j=1}^{n} \sum_{\substack{n_1+n_2+\cdots+n_j \leq n \\ \forall i, n_i > 1 \quad n_i \geq n_{i+1}}} n_1! n_2! \ldots n_j! \binom{n}{n_1, n_2, \ldots, n_j} \mathbb{P}(T_1 = T_2)^{\sum_{i=1}^{j} n_i - 1}$$

$$\leq \left(1 + \sum_{k=2}^{n} k! \binom{n}{k} \mathbb{P}(T_1 = T_2)^{k-1}\right)^n.$$

The last line, after expanding the product, contains terms which upper bound each term in the previous line upon applying the bound $\binom{n}{n_1, n_2, \ldots, n_j} \leq \binom{n}{n_1} \ldots \binom{n}{n_j}$. Reusing the bound from (3) we get that

$$\mathbb{P}(T_1 = T_2) \leq \left(\frac{C}{(np)^{1/4}}\right)^{\sqrt{np}} + (\sqrt{np}) e^{-\frac{1}{6}\gamma\sqrt{\frac{n}{p}}} \leq n^{-(1+o(1))3\sqrt{1+\epsilon}},$$

when $p \geq \frac{36(1+\epsilon)\log^2(n)}{n(\log\log(n))^2}$. Combining this with the above approximation for $\mathbb{E}[P]$ we see that

$$\begin{aligned}
\mathbb{E}[P] &\leq \left(1 + \sum_{k=2}^{n} k! \binom{n}{k} \left(\frac{1}{n^{(1+o(1)3\sqrt{1+\epsilon}}}\right)^{k-1}\right)^n \\
&\leq \left(1 + \sum_{k=2}^{n} n^k \left(\frac{1}{n^{(1+o(1))2\sqrt{1+\epsilon}}}\right)^{k-1}\right)^n \\
&= \left(1 + n \sum_{k=1}^{n-1} \left(\frac{1}{n^{(1+o(1))2\sqrt{1+\epsilon}}}\right)^{k}\right)^n \\
&\leq \left(1 + n\left(\frac{1}{1 - n^{-(1+o(1))2\sqrt{1+\epsilon}}} - 1\right)\right)^n \\
&\leq \exp\left\{\frac{n^2}{n^{(1+o(1))2\sqrt{1+\epsilon}}}\right\} \\
&\to 1,
\end{aligned}$$

as $n \to \infty$. Hence $\mathbb{E}[P] \to 1$ as $n \to \infty$ as there is always at least one valid permutation (the original ordering before shredding).

□

5 Unique reconstructibility

A common problem of interest in most reconstruction models is that of finding which parameters $p = p(n)$ are such that the reconstructibility of the structure being studied is guaranteed with high probability. Our algorithm gives an upper bound of $\frac{16\log^2(n)}{n(\log\log(n))^2}$ for the critical value at which re-constructibility can be ensured, though with the first-moment method approach we can improve that bound.

Theorem 4. *Let M be an $n \times n$ random binary matrix with i.i.d. entries m_{ij} with $\mathbb{P}(m_{ij} = 1) = p$ and $\mathbb{P}(m_{ij} = 0) = 1 - p$. Then, for any $\epsilon > 0$, $\mathbb{P}(M$ is reconstructible$) \to 1$ as $n \to \infty$ for $p \geq \frac{2(1+\epsilon)\log(n)}{n}$.*

The following lemma offers us a second equivalent definition for constructibility that is better for completing the computations in the proof of Theorem 4

Lemma 5. *Let M be an $n \times n$ binary matrix with shredded column and row collections given by $\gamma_1, ..., \gamma_n$ and $\rho_1, ..., \rho_n$ respectively, and let $M_{\sigma,\tau}$ denote the matrix obtained from permuting the rows by σ and the columns by τ, $M_{\sigma,\tau} = (m_{\sigma(i),\tau(j)})_{i,j=1}^n$ for a particular pair $(\sigma, \tau) \in S_n^2 \setminus \{(\mathrm{Id}, \mathrm{Id})\}$ (here Id just means the identity permutation that sends each $i \in [n]$ to itself). Then,*

$$\{M \text{ is not reconstructible}\} = \bigcup_{\substack{(\sigma,\tau)\in S_n^2 \\ (\sigma,\tau)\neq(\mathrm{Id},\mathrm{Id})}} \{M_{\sigma,\tau} = M\}.$$

Proof of Theorem 4. Define,

$$N = \sum_{(\sigma,\tau)\in(S_n\setminus\{\mathrm{Id}\})^2} \mathbb{1}_{\{M_{\sigma,\tau}=M\}}.$$

A quick computation shows that $\mathbb{E}[N] = (n! - 1)^2 \mathbb{P}(M_{\sigma,\tau} = M)$, where (σ, τ) are independent and both uniform over $S_n \setminus \{\mathrm{Id}\}$. Before bounding this expression, we need some further exploration of the events $\{M_{\sigma,\tau} = M\}$.

We define the permutation graph of a pair $\sigma, \tau \in S_n$ to be the directed graph on $[n]^2 = \{(i,j) : 1 \le i, j \le n\}$ where each vertex (i,j) has an outgoing edge pointing to $(\sigma(i), \tau(j)) = (\sigma_i, \tau_j)$. If $\sigma, \tau \in S_n$ have cyclic decompositions $\sigma = a_1 \cdots a_m$ and $\tau = b_1 \cdots b_k$, a particular pair of cycles a_i and b_j acts on exactly the $|a_i| \times |b_j|$ sub-matrix of M that corresponds to the rows that a_i acts on and the columns that b_j acts on (here $|\cdot|$ denotes the length). In the permutation graph, this $|a_i| \times |b_j|$ sized region corresponds exactly to a collection of $\gcd(|a_i|, |b_j|)$ disjoint cycles, all of length $\mathrm{lcm}(|a_i|, |b_j|)$. To have $M_{\sigma,\tau} = M$, it is necessary to have equality between all entries in M that exist within the same cycle in the permutation graph. That is,

$$\{M_{\sigma,\tau} = M\} \subseteq \bigcap_{i,j=1}^{n} \{m_{i,j} = m_{\sigma^\ell(i),\tau^\ell(j)} \text{ for all } \ell \in \mathbb{N}\}. \tag{4}$$

Since the cycles are disjoint, the events in the intersection are all independent. Using (4) along with our original expression for $\mathbb{E}[N]$ we get that

$$\mathbb{E}[N] \le (n!)^2 \mathbb{E}\left[\prod_{(i,j)\in S} \left(p^{\mathrm{lcm}(|a_i|,|b_j|)} + (1-p)^{\mathrm{lcm}(|a_i|,|b_j|)}\right)^{\gcd(|a_i|,|b_j|)}\right],$$

where $S = \{(i,j) : 1 \le i \le m, 1 \le j \le k, (|a_i|, |b_j|) \ne (1,1)\}$. If we let $c_1(\sigma)$ and $c_1(\tau)$ denote the number of singleton cycles in σ and τ, then we can factor out powers of $(1-p)$ and use the fact that $|a_i| \cdot |b_j| \ge 2$ for $(i,j) \in S$,

$$\mathbb{E}[N] \le (n!)^2 \mathbb{E}\left[(1-p)^{n^2 - c_1(\sigma)c_1(\tau)} \prod_{(i,j)\in S} \left(1 + \left(\frac{p}{1-p}\right)^{\mathrm{lcm}(|a_i|,|b_j|)}\right)^{\gcd(|a_i|,|b_j|)}\right]$$

$$\leq (n!)^2 \mathbb{E}\left[(1-p)^{n^2-c_1(\sigma)c_1(\tau)} \exp\left\{\sum_{(i,j)\in S}\left(\frac{p}{1-p}\right)^2\right\}\right]$$
$$\leq (n!)^2 \mathbb{E}\left[e^{-pn^2+pc_1(\sigma)c_1(\tau)} e^{4(n^2-c_1(\sigma)c_1(\tau))p^2}\right].$$

By bounding the expected value in the final upper bound, one can show that $\mathbb{E}[N] \to 0$ as $n \to \infty$ for

$$\frac{(2+\epsilon)\log(n)}{n} \leq p \leq \frac{17\log^2(n)}{n(\log\log(n))^2},$$

which is sufficient because Theorem 1 covers the case where $p \geq \frac{16(1+\epsilon)\log^2(n)}{n(\log\log(n))^2}$ for any $\epsilon > 0$. Applying Lemma 5 with the union bound we see that

$$\mathbb{P}(M \text{ is reconstructible}) \leq \mathbb{E}[N] + \mathbb{P}\left(\bigcup_{\substack{(\sigma,\tau)\in S_n^2\setminus(\mathrm{Id},\mathrm{Id}) \\ \sigma=\mathrm{Id} \text{ or } \tau=\mathrm{Id}}} \{M_{\sigma,\tau} = M\}\right).$$

However, if one of σ or τ is the identity there must be at least two rows or columns that are identical in M as the other cannot be the identity. Thus by Lemma 3

$$\mathbb{P}\left(\bigcup_{\substack{(\sigma,\tau)\in S_n^2\setminus(\mathrm{Id},\mathrm{Id}) \\ \sigma=\mathrm{Id} \text{ or } \tau=\mathrm{Id}}} \{M_{\sigma,\tau} = M\}\right) \to 0 \text{ as } n \to \infty \text{ for } p \geq \frac{2(1+\epsilon)\log(n)}{n}.$$

Combining this with the above we obtain that $\mathbb{P}(M \text{ is not reconstructible}) \to 0$ as $n \to \infty$ for $p \geq \frac{2(1+\epsilon)\log(n)}{n}$. □

6 Appendix: proofs

Lemma 2. *Suppose that $p = p(n)$ is some sequence such that $np \geq 16$. There exists a positive constant $\gamma > 0$ such that $b_{n,p,\lfloor np \rfloor + i} \geq 2\gamma \frac{1}{\sqrt{np}}$ for all $i \in [0, \lfloor \sqrt{np} \rfloor]$. Furthermore,*

$$\mathbb{P}\left(|\mathcal{P}_{\lfloor np \rfloor + i}| \leq \gamma \sqrt{\frac{n}{p}} \right) \leq e^{-\frac{1}{6}\gamma \sqrt{\frac{n}{p}}},$$

where $\mathcal{P}_i = \{1 \leq j \leq n : |\gamma_j| = i\}$ and $\gamma_1, ..., \gamma_n$ are the shredded columns.

Proof. Since $|\mathcal{P}_k| = \sum_{i=1}^{n} \mathbb{1}_{\{\text{column } i \text{ has weight } k\}}$, and each column has weight k with probability $b_{n,p,k} = \binom{n}{k} p^k (1-p)^{n-k}$, it holds that $|\mathcal{P}_k| \sim \text{binomial}(n, b_{n,p,k})$. By a Chernoff bound we obtain,

$$\mathbb{P}\left(|\mathcal{P}_k| \leq \frac{1}{2} n b_{n,p,k} \right) \leq e^{-\frac{1}{12} n b_{n,p,k}}.$$

From here it suffices to show that there is a constant $\gamma > 0$ such that $\frac{1}{2} n b_{n,p,k} \geq \gamma \sqrt{\frac{n}{p}}$ when $k = \lfloor np \rfloor + i$, $i \in [0, \lfloor \sqrt{np} \rfloor]$. To do this we show the following: for any $0 \leq x \leq \sqrt{np}$ such that $np + x$ is integer-valued, $b_{n,p,np+x} \geq \frac{\alpha}{\sqrt{np}}$ for some $\alpha > 0$. Repeatedly applying Stirling's bounds

$$\sqrt{2\pi n} \left(\frac{n}{e} \right)^n \leq n! \leq e\sqrt{2\pi n} \left(\frac{n}{e} \right)^n,$$

yields

$$b_{n,p,np+x} \geq e^{-2} A_1 A_2 A_3,$$

where

$$A_1 = \frac{1}{\left(1 + \frac{x}{np}\right)^{np} \left(1 - \frac{x}{n(1-p)}\right)^{n(1-p)}},$$

$$A_2 = \left(\frac{1 - \frac{x}{n(1-p)}}{1 + \frac{x}{np}}\right)^x,$$
$$A_3 = \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{(np+x)(n(1-p)-x)}}.$$

Using the fact that $1 + y \leq e^y$ for all $y \in \mathbb{R}$, we get

$$A_1 \geq \frac{1}{e^x e^{-x}} = 1.$$

Next, since $p \geq \frac{16}{n}$,

$$\begin{aligned} A_2 &\geq \left(\left(1 - \frac{x}{n(1-p)}\right)\left(1 - \frac{x}{np}\right)\right)^x \\ &\geq \left(1 - \frac{x}{np(1-p)}\right)^x \geq \left(1 - \frac{2}{\sqrt{np}}\right)^{\sqrt{np}} \\ &\geq \frac{1}{2^4} = \frac{1}{16}. \end{aligned}$$

Finally, using again the fact that $\frac{1}{2} \geq p \geq 16/n$,

$$\begin{aligned} A_3 &\geq \frac{1}{\sqrt{2\pi np(1-p)}} \frac{1}{\sqrt{(1 + \frac{x}{np})(1 - \frac{x}{n(1-p)})}} \\ &\geq \frac{1}{\sqrt{2\pi np(1 + \frac{x}{np(1-p)})}} \\ &\geq \frac{1}{\sqrt{2\pi np(1 + \frac{2}{\sqrt{np}})}} \\ &\geq \frac{1}{\sqrt{3\pi np}}. \end{aligned}$$

Putting everything together we get that

$$b_{n,p,np+x} \geq \left(\frac{1}{16e^2\sqrt{3\pi}}\right)\frac{1}{\sqrt{np}}.$$

□

Lemma 3. *Let M be an $n \times n$ random binary matrix with i.i.d. entries m_{ij} such that $\mathbb{P}(m_{ij} = 1) = p$ and $\mathbb{P}(m_{ij} = 0) = 1 - p$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(M \text{ has at least two equal rows or columns}) \to 0$$

as $n \to \infty$ if $p \geq \frac{(1+\epsilon)\log(n)}{n}$.

Proof. Let $r_1, ..., r_n$ be the rows of M, $A_{i,j} = \{r_i = r_j\}$, and let $N = \sum_{i \neq j} \mathbb{1}_{A_{i,j}}$. Then,

$$\begin{aligned}
\mathbb{E}[N] &= \binom{n}{2}\mathbb{P}(A_{1,2}) = \binom{n}{2}(p^2 + (1-p)^2)^n \\
&= \binom{n}{2}(1-p)^{2n}\left(1 + \frac{p^2}{(1-p)^2}\right)^n \\
&\leq \binom{n}{2}e^{-2np+4np^2} \to 0 \quad \text{as } n \to \infty,
\end{aligned}$$

when $p \geq (1+\epsilon)\frac{\log(n)}{n}$. Since the columns have the same distribution as the rows, the result follows by Markov's inequality. □

Lemma 5. *Let M be an $n \times n$ binary matrix with shredded column and row collections given by $\gamma_1, ..., \gamma_n$ and $\rho_1, ..., \rho_n$ respectively, and let $M_{\sigma,\tau}$ denote the matrix obtained from permuting the rows by σ and the columns by τ, $M_{\sigma,\tau} = (m_{\sigma(i),\tau(j)})_{i,j=1}^n$ for a particular pair $(\sigma, \tau) \in S_n^2 \setminus \{(\mathrm{Id}, \mathrm{Id})\}$ (here Id just means the identity permutation that sends each $i \in [n]$ to itself). Then,*

$$\{M \text{ is not reconstructible}\} = \bigcup_{\substack{(\sigma,\tau)\in S_n^2 \\ (\sigma,\tau)\neq(\mathrm{Id},\mathrm{Id})}} \{M_{\sigma,\tau} = M\}. \tag{5}$$

Proof. Suppose that M is not reconstructible. Then, there exists two distinct

pairs of permutations (σ, τ) and (σ', τ') that satisfy (1). That is, there is some matrix M' (possibly equal to M) such that

$$M = \begin{bmatrix} \rho_{\sigma_1} \\ \vdots \\ \rho_{\sigma_n} \end{bmatrix} = \begin{bmatrix} \gamma_{\tau_1} & \cdots & \gamma_{\tau_n} \end{bmatrix}, \text{ and } M' = \begin{bmatrix} \rho_{\sigma'_1} \\ \vdots \\ \rho_{\sigma'_n} \end{bmatrix} = \begin{bmatrix} \gamma_{\tau'_1} & \cdots & \gamma_{\tau'_n} \end{bmatrix}.$$

Suppose that $r_1, ..., r_n$ and $c_1, ..., c_n$ are the rows and columns in their original, pre-shredding order. Applying $\sigma' \circ \sigma^{-1}$ to $(r_1, ..., r_n)$ and $\tau \circ (\tau')^{-1}$ to $(c_1, ..., c_n)$ must necessarily send M back to itself:

(i) Applying σ^{-1} to the rows of M yields the matrix associated with the shredded rows $R = [\rho_1 \cdots \rho_n]^T$;

(ii) Applying σ' to R gives $[\rho_{\sigma'_1} \cdots \rho_{\sigma'_n}]^T = [\gamma_{\tau'_1} \cdots \gamma_{\tau'_n}] = M'$ by the above identity;

(iii) Applying $(\tau')^{-1}$ to M' brings us to the matrix associated with the shredded columns $C = [\gamma_1 \cdots \gamma_n]$;

(iv) Finally, applying τ to C brings us back to our original matrix $[\gamma_{\tau_1} \cdots \gamma_{\tau_n}] = M$.

Both of these two permutation pairs cannot be the identity because we assume the pairs are distinct, and so the inclusion $\subseteq$ holds in (5).

For the other direction in (5) we suppose we have $(\sigma, \tau) \in S_n^2 \setminus (\mathrm{Id}, \mathrm{Id})$ such that $M_{\sigma,\tau} = M$. Then if we are given an arbitrary shredded ordering $\gamma_1, ..., \gamma_n$ and $\rho_1, ..., \rho_n$, we can by assumption always apply (σ, τ) to the correct ordering to obtain a new non-equal ordering that is valid. Hence, M is not reconstructible. □

Lemma 6. *Suppose* $X \sim \text{binomial}(n, p)$ *and that* $np \geq 16$*. Then,*

$$\mathbb{E}\left[\frac{1}{\sqrt{X \vee 1}}\right] \leq \frac{3}{\sqrt{np}}.$$

Proof. Splitting the expectation into two pieces and then applying the Chebyshev-Cantelli inequality gives us the upper bound

$$\mathbb{E}\left[\frac{1}{\sqrt{X \vee 1}}\right] \leq \sqrt{\frac{2}{np}} + \mathbb{P}\left(X \leq \frac{np}{2}\right) \leq \sqrt{\frac{2}{np}} + \frac{np(1-p)}{np(1-p) + (\frac{np}{2})^2}.$$

Utilizing the fact that $np \geq 16$ we can see that

$$\frac{np(1-p)}{np(1-p) + (\frac{np}{2})^2} \leq \frac{4}{4 + np} \leq \frac{1}{\sqrt{np}},$$

which combined with the above gives

$$\mathbb{E}\left[\frac{1}{\sqrt{X \vee 1}}\right] \leq \frac{\sqrt{2}}{\sqrt{np}} + \frac{1}{\sqrt{np}} \leq \frac{3}{\sqrt{np}}.$$

□

Lemma 7. *Let* σ, τ *be independent uniform permutations over* $S_n \setminus \{\text{Id}\}$*, and let* $c_1(\sigma), c_1(\tau)$ *be the number of singleton cycles in both* σ *and* τ *respectively. Then, for any* $\epsilon > 0$*,*

$$a_n := (n!)^2 \mathbb{E}\left[e^{-pn^2 + pc_1(\sigma)c_1(\tau)} e^{4(n^2 - c_1(\sigma)c_1(\tau))p^2}\right] \to 0$$

as $n \to \infty$ *for*

$$\frac{(2+\epsilon)\log(n)}{n} \leq p \leq \frac{17\log^2(n)}{n(\log\log(n))^2}. \tag{6}$$

Proof. First, we write the expression in the statement of the lemma as

$$a_n = \sum_{0\le x,y\le n-1} (n!)^2 e^{-pn^2+pxy+4(n^2-xy)p^2}\,\mathbb{P}(c_1(\sigma)=x)\,\mathbb{P}(c_1(\tau)=y).$$

We split off the terms with $xy=0$ and upper-bound by

$$\begin{aligned} a_n \le & 2n(n!)^2\exp\left\{-pn^2+4n^2p^2\right\} \\ & + C\sum_{1\le x,y\le n-1}\frac{(n!)^2}{x!y!}\exp\left\{-pn^2+pxy+4(n^2-xy)p^2\right\}, \end{aligned}$$

for some $C>0$ such that $\mathbb{P}(c_1(\sigma)=x)\le\sqrt{C}\frac{1}{x!}$. Such a C exists because $\mathbb{P}(c_1(\sigma')=x)\sim\frac{1}{x!}$ for $\sigma'\in S_n$ uniformly drawn (see Arratia and Tavaré (1992) and Ford (2022) for a discussion of random permutation statistics). One can see immediately that the first term tends to 0 for p in the described range, so we are left with the second term. Relabelling $x=n-k$ and $y=n-\ell$ we can upper bound the sum by

$$\begin{aligned} & C\sum_{1\le k,\ell\le n-1} n^{k+\ell}\exp\left\{-p(n(k+\ell)-k\ell)(1+o(1))\right\} \\ & \qquad \le C\sum_{1\le \ell\le n-1}\left(n\sup_{0\le k\le n} f_\ell(k)\right), \end{aligned}$$

where $f_\ell(k)=e^{-((np-\log(n))(k+\ell)-pk\ell)(1+o(1))}$ with k now being allowed to take on real values. To find $\max_{0\le k\le n} f_\ell(k)$ it suffices to find $\min_{0\le k\le n}((np-\log(n))(k+\ell)-pk\ell) := \min_{0\le k\le n} g_\ell(k)$. Since $g_\ell(k)$ is linear in k it is monotone, and so

$$\min_{0\le k\le n} g_\ell(k) = \min\left\{(np-\log(n))\ell, (n^2p-n\log(n)-\ell\log(n))\right\}$$

$$\geq \min\left\{(1+\epsilon)\log(n)\ell, \epsilon n\log(n)\right\}.$$

The above inequality uses the assumptions on p from (6). Combining this bound with (**??**) gives

$$a_n \leq C\sum_{1\leq\ell\leq n-1} n^{-\epsilon\ell(1+o(1))} + C\sum_{1\leq\ell\leq n-1} n^{-n\epsilon(1+o(1))-1} = o(1).$$

□

7 Future research

As this is the first work exploring this model, there are several future avenues of research.

Suppose that our matrix M, instead of being drawn from a distribution where each entry is 1 with probability p, is drawn from a distribution that is uniform over all matrices with column and row weights equal to d. For which values of d is a reconstruction of M possible?

Suppose that, instead of being given all of $\mathcal{R}(M)$ and $\mathcal{C}(M)$ to use as information for reconstruction, we are only given sub-multisets of the two, where the number of vectors given is a $\text{binomial}(n, q)$ random variable. For what range of q and p is it possible to reconstruct M? This particular question was also posed in Balister, Kronenberg, et al. (2024).

Another natural extension is to consider the case where our entries are no longer binary, but are rather drawn from the set $\{1, ..., k\}$ for some $k \geq 3$ with some distribution $(p_1, ..., p_k)$. Can one find distinct regions of the unit k-simplex for which matrix reconstruction is possible? This is similar to the questions about the reconstruction of k-colourings addressed in Demidovich,

Panichkin, and Zhukovskii (2023). An additional extension that could be borrowed from Demidovich, Panichkin, and Zhukovskii (2023) is to study the case when M is not an $n \times n$ matrix, but instead, a higher dimensional equal-sized array M with n^k elements, with $k > 2$, where every entry is still independently 1 with probability p and 0 with probability $1 - p$. For these cases, one could investigate the shredded reconstruction problem with many different forms of given information. For example, if we denote the i-th row in a $k = 2$ matrix by $(i, *)$. where the wildcard $*$ matches any integer in $\{1, \ldots, n\}$, then the problem dealt with in this paper is when all $(i, *)$ and $(*, j)$ are given. For general k, we can be given between 1 and $k - 1$ wildcards in arbitrary positions. When $k = 3$, for example, we could be given row-like chains like $(i, j, *)$, or slabs like $(*, j, *)$, or combinations of both.

Acknowledgements. The authors thank the anonymous reviewer for several helpful suggestions. The authors also thank Omer Angel for posing the first open problem in section 7.

References

Adhikari, Kartick and Sukrit Chakraborty (2022). *Shotgun assembly of linial-meshulam model*. arXiv: 2209.10942 [math.CO].

Arratia, Richard and Gesine Reinert (1996). "Poisson process approximation for repeats in one sequence and its application to sequencing by hybridization". In: *Combinatorial Pattern Matching (Laguna Beach, CA, 1996)*. Vol. 1075. Lecture Notes in Computer Science. Springer, Berlin, pp. 209–219. ISBN: 3-540-61258-0. DOI: 10.1007/3-540-61258-0_16. URL: https://doi.org/10.1007/3-540-61258-0_16.

Arratia, Richard and Simon Tavaré (1992). “The cycle structure of random permutations”. In: *Ann. Probab.* 20.3, pp. 1567–1591. ISSN: 0091-1798,2168-894X. URL: http://links.jstor.org/sici?sici=0091-1798(199207)20:3%3C1567:TCSORP%3E2.0.CO;2-T&origin=MSN.

Babai, László (1980). “On the complexity of canonical labeling of strongly regular graphs”. In: *SIAM Journal on Computing* 9.1, pp. 212–216. DOI: 10.1137/0209018. eprint: https://doi.org/10.1137/0209018. URL: https://doi.org/10.1137/0209018.

Babai, László, Paul Erdös, and Stanley Selkow (Aug. 1980). “Random graph isomorphism”. In: *SIAM J. Comput.* 9, pp. 628–635. DOI: 10.1137/0209047.

Babai, László and Eugene M. Luks (1983). “Canonical labeling of graphs”. In: *Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing*. STOC ’83. New York, NY, USA: Association for Computing Machinery, pp. 171–183. ISBN: 0897910990. DOI: 10.1145/800061.808746. URL: https://doi.org/10.1145/800061.808746.

Balister, Paul, Béla Bollobás, and Bhargav Narayanan (2019). “Reconstructing random jigsaws”. In: *Multiplex and Multilevel Networks*. Oxford Univ. Press, Oxford, pp. 31–50. ISBN: 978-0-19-880945-6.

Balister, Paul, Gal Kronenberg, Alex Scott, and Youri Tamitegama (2024). *Reconstruction of shredded random matrices*. arXiv: 2401.05058.

Bollobás, Béla (1990). “Almost every graph has reconstruction number three”. In: *J. Graph Theory* 14.1, pp. 1–4. ISSN: 0364-9024,1097-0118. DOI: 10.1002/jgt.3190140102. URL: https://doi.org/10.1002/jgt.3190140102.

Briandais, René de la (1959). “File searching using variable length keys”. In: *1959 Proceedings of the Western Joint Computer Conference*. Vol. 1, pp. 295–298.

Demidovich, Yury, Yaroslav Panichkin, and Maksim Zhukovskii (2023). *Reconstruction of graph colourings*. arXiv: 2308.01671.

Ding, Jian, Yiyang Jiang, and Heng Ma (2022). *Shotgun threshold for sparse Erdős-Rényi graphs*. arXiv: 2208.09876 [math.PR].

Ding, Jian and Haoyu Liu (2022). *Shotgun assembly threshold for lattice labeling model*. arXiv: 2205.01327 [math.PR].

Dyer, Martin, Alan Frieze, and Stephen Suen (1994). "The probability of unique solutions of sequencing by hybridization". In: *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology* 1, pp. 105–10. DOI: 10.1089/cmb.1994.1.105.

Ford, Kevin (2022). "Cycle type of random permutations: a toolkit". In: *Discrete Analysis*. DOI: 10.19086/da.38090.

Fredkin, Edward (1960). "Trie memory". In: *Communications of the ACM* 3.9, pp. 490–499.

Gaudio, Julia and Elchanan Mossel (2022). "Shotgun assembly of Erdős-Rényi random graphs". In: *Electron. Commun. Probab.* 27, Paper No. 5, 14. ISSN: 1083-589X. DOI: 10.1214/22-ecp445. URL: https://doi.org/10.1214/22-ecp445.

Gaudio, Julia, Miklós Z. Rácz, and Anirudh Sridhar (2022). *Local canonical labeling of Erdős–Rényi random graphs*. arXiv: 2211.16454 [math.PR].

Harary, Frank (1974). "A survey of the reconstruction conjecture". In: *Graphs and combinatorics (Proc. Capital Conf., George Washington Univ., Washington, D.C., 1973)*. Vol. 406. Lecture Notes in Math. Springer, Berlin, pp 18–28.

Harary, Frank and Michael Plantholt (1985). "The graph reconstruction number". In: *J. Graph Theory* 9.4, pp. 451–454. ISSN: 0364-9024,1097-0118.

DOI: 10.1002/jgt.3190090403. URL: https://doi.org/10.1002/jgt.3190090403.

Huang, Han and Konstantin Tikhomirov (2022). *Shotgun assembly of unlabeled Erdős-Rényi graphs*. arXiv: 2108.09636 [math.PR].

Johnston, Tom, Gal Kronenberg, Alexander Roberts, and Alex Scott (2023). *Shotgun assembly of random graphs*. arXiv: 2211.14218 [math.CO].

Kelly, Paul J. (1957). "A congruence theorem for trees". In: *Pacific J. Math.* 7, pp. 961–968. ISSN: 0030-8730,1945-5844. URL: http://projecteuclid.org/euclid.pjm/1103043674.

Knuth, Donald E. (1997). *The Art of Computer Programming Volume 3: Sorting and Searching (2nd ed.)* Addison-Wesley.

— (1998). *The Art of Computer Programming. Vol. 3*. Second. Sorting and Searching. Addison-Wesley, Reading, MA, pp. 492+507. ISBN: 0-201-89685-0.

Martinsson, Anders (2016). *Shotgun edge assembly of random jigsaw puzzles*. arXiv: 1605.07151 [math.PR].

— (2019). "A linear threshold for uniqueness of solutions to random jigsaw puzzles". In: *Combinatorics, Probability and Computing* 28.2, pp. 287–302. DOI: 10.1017/s0963548318000391. URL: https://doi.org/10.1017%2Fs0963548318000391.

Morin, Pat (2013). *Open Data Structures: An Introduction*. AU Press, Athabasca University.

Mossel, Elchanan and Nathan Ross (Apr. 2019). "Shotgun sssembly of labeled graphs". In: *IEEE Transactions on Network Science and Engineering* 6.2, pp. 145–157. DOI: 10.1109/tnse.2017.2776913. URL: https://doi.org/10.1109%2Ftnse.2017.2776913.

Mossel, Elchanan and Nike Sun (2015). *Shotgun assembly of random regular graphs*. arXiv: 1512.08473 [math.PR].

Motahari, Abolfazl S., Guy Bresler, and David N. C. Tse (2013). "Information Theory of DNA Shotgun Sequencing". In: *IEEE Transactions on Information Theory* 59.10, pp. 6273–6289. DOI: 10.1109/TIT.2013.2270273.

Müller, Vladimír (1976). "Probabilistic reconstruction from subgraphs". eng. In: *Commentationes Mathematicae Universitatis Carolinae* 017.4, pp. 709–719. URL: http://eudml.org/doc/16787.

Narayanan, Bhargav and Corrine Yap (2023). *Reconstructing random pictures*. arXiv: 2210.09410 [math.CO].

Nenadov, Rajko, Pascal Pfister, and Angelika Steger (2017). "Unique reconstruction threshold for random jigsaw puzzles". In: *Chic. J. Theoret. Comput. Sci.*, Art. 2, 16. ISSN: 1073-0486. DOI: 10.4086/cjtcs.2017.002. URL: https://doi.org/10.4086/cjtcs.2017.002.

Przykucki, Michał, Alexander Roberts, and Alex Scott (2022). "Shotgun reconstruction in the hypercube". In: *Random Structures Algorithms* 60.1, pp. 117–150. ISSN: 1042-9832,1098-2418. DOI: 10.1002/rsa.21028. URL: https://doi.org/10.1002/rsa.21028.

Ramachandran, S. (1981). "On a new digraph reconstruction conjecture". In: *Journal of Combinatorial Theory, Series B* 31.2, pp. 143–149. ISSN: 0095-8956. DOI: https://doi.org/10.1016/S0095-8956(81)80019-6. URL: https://www.sciencedirect.com/science/article/pii/S0095895681800196.

Ramachandran, S. and S. Arumugam (2004). "Graph reconstruction-some new developments". In: *AKCE International Journal of Graphs and Combinatorics* 1.1, pp. 51–61.

Soudry, Daniel, Suraj Keshri, Patrick Stinson, Min-hwan Oh, Garud Iyengar, and Liam Paninski (Oct. 2015). “Efficient "shotgun" inference of neural connectivity from highly sub-sampled activity data”. In: *PLOS Computational Biology* 11, e1004464. DOI: 10.1371/journal.pcbi.1004464.

Ulam, S. M. (1960). *A Collection of Mathematical Problems*. Interscience Publishers, New York-London, pp. xiii+150.