Reconstruction of shredded random matrices

Paul Balister[†], Gal Kronenberg[†*], Alex Scott[†‡], Youri Tamitegama[†]

11 January 2024

Abstract

A matrix is given in "shredded" form if we are presented with the multiset of rows and the multiset of columns, but not told which row is which or which column is which. The matrix is reconstructible if it is uniquely determined by this information. Let M be a random binary $n \times n$ matrix, where each entry independently is 1 with probability $p = p(n) \leqslant \frac{1}{2}$. Atamanchuk, Devroye and Vicenzo introduced the problem and showed that M is reconstructible with high probability for $p \geqslant (2+\varepsilon)\frac{1}{n}\log n$. Here we find that the sharp threshold for reconstructibility is at $p \sim \frac{1}{2n}\log n$.

1 Introduction

Let M be an $n \times n$ matrix with entries all either 0 or 1 and let $\mathcal{R}$ and $\mathcal{C}$ be the collections (multisets) of the n binary strings of length n representing the rows and columns of M respectively. When is it possible to reconstruct M just from the knowledge of $\mathcal{R}$ and $\mathcal{C}$?

The matrix M is *reconstructible* (or *weakly reconstructible*) if M is uniquely determined by the multisets $\mathcal{R}$ and $\mathcal{C}$ of its rows and columns. We say M is *strongly reconstructible* if the positions of all rows and columns are determined by $\mathcal{R}$ and $\mathcal{C}$, that is, for each row $r = (r_1, \dots, r_n) \in \mathcal{R}$ we can determine a unique i such that $M_{i,j} = r_j$ for all j, and similarly for columns. Clearly a weakly reconstructible matrix M is strongly reconstructible if and only if there are no two identical rows and no two identical columns, i.e., $\mathcal{R}$ and $\mathcal{C}$ are actually sets.

We study the threshold for reconstructibility when entries of M are random, independently chosen to be 1 with probability $p \leqslant \frac{1}{2}$. Note that corresponding results for $p \geqslant \frac{1}{2}$ can be obtained by exchanging 0 and 1; so we will always assume $p \leqslant \frac{1}{2}$. In [ADV23], Atamanchuk, Devroye and Vicenzo showed that for any $\varepsilon > 0$, M is strongly reconstructible with high probability whenever $(2+\varepsilon)\frac{1}{n}\log n \leqslant p \leqslant \frac{1}{2}$. (We say that an event E happens *with high probability* (*w.h.p.*) if $\mathbb{P}(E) \to 1$ as $n \to \infty$.)

As mentioned above, a simple obstruction to strong reconstructibility is when two rows or two columns are equal, and $p \sim \frac{1}{n}\log n$ is a sharp threshold for such an event (see

*Supported by the Royal Commission for the Exhibition of 1851.

†Supported by EPSRC grant EP/X013642/1.

†Mathematical Institute, University of Oxford, United Kingdom ({balister,kronenberg,scott,tamitegama}@maths.ox.ac.uk).

Lemma 7). Here we show that this is the main obstacle, locating the threshold for strong reconstructibility at $p \sim \frac{1}{n} \log n$. In fact, we show a stronger statement. We prove that the threshold for (weak) reconstructibility is $p \sim \frac{1}{2n} \log n$, so above this value of p, with high probability, the only likely obstruction to strong reconstructibility is the presence of duplicate rows and/or columns. Moreover, we also identify the main obstacle to weak reconstructibility as a pair of 1s in M, each of which is the only 1 in its row and column. The threshold $p \sim \frac{1}{2n} \log n$ for weak reconstructibility is simply the threshold for the disappearance of these obstacles (see Lemma 8).

Our main result is the following.

Theorem 1. *Suppose that* $p = \frac{1}{2n}(\log n + \log\log n + c_n) \leqslant \frac{1}{2}$.

(a) If $c_n \to \infty$ *as* $n \to \infty$*, then with high probability* M *is reconstructible from the collection of its rows and columns.*

(b) If $c_n \to -\infty$ *as* $n \to \infty$*, and assuming that* M *has at least two* 1*'s, then with high probability* M *is not reconstructible from the collection of its rows and columns.*

(c) If $c_n \to c$ *as* $n \to \infty$*, then the probability that* M *is reconstructible tends to an explicit constant depending on* c *that is strictly between* 0 *and* 1*.*

The proof of Theorem 1 runs in two phases. We begin by knowing the "row values" $\mathcal{R}$ and "column values" $\mathcal{C}$, but not where the rows and columns are placed. In the first phase, we attempt to assign most of the rows and columns to the correct "row position" or "column position" using "local" information. It is helpful here to consider the matrix as the adjacency matrix of a bipartite graph G with bipartition $(\mathcal{I}, \mathcal{J})$, where the vertex class $\mathcal{I}$ corresponds to the indices of the rows of M, the vertex class $\mathcal{J}$ corresponds to the indices of the columns of M, and the edges correspond to the 1s in the matrix. Hence for $i \in \mathcal{I}$, $j \in \mathcal{J}$, ij is an edge if and only if $M_{i,j} = 1$. (We assume $\mathcal{I}$ and $\mathcal{J}$ are disjoint, but both have natural bijections to $[n] := \{1, 2, \ldots, n\}$.) We deduce information about the local structure of G in two different ways:

- For row values $r \in \mathcal{R}$, we deduce the structure of the ball of radius 3 around r by looking at the multiset $\mathcal{R}$ of rows of M, which gives the adjacencies in G between the row values in $\mathcal{R}$ and the column indices (positions) in $\mathcal{J}$.
- For row positions $i \in \mathcal{I}$, we deduce the structure of the ball of radius 3 around the ith row of M by looking at the multiset $\mathcal{C}$ of columns of M, which gives the adjacencies between the row indices in $\mathcal{I}$ and the column values in $\mathcal{C}$.

We complete the first part of the argument by matching up most row values with their row positions and most column values with their column positions. To do this we show that with high probability (for $p > \frac{\delta}{n} \log n$ with any $\delta > 0$), the ball of radius 3 is enough to uniquely identify most vertices of G. Here we use a method similar to Johnston, Kronenberg, Roberts, and Scott [JKRS23] for reconstructing Erdős-Renyi random graphs.

Once we have identified the correct position for most rows and columns, we show in the second phase that we now have enough information to fill in the remaining rows and columns with high probability unless certain substructures occur, and identify the thresholds for these substructures occurring.

In [ADV23], Atamanchuk, Devroye and Vicenzo also proved that for $p \geqslant \frac{(16+\varepsilon)\log^2 n}{n(\log\log n)^2}$ there is an algorithm that succeeds in producing a strong reconstruction of the matrix in $O(n^2)$ time with high probability and in expectation. Our proof gives an algorithm that also produces the matrix in $O(n^2)$ time with high probability above the threshold for weak reconstructibility. See Lemma 18 for more details.

As mentioned above, in light of Lemma 7, we obtain the threshold for strong reconstructibility as a simple corollary of Theorem 1.

Corollary 2. *Suppose that* $p = \frac{1}{n}(\log n + c_n) \leqslant \frac{1}{2}$.

(a) If $c_n \to \infty$ *as* $n \to \infty$, *then with high probability, M is strongly reconstructible from the collection of its rows and columns.*

(b) If $c_n \to -\infty$ *as* $n \to \infty$, *then with high probability, M is not strongly reconstructible from the collection of its rows and columns.*

(c) If $c_n \to c$ *as* $n \to \infty$, *then the probability that M is strongly reconstructible tends to an explicit constant depending on c that is strictly between* 0 *and* 1.

1.1 Discussion and related results

Reconstruction of binary matrices is closely connected to graph reconstruction.

Reconstruction of bipartite graphs. We note that every binary matrix can be viewed as a bipartite graph where one part acts as the rows of the graph, and the other as the columns. There is an edge ij in this graph if there is 1 at the (i, j) entry of the matrix. Thus, reconstruction of a binary matrix by the collection of its rows and columns, can be achieved by the reconstruction of the corresponding balanced bipartite graph from the collection of its 1-balls, where the centred vertex is unlabelled, but the other vertices are labelled. We will use this connection in our proofs. Our main theorem thus also says the following.

Corollary 3. *Suppose that* $p = \frac{1}{2n}(\log n + \log\log n + c_n) \leqslant \frac{1}{2}$ *and let G be a random subgraph of* $K_{n,n}$ *obtained by keeping each edge independently with probability p.*

(a) If $c_n \to \infty$ *as* $n \to \infty$, *then with high probability G is reconstructible from the collection of its* 1*-balls with unlabelled centres.*

(b) If $c_n \to -\infty$ *as* $n \to \infty$, *and assuming that G has at least two edges, then with high probability M is not reconstructible from the collection of its* 1*-balls with unlabelled centres.*

(c) If $c_n \to c$ *as* $n \to \infty$, *then the probability that G is reconstructible from the collection of its* 1*-balls with unlabelled centres tends to an explicit constant depending on c that is strictly between* 0 *and* 1.

Reconstruction of directed graphs. Every $n \times n$ binary matrix is equivalent to a directed graph on n vertices, where a directed edge $\vec{ij}$ appears if and only if there is 1 in the (i, j) entry of the matrix (possibly with self-loops). Thus, reconstruction of a binary matrix by the collection of its rows and columns, is equivalent to the reconstruction of the a directed graph by the collection of its in- and out- neighbourhoods, where the centre

vertex is unlabelled, but the other vertices are labelled. As above, it is straightforward to write down a corollary of Theorem 1 for random directed graphs.

Reconstruction of (random) graphs. There is a huge literature on graph reconstruction, and a growing body of work on reconstructing random graphs and other random combinatorial structures. Kelly and Ulam conjectured in 1941 that every finite simple graph on at least 3 vertices can be determined (up to isomorphism) by its collection of vertex-deleted subgraphs, that is the multiset $\{G \setminus \{v\} : v \in V(G)\}$ of subgraphs of G obtaining by deleting one vertex each time [Kel42, Ula60]. The Reconstruction Conjecture has a long history and was proved in some special cases, but the general statement is still open (see e.g., [BH77, Bon91, AFLM10, LS16] for surveys and background).

Müller [Mül76] proved in 1976 that the Reconstruction Conjecture holds for almost all graphs, in the sense that it holds with high probability for the binomial random graph $G(n, \frac{1}{2})$. Bollobás [Bol90] subsequently showed a much stronger result: in fact, with high probability, the graph can be reconstructed from any three of the subgraphs $G \setminus v$. This led to the understanding that for the reconstruction of random combinatorial objects, much less information is needed.

A significant line of recent research looks at when graphs can be reconstructed from "local" information. Mossel and Ross [MR19] introduced the "shotgun reconstruction" problem. The terminology was motivated by the shotgun assembly problem for DNA sequences, where the goal is to reconstruct a DNA sequence from random local "reads", corresponding to short subsequences (see [DFS94, AMRW96, MBT13] among many references). In the context of graphs, the goal is to reconstruct the graph from balls of small radius. For a graph G and a vertex v of G, let $N_G^r(v)$ be the induced graph of the vertices of distance at most r from v (where only v is labelled). Then G is *r-reconstructible* if every graph with the same multiset of r-balls as G is isomorphic to G. In other words, G can be identified up to isomorphism from the multiset $\{N_G^r(v) : v \in V(G)\}$.

Mossel and Ross [MR19] proved that if $p = \frac{\lambda}{n}$, for $\lambda > 1$, then $r = \Theta(\log n)$ is enough for r-reconstruction of $G(n, p)$ with high probability. Sharp asymptotics was obtained by Ding, Jiang, and Ma in 2021 [DYM22]. Mossel and Ross also looked at the problem of reconstructing from balls of constant radius. They showed that if $np/\log^2 n \to \infty$, then $r = 3$ is enough for $G(n, p)$ with high probability. This was later improved by Gaudio and Mossel [GM22] who also obtained bounds on p for the cases $r = 1$ and 2. The case $r = 1$ was improved by Huang and Tikhomorov [HT21], who showed that there is a phase transition in 1-reconstructibility around $p = n^{-1/2}$, where the upper and lower bounds differ by a polylogarithmic factor. In [JKRS23], this problem was settled for $r \geqslant 3$, and improved bounds were obtained for $r = 1$ and 2.

There is also work on other graph models including random regular graphs [MS15], random geometric graphs [AC22b] and random simplicial complexes [AC22a].

Going back to reconstruction of matrices, one may see this study as the complement of the previous one. Here, the subgraphs given have all vertices labelled except of the centre, while in the Mossel-Ross type of random graph reconstruction the only labelled vertex in each given subgraph is the middle one. As we will see later, the fact that in the matrix case the neighbourhood is labelled, will allow us to obtain information on

larger and larger balls, and by that to apply some of the techniques that were presented in [JKRS23] for r-reconstruction, where $r \geqslant 3$. In this case, however, the notion of r-neighbourhoods is slightly different, as all vertices at odd distance from the root of the ball are labelled but all vertices at even distance are not. But the informative structure of the matrix does allow us to obtain statistics on the vertices of distance at most r of each vertex, which will be sufficient for reconstruction.

1.2 Organisation

The paper is organised as follows. In Section 2 we give classical probabilistic results and thresholds for the substructures that provide obstacles to either weak or strong reconstructibility. In Section 3 we start the proof of Theorem 1 by showing how to reconstruct almost all rows and columns. In Section 4 we show that we can complete this into a full reconstruction unless some specific substructures occur, and complete the proof of Theorem 1. Finally in Section 5 we show that reconstruction can be achieved w.h.p. in $O(n^2)$ time.

2 Preliminaries

In this section we state probabilistic bounds which will be useful later in the paper.

We make frequent use of the following well-known bounds on the tails of the binomial distribution, known as Chernoff bounds (see e.g., [MU17], Theorem 4.4).

Lemma 4. *Let $0 < p \leqslant \frac{1}{2}$, $X \sim \mathrm{Bin}(n,p)$ and $\varepsilon > 0$. Then,*

$$\mathbb{P}\big(X \geqslant (1+\varepsilon)np\big) \leqslant \exp\Big(-\tfrac{\varepsilon^2 np}{2+\varepsilon}\Big),$$
$$\mathbb{P}\big(X \leqslant (1-\varepsilon)np\big) \leqslant \exp\Big(-\tfrac{\varepsilon^2 np}{2}\Big).$$

We will also be interested in tail bounds for binomial distributions where $np \to 0$ as $n \to \infty$, for which we use the following simple observation.

Lemma 5. *Let $X \sim \mathrm{Bin}(n,p)$ and $k \in \mathbb{N}$. Then*

$$\mathbb{P}(X \geqslant k) \leqslant \tbinom{n}{k} p^k \,.$$

Proof. Indeed, $\mathbb{P}(X \geqslant k)$ is at most the expected number of k-tuples of trials that all succeed, which is $\binom{n}{k}p^k$. □

We now quote the following result on convergence to a Poisson distribution which follows directly from [Bol01, Theorem 1.23]. Here we use the *falling factorial* notation $(n)_r := n(n-1)\cdots(n-r+1)$.

Lemma 6. *Let $\lambda_1, \dots, \lambda_k$ be non-negative reals and $X_n^{(i)}$, $i = 1, \dots, k$, be sequences of random variables such that for all k-tuples $(r_1, \dots, r_k)$ of non-negative integers,*

$$\mathbb{E}\big[(X_n^{(1)})_{r_1} \cdots (X_n^{(k)})_{r_k}\big] \to \lambda_1^{r_1} \cdots \lambda_k^{r_k} \qquad \text{as } n \to \infty,$$

Then $(X_n^{(1)}, \dots, X_n^{(k)})$ *converges jointly in distribution to independent Poisson random variables with means* $\lambda_1, \dots, \lambda_k$*. Namely, for any non-negative integers* $s_1, \dots, s_k$*,*

$$\lim_{n\to\infty} \mathbb{P}\big(X_n^{(1)} = s_1, \dots, X_n^{(k)} = s_k\big) = \prod_{i=1}^{k} \frac{e^{-\lambda_i}\lambda_i^{s_i}}{s_i!}\,.$$

Recall that M is an $n \times n$ matrix with i.i.d. Bernoulli random entries that are 1 with probability p. The following lemma gives the probability for the main obstruction to strong reconstructibility, which is the appearance of two identical rows or two identical columns.

Lemma 7. *Assume* $p = \frac{1}{n}(\log n + c_n) \leqslant \frac{1}{2}$*. Then, in the matrix* M*,*

$$\mathbb{P}(\exists \textit{ two equal rows or two equal columns}) \to \begin{cases} 1, & \textit{if } c_n \to -\infty, \\ 1 - \big((1+e^{-c})e^{-e^{-c}}\big)^2, & \textit{if } c_n \to c, \\ 0, & \textit{if } c_n \to \infty. \end{cases}$$

Proof. Let N be the number of pairs of identical rows in M that are *not* entirely zero. The probability that two given rows are identical is $((1-p)^2 + p^2)^n$ and the probability that the two rows are identically zero is $(1-p)^{2n}$. Hence,

$$\begin{aligned} \mathbb{E}[N] &= \binom{n}{2}\Big(\big((1-p)^2+p^2\big)^n - \big((1-p)^2\big)^n\Big) \\ &\leqslant n^2 \cdot np^2 \cdot \big((1-p)^2+p^2\big)^{n-1} \\ &\leqslant n^2 \cdot np^2 \cdot e^{-2(n-1)p(1-p)}, \end{aligned}$$

where we have used comparison with a geometric series (or the Mean Value Theorem) in the second line, and the inequality $1 - x \leqslant e^{-x}$ in the third line. This last expression tends to zero except in the case $p \leqslant (\frac{1}{2} + o(1))\frac{1}{n}\log n$, in which case $c_n \to -\infty$.

Hence it is enough to consider pairs of identically zero rows or columns. As the number of zero rows and/or zero columns is stochastically decreasing as p increases, it is enough to prove the result just in the case when $c_n \to c$ as the other two cases follow from stochastic domination and taking limits $c \to \pm\infty$.

Let X be the number of zero rows and Y the number of zero columns. Then for $r, s \geqslant 0$, $(X)_r(Y)_s$ counts the number of choices of r-tuples of rows and s-tuples of columns, all filled with zero. Such a configuration consists of a union of r rows and s columns with all $rn + sn - rs$ entries equal to zero. There are $(n)_r$ r-tuples of distinct rows and $(n)_s$ s-tuples of distinct columns so, for fixed r and s,

$$\mathbb{E}\big[(X)_r(Y)_s\big] = (n)_r(n)_s(1-p)^{rn+sn-rs} \to e^{-c(r+s)} \qquad \text{as } n \to \infty,$$

as $(n)_r/n^r, (n)_s/n^s, (1-p)^{-rs} \to 1$ and $n(1-p)^n = \exp(\log n - pn + O(p^2 n)) \to e^{-c}$ as $n \to \infty$. Thus by Lemma 6, (X, Y) converges in distribution to i.i.d. $\mathrm{Po}(e^{-c})$ random variables. The probability that there are either two empty rows or two empty columns is $\mathbb{P}(X \geqslant 2 \text{ or } Y \geqslant 2)$, which converges to the expression given in the statement of the lemma. □

Similarly, the following lemma gives the probability of the main obstruction to weak reconstructibility. We say an entry in M is an *isolated* 1 if the entry is a 1 but all other entries in the same row or column are zero. In the graph G this corresponds to an isolated edge, i.e., a component consisting of a single edge. The main obstruction to weak reconstructibility turns out to be the appearance of two or more isolated 1s in the matrix. Indeed, in such a case the isolated 1s must appear in distinct rows and columns, and any permutation of their rows, say, results in a matrix distinct from M, but with identical multisets of rows and columns.

Lemma 8. *Suppose* $p = \frac{1}{2n}(\log n + \log\log n + c_n) \leqslant \frac{1}{2}$. *Let* X *be the number of* 1*s in the matrix* M *and let* Y *be the number of isolated* 1*s in* M. *Then,*

$$\mathbb{P}(Y \geqslant 2 \text{ or } X < 2) \to \begin{cases} 1, & \text{if } c_n \to -\infty, \\ 1 - (1 + e^{-c}/2)e^{-e^{-c}/2}, & \text{if } c_n \to c, \\ 0, & \text{if } c_n \to \infty. \end{cases}$$

Proof. Note that $X \sim \mathrm{Bin}(n^2, p)$, so the $X < 2$ condition is only significant if $n^2 p$ is bounded, i.e., only in the $c_n \to -\infty$ case.

Now $(Y)_r$ counts the number of r-tuples of isolated 1s, all of which must lie in distinct rows and columns. Hence,

$$\mathbb{E}\big[(Y)_r\big] = (n)_r (n)_r p^r (1-p)^{2nr - r^2 - r}.$$

If $c_n \to \infty$ then when $p \leqslant n^{-1/2}$,

$$\mathbb{E}[Y] = n^2 p (1-p)^{2n-2} = n^2 p e^{-2np - O(np^2 + p)} = \frac{np}{\log n} e^{-c_n + O(1)} \to 0,$$

and clearly $\mathbb{E}[Y] = O(n^2 e^{-2np}) \to 0$ for larger p. Hence $\mathbb{P}(Y \geqslant 2) \to 0$ by Markov and, as noted above, $\mathbb{P}(X < 2) \to 0$ as well.

If $c_n \to c$ then,

$$n^2 p (1-p)^{2n} = n^2 p e^{-2np - O(np^2)} = \frac{np}{\log n} e^{-c_n + o(1)} \to e^{-c}/2.$$

As r is fixed and $p \to 0$, we then have $\mathbb{E}[(Y)_r] \to (e^{-c}/2)^r$. Thus Y tends in distribution to a $\mathrm{Po}(e^{-c}/2)$ random variable. Also $\mathbb{P}(X < 2) \to 0$, so $\mathbb{P}(Y \geqslant 2 \text{ or } X < 2)$ converges to the expression given.

If $c_n \to -\infty$ but $n^2 p \to \infty$ then $\mathbb{E}[Y] \to \infty$. However,

$$\frac{\mathbb{E}[Y(Y-1)]}{\mathbb{E}[Y]^2} = \frac{(n-1)^2}{n^2}(1-p)^{-2} \to 1.$$

Thus $\mathrm{Var}[Y] = \mathbb{E}[Y(Y-1)] + \mathbb{E}[Y] - \mathbb{E}[Y]^2 = o(\mathbb{E}[Y]^2)$. Hence by Chebychev's inequality (i.e., the second moment method),

$$\mathbb{P}(Y < 2) \leqslant \mathbb{P}\big(|Y - \mathbb{E}[Y]| > \mathbb{E}[Y] - 2\big) \leqslant \frac{\mathrm{Var}[Y]}{(\mathbb{E}[Y] - 2)^2} \to 0,$$

so $\mathbb{P}(Y \geqslant 2) \to 1$.

Finally we may assume $n^2 p = O(1)$, in which case $Y = X$ w.h.p. as the probability of any row or column containing at least two 1s is $O(n^3 p^2) = o(1)$. Hence in this case $\mathbb{P}(Y \geqslant 2 \text{ or } X < 2) \geqslant \mathbb{P}(X = Y) \to 1$. □

The following lemma will allow us to bound the probability that two multisets of i.i.d. random variables are the same. We shall use the notation $[x_1, \dots, x_d]$ to denote the multiset consisting of the elements $x_1, \dots, x_d$.

Lemma 9. *Let $X_1, \dots, X_d$ be i.i.d. discrete random variables with $\mathbb{P}(X_i = x) \leqslant p_0$ for all x. Then for any multiset $\mathcal{M}$ of d possible values of X_i,*

$$\mathbb{P}\big([X_1, \dots, X_d] = \mathcal{M}\big) \leqslant \frac{(2\pi d + 2)^{1/2}}{(2\pi p_0 d + 1)^{1/(2p_0)}} = O\big(\sqrt{d}(2\pi p_0 d)^{-1/(2p_0)}\big).$$

We note that if $p_0 d \leqslant 1$ then we have a better simple bound of $d! p_0^d$ obtained by summing over all permutations $\sigma \in S_d$ the probability $\mathbb{P}(X_1 = x_{\sigma(1)}, \dots, X_d = x_{\sigma(d)})$, where $\mathcal{M} = [x_1, \dots, x_d]$.

Proof. Let the multiset $\mathcal{M}$ with the highest probability have d_i copies of an element x_i where x_i occurs with probability p_i, $i = 1, 2, \dots$. We use the following version of Stirling's formula which holds for all $d \geqslant 0$,

$$(d/e)^d \sqrt{2\pi d + 1} \leqslant d! \leqslant (d/e)^d \sqrt{2\pi d + 2}.$$

We note that this clearly holds for $d = 0$ (with the interpretation that $0^0 = 1$) and follows easily for $d \geqslant 1$ from the explicit bounds

$$(d/e)^d \sqrt{2\pi d}\, e^{1/(12d+1)} \leqslant d! \leqslant (d/e)^d \sqrt{2\pi d}\, e^{1/12d}$$

proved by Robbins [Rob55]. Thus,

$$\mathbb{P}\big([X_1, \dots, X_d] = \mathcal{M}\big) = \frac{d!}{d_1! \cdots d_n!} p_1^{d_1} \cdots p_n^{d_n} \leqslant \sqrt{2\pi d + 2} \cdot \prod_{i=1}^{n} \Big(\frac{d p_i}{d_i}\Big)^{d_i} \frac{1}{\sqrt{2\pi d_i + 1}}.$$

where, without loss of generality, $d_1, \dots, d_n > 0$ and $d_i = 0$ for $i > n$. Now set $\alpha_i = p_i / p_0$ and $d_i' = d_i / \alpha_i$. We note that $d p_i / d_i = d p_0 / d_i'$ and $2\pi d_i + 1 \geqslant (2\pi d_i' + 1)^{\alpha_i}$ as $0 < \alpha_i \leqslant 1$. Hence we can rewrite the bound as

$$\log \mathbb{P}\big([X_1, \dots, X_d] = \mathcal{M}\big) \leqslant \log \sqrt{2\pi d + 2} + \sum_i \alpha_i \Big(d_i' \log \tfrac{d p_0}{d_i'} - \tfrac{1}{2} \log\big(2\pi d_i' + 1\big)\Big).$$

Now maximise over the d_i', assumed just to be non-negative reals with $\sum \alpha_i d_i' = d$ (and include d_i' with $i > n$ here as well). It is easy to check that $f(x) = x \log \frac{d p_0}{x} - \frac{1}{2} \log(2\pi x + 1)$ is concave, where we set $f(0) = 0$. Indeed $f(x) \to 0$ as $x \to 0^+$ and $f''(x) = -\frac{1}{x} + \frac{2\pi^2}{(2\pi x + 1)^2} = -\frac{(2\pi x - 1)^2 + (4 - \pi)(2\pi x)}{x(2\pi x + 1)^2} < 0$ for all $x > 0$. Hence to maximise

$\sum \alpha_i f(d_i')$ subject to $\sum \alpha_i d_i' = d$ requires taking all d_i' to be equal, say $d_i' = d'$. We do this for all i, even $i > n$. But then $\frac{d}{d'} = \sum \alpha_i = \sum \frac{p_i}{p_0} = \frac{1}{p_0}$, so all $d_i' = d' = p_0 d$. Thus,

$$\log \mathbb{P}\big([X_1, \dots, X_d] = \mathcal{M}\big) \leqslant \log \sqrt{2\pi d + 2} - \tfrac{1}{2p_0} \log\big(2\pi p_0 d + 1\big),$$

as required. □

3 Reconstructing almost all rows and columns

For this section, let M be a binary matrix whose entries are i.i.d. Bernoulli random variables taking value 1 with probability p, where $\frac{\delta}{n} \log n \leqslant p \leqslant \frac{1}{2}$ for some fixed small constant $\delta > 0$.

Recall that the matrix M corresponds to a bipartite graph G with vertex classes $\mathcal{I}$ and $\mathcal{J}$, both of size n, corresponding to the indices of the rows and columns. For $i \in \mathcal{I}$, $j \in \mathcal{J}$, ij is an edge of G if and only if the matrix M has 1 in the entry (i, j).

Given the multisets $\mathcal{R}$ and $\mathcal{C}$, we can reconstruct two graphs, both isomorphic to G. The first is G_R, which is a bipartite graph with vertex classes $\mathcal{R}$ and $\mathcal{J}$ where the row value $r \in \mathcal{R}$, which is a binary vector $r = (r_1, \dots, r_n)$, is joined to all columns $j \in \mathcal{J}$ where $r_j = 1$. The second is G_C, which is a bipartite graph with vertex classes $\mathcal{I}$ and $\mathcal{C}$ where the column value $c = (c_1, \dots, c_n)^T \in \mathcal{C}$ is joined to all rows $i \in \mathcal{I}$ where $c_i = 1$. Clearly G_R and G_C are both isomorphic to G by the correct identification of the row values in $\mathcal{R}$ with their indices in $\mathcal{I}$, and the column values in $\mathcal{C}$ with their indices in $\mathcal{J}$ respectively.

For a vertex v of G (or G_R or G_C), define its *kth degree statistics* inductively as follows:

$$\begin{aligned} \mathcal{D}_0(v) &= \deg(v), \\ \mathcal{D}_{k+1}(v) &= [\mathcal{D}_k(u) \colon u \in N(v)], \text{ for } k > 0. \end{aligned}$$

Note that as G_R and G_C are isomorphic to G, $\mathcal{D}_k(v)$ can be reconstructed from $\mathcal{R}$, $\mathcal{C}$, and *either* the index *or* the value of the row or column v. In particular, we observe that the index of a row value r, say, can be correctly identified if for all rows $r' \neq r$ there is some k such that $\mathcal{D}_k(r) \neq \mathcal{D}_k(r')$.

We first show that for large p, w.h.p. even $\mathcal{D}_1(v)$ is enough to uniquely identify all rows and columns.

Lemma 10. *There exists a $C > 0$ such that for $\frac{C}{n} \log^2 n \leqslant p \leqslant \frac{1}{2}$ and any two distinct rows r and r' of M,*

$$\mathbb{P}\big(\mathcal{D}_1(r) = \mathcal{D}_1(r')\big) = o(n^{-2}).$$

In particular, w.h.p. M is strongly reconstructible.

Proof. Let $r, r' \in \mathcal{R}$. If $\deg(r) \neq \deg(r')$ then clearly $\mathcal{D}_1(r) \neq \mathcal{D}_1(r')$, so we may assume $\deg(r) = \deg(r')$. If $\mathcal{D}_1(r) = \mathcal{D}_1(r')$ then the multiset $\mathcal{M}_1$ of degrees of vertices (columns) in $N(r) \setminus N(r')$ is the same as the multiset $\mathcal{M}_2$ of degrees of vertices

in $N(r') \setminus N(r)$. The multiset of degrees of vertices in $N(r) \cap N(r')$ contributes equally to $\mathcal{D}_1(r)$ and $\mathcal{D}_1(r')$, so we can ignore these.

If we condition on $N(r)$ and $N(r')$ (i.e., the entries in rows r and r') and write $d = |N(r) \setminus N(r')| = |N(r') \setminus N(r)|$ then $\mathcal{M}_1$ and $\mathcal{M}_2$ are i.i.d. random multisets of size d, each of which consists of i.i.d. $\mathrm{Bin}(n-2,p)+1$ random variables. The $+1$ is because we have exactly one of r or r' counted in the degrees. The elements of the multisets and the multisets themselves are independent as they depend only on distinct columns, respectively disjoint rectangles $(\mathcal{I} \setminus \{r,r'\}) \times (N(r) \setminus N(r'))$ and $(\mathcal{I} \setminus \{r,r'\}) \times (N(r') \setminus N(r))$, of unconditioned entries in M.

Now if $X \sim \mathrm{Bin}(n-2,p)$ then $p_0 := \max_x \mathbb{P}(X = x) = \Theta(1/\sqrt{np})$, which we may assume is at most $1/(4\log n)$ if C is large enough. Hence, by Lemma 9,

$$\mathbb{P}(\mathcal{M}_1 = \mathcal{M}_2) = O(\sqrt{d}) \exp\big(-\tfrac{1}{2p_0}\log(2\pi p_0 d)\big) = O\big(n^{1-2\log(2\pi p_0 d)}\big) = o(n^{-2}),$$

provided $p_0 d > 2$, say. But $d \sim \mathrm{Bin}(n, p(1-p))$ stochastically dominates $\mathrm{Bin}(n, p/2)$ and thus by Lemma 4,

$$\mathbb{P}(d < np/4) \leqslant \exp(-np/16) = o(n^{-2}).$$

But $d \geqslant np/4$ implies $p_0 d = \Theta(\sqrt{np}) > 2$ for large n. Hence unconditionally,

$$\mathbb{P}\big(\mathcal{D}_1(r) = \mathcal{D}_1(r')\big) = \mathbb{P}\big(\mathcal{M}_1 = \mathcal{M}_2\big) = o(n^{-2}).$$

The union bound now shows that the probability that there are two rows with the same value of $\mathcal{D}_1(r)$ is $o(1)$, and the same also holds for columns. Since the $\mathcal{D}_1(v)$ can be determined either from the values or the indices of the rows or columns v, w.h.p. each row or column value can be associated with a unique index, and so M is strongly reconstructible. □

For smaller values of p we will need to consider $\mathcal{D}_2(v)$. Before we do so, it will be convenient to prove some results about the typical structure of M when $p = O(\frac{1}{n}(\log n)^2)$.

Lemma 11. *Let $p = O(\frac{1}{n}\log^2 n)$. Then with high probability, every pair of distinct rows r and r' satisfy $|N(r) \cap N(r')| \leqslant 2$.*

Proof. We have $|N(r) \cap N(r')| \sim \mathrm{Bin}(n, p^2)$, so

$$\mathbb{P}\big(|N(r) \cap N(r')| \geqslant 3\big) \leqslant \tbinom{n}{3} p^6 = o(n^{-2}).$$

A union bound now shows that w.h.p. there is no pair of rows r, r' for which $|N(r) \cap N(r')| \geqslant 3$. □

The next fact roughly states that the 2-balls in our bipartite graph are essentially cycle-free. We will use this to argue that the degree distributions of leaves in these balls are essentially independent.

Lemma 12. *For $p = O(\frac{1}{n}\log^2 n)$ w.h.p. there does not exist a pair r, r' of rows for which there are more than two rows $s \neq r, r'$ with at least two neighbours in $N(r) \cup N(r')$.*

Proof. Consider a pair of rows r and r'. If rows $s_1, s_2, s_3 \neq r, r'$ all have at least two neighbours in $N(r) \cup N(r')$, then choose two such neighbours for each of these rows to form a subset $C \subseteq N(r) \cup N(r')$ of columns of size $k := |C| \leqslant 6$ with at least 6 edges from $\{s_1, s_2, s_3\}$ to C. Adding r and r', it follows that there is a set of 5 rows and k columns with $k+6$ entries equal to 1. The expected number of these configurations is $O(n^{k+5}p^{k+6}) = o(1)$. Summing over $k \leqslant 6$, we see that with high probability there is no configuration of this type, and so no choice of r, r' and the s_i. □

A standard Chernoff bound argument shows the following bound on the maximum number of 1s in any row or column.

Lemma 13. *Fix $\delta > 0$. Then there exists a constant $K = K(\delta)$ such that for $p \geqslant \frac{\delta}{n}\log n$, w.h.p., no row or column has degree more than Knp in G.*

Proof. Set $\varepsilon = K - 1$ in Lemma 4 and note that $\frac{\varepsilon^2 np}{2+\varepsilon} \geqslant \frac{(K-1)^2\delta}{K+1}\log n \geqslant 2\log n$ for sufficiently large K. Thus the probability that a fixed row or column has degree more than Knp is at most $e^{-2\log n} = 1/n^2$. The result follows from the union bound over the $2n$ rows and columns. □

Call a row or column *heavy* if it has at least $\frac{1}{2}np$ ones, and *light* otherwise.

Lemma 14. *Fix $\delta > 0$ and assume $p \geqslant \frac{\delta}{n}\log n$. Then w.h.p. there are at most $n^{1-\delta/9} = o(n)$ light rows. Moreover, w.h.p. each row r is adjacent to at most $K' = K'(\delta)$ light columns.*

Proof. By Lemma 4, each row is light with probability at most $e^{-np/8} \leqslant n^{-\delta/8}$, so by Markov there are w.h.p. at most $n^{1-\delta/9}$ light rows.

Now given a column c and a row index $i_0 \in \mathcal{I}$, Lemma 4 gives

$$\mathbb{P}\Big(\sum_{i\neq i_0} c_i < \tfrac{1}{2}np - 1\Big) \leqslant \mathbb{P}\Big(\sum_{i\neq i_0} c_i < \tfrac{1}{2}(n-1)p\Big) \leqslant e^{-(n-1)p/8},$$

as $\sum_{i\neq i_0} c_i \sim \mathrm{Bin}(n-1, p)$. Therefore, given a row r, we have that, conditioned on $N(r)$, the probability that a fixed column $c \in N(r)$ is light is at most $e^{-(n-1)p/8}$. Hence, as distinct columns are independent,

$$\mathbb{P}(\text{There are at least } K' \text{ light columns in } N(r)) \leqslant \binom{|N(r)|}{K'}\big(e^{-(n-1)p/8}\big)^{K'}.$$

Now w.h.p. every row r satisfies $|N(r)| \leqslant Knp$ by Lemma 13, so for $K' := \lceil 9/\delta \rceil$, say, we have

$$\binom{|N(r)|}{K'}\big(e^{-(n-1)p/8}\big)^{K'} \leqslant e^{K'(\log(Knp)-(n-1)p/8)} = o(n^{-1}).$$

Taking a union bound over the rows of M finishes the argument. □

Thus, it is enough to reconstruct only heavy rows and columns in order to reconstruct $(1 - o(1))$ of the matrix.

Lemma 15. *Fix $C, \delta > 0$ and suppose $\frac{\delta}{n} \log n \leqslant p = p(n) \leqslant \frac{C}{n} \log^2 n$. Then w.h.p. for any pair of distinct rows r and r' with r heavy, $\mathcal{D}_2(r) \neq \mathcal{D}_2(r')$. In particular, w.h.p. all heavy row and column values can be matched with their indices.*

Proof. We may assume $\deg(r) = \deg(r')$ as otherwise $\mathcal{D}_2(r) \neq \mathcal{D}_2(r')$. We may also assume the conclusions of Lemmas 11–14 all hold. Reveal all entries in columns of $N(r) \cup N(r')$ and in rows of $N(N(r'))$, so that the full information determining $\mathcal{D}_2(r')$ is revealed. For the second degree statistics to coincide, there needs to be a matching $\sigma\colon N(r) \to N(r')$ such that $\mathcal{D}_1(c) = \mathcal{D}_1(\sigma(c))$ for all columns $c \in N(r)$. We take a union bound over all $d!$ such matchings where $d = |N(r)| = |N(r')|$.

$$\begin{aligned}\mathbb{P}\big(\mathcal{D}_2(r) = \mathcal{D}_2(r')\big) &\leqslant \sum_{\sigma} \mathbb{P}\big(\forall c \in N(r),\ \mathcal{D}_1(c) = \mathcal{D}_1(\sigma(c))\big) \\ &\leqslant d! \cdot \max_{\sigma} \mathbb{P}\big(\forall c \in N(r),\ \mathcal{D}_1(c) = \mathcal{D}_1(\sigma(c))\big).\end{aligned}$$

It now suffices to bound the last probability for any matching σ.

We first show that essentially all degree distributions $\mathcal{D}_1(c)$ are independent from each other and crucially from the information revealed so far. The probability of each $\mathcal{D}_1(c)$ hitting its prescribed degree distribution is bounded by Lemma 9. Taking a product over the valid c yields the desired bound.

Say that a column $c \in N(r)$ is *admissible* if none of the following occurs:

(i) $c \in N(r')$,
(ii) there is a $c' \in N(r) \cup N(r')$ with $c' \neq c$ such that $(N(c') \cap N(c)) \setminus \{r, r'\} \neq \varnothing$,
(iii) $\deg(c) < \frac{1}{2}np$.

By Lemma 11, $N(r)$ and $N(r')$ share at most two elements, so at most two columns c fail because of condition (i). By Lemma 12 there are at most two rows $s \neq r, r'$ whose neighbourhood intersects $N(r) \cup N(r')$ in at least two columns. Since these intersections have size at most 4 each, there are at most 8 columns $c \in N(r) \cup N(r')$ for which there is a $c' \in N(r) \cup N(r')$, $c' \neq c$, with $N(c) \cap N(c')$ containing such a row s. In other words, there are at most 8 columns which fail the condition (ii). Lastly, by Lemma 14, only constantly many $c \in N(r)$ fail condition (iii).

As r is assumed heavy, for large n there are at least $\frac{1}{3}np$ admissible vertices in $N(r)$.

Notice that if a column c is admissible, then, after omitting r, all rows in its neighbourhood $N(c) \setminus \{r\}$ have had exactly one non-zero entry revealed so far. In particular, the degree of each such row only depends on its $N := n - |N(r) \cup N(r')|$ entries yet to be revealed, so there are i.i.d. random variables $X_1, \dots, X_{\deg(c)} \sim \mathrm{Bin}(N, p)$ such that $\mathcal{D}_1(c) = [\deg(r), X_1 + 1, \dots, X_{\deg(c)} + 1]$. Additionally, the second admissibility condition ensures that neighbourhoods of distinct admissible columns are disjoint. Namely, the degrees of the neighbours of distinct admissible columns c, c' depend on disjoint subsets of unconditioned entries and are therefore independent.

Now for any admissible c and choosing any fixed choice of $\mathcal{D}_1(\sigma(c))$ we have that

$$\mathbb{P}(\mathcal{D}_1(c) = \mathcal{D}_1(\sigma(c))) = O(\deg(c)^{1/2})(2\pi p_0 \deg(c))^{-1/p_0},$$

where $p_0 = \Theta(1/\sqrt{Np})$. However, $|N(r) \cup N(r')| \leqslant 2Knp$, so $p_0 = \Theta(1/\sqrt{np})$ and by assumption $\deg(c) \geqslant \frac{1}{2}np$. Hence,

$$\mathbb{P}(\mathcal{D}_1(c) = \mathcal{D}_1(\sigma(c))) \leqslant \exp\big(-\Theta(\sqrt{np}\log(np))\big).$$

Now, if $\mathcal{D}_1(c) = \mathcal{D}_1(\sigma(c))$ holds for every $c \in N(r)$ then certainly it holds for every admissible c, so,

$$\begin{aligned}\mathbb{P}\big(\forall c \in N(r),\ \mathcal{D}_1(c) = \mathcal{D}_1(\sigma(c))\big) &\leqslant \mathbb{P}\big(\forall c \text{ admissible},\ \mathcal{D}_1(c) = \mathcal{D}_1(\sigma(c))\big)\\ &= \prod_{c \text{ admissible}} \mathbb{P}\big(\mathcal{D}_1(c) = \mathcal{D}_1(\sigma(c))\big)\\ &= \exp\big(-\Omega((np)^{3/2}\log(np))\big),\end{aligned}$$

where we used that there are at least $np/3$ admissible columns. Finally, note that $d! \leqslant (Knp)! = \exp\big(O(np\log(np))\big)$, so that

$$\begin{aligned}\mathbb{P}\big(\mathcal{D}_2(r) = \mathcal{D}_2(r')\big) &\leqslant \exp\big(-\Omega((np)^{3/2}\log(np)) + O(np\log(np))\big)\\ &\leqslant \exp\big(-\Omega(\log^{3/2} n)\big) = o(n^{-2}).\end{aligned}$$

Taking a union bound over all choices of r and r' then gives the result. □

4 Full reconstruction

In this section we show that once all but $o(n)$ of the rows and columns have been reconstructed, then with high probability the remaining entries of M can be deduced unless there is some simple obstruction. Recall that we say an entry in M is an isolated 1 if the entry is a 1 but all other entries in the same row or column are zeros.

Lemma 16. *Fix $\varepsilon, C > 0$ and suppose $(\frac{1}{3}+\varepsilon)\frac{1}{n}\log n \leqslant p \leqslant \frac{C}{n}\log^2 n$. Then w.h.p. every row and column of M can be determined except possibly for rows and columns where there is an isolated 1.*

Proof. We can assume, by Lemma 15, that we have already placed all heavy rows and columns in their correct positions. Write $\mathcal{X}$ for the set of rows and $\mathcal{Y}$ for the set of columns that have been placed.

We note that we can place any row r which has a unique neighbourhood $N(r) \cap \mathcal{Y}$ in the already placed columns as we can determine this set from either the value or the index of r. Moreover, we may assume that in the $k \times \ell$ submatrix A formed from the unplaced rows and columns that every row and every column contains a 1. Indeed, if A contained a zero row, corresponding to the row r of M, say, then $N(r) = N(r) \cap \mathcal{Y}$. But for any row r' with $N(r') \cap \mathcal{Y} = N(r) \cap \mathcal{Y}$, either $|N(r')| > |N(r)|$, in which case the rows are distinguished by their degrees (and we can determine the degrees of rows in any position from $\mathcal{C}$), or $N(r') = N(r)$, in which case the row values of r and r' are identical. Thus the position of any row value r is uniquely identified up to permutation of equal rows.

Now suppose the row r is still unplaced but $N(r) \cap \mathcal{Y} \neq \varnothing$. Then there must be another row r' with $|N(r)| = |N(r')|$ and $N(r) \cap \mathcal{Y} = N(r') \cap \mathcal{Y}$, but $N(r) \neq N(r')$. As $|N(r)| = |N(r')|$ there are distinct columns c, c' with $c \in N(r) \setminus N(r')$ and $c' \in N(r') \setminus N(r)$. Clearly $c, c' \notin \mathcal{Y}$ so are also unplaced. We may also choose c and c' so that $N(c) \cap \mathcal{X} = N(c') \cap \mathcal{X}$ as otherwise r and r' could be distinguished by the multisets $[N(c) \cap \mathcal{X} : c \in N(r) \setminus N(r')]$ and $[N(c) \cap \mathcal{X} : c \in N(r') \setminus N(r)]$, both of which can be identified from either the values or positions of r and r'.

We now bound the number of 4-tuples (r, r', c, c') which could satisfy these conditions. More specifically we count the number of 4-tuples (r, r', c, c') satisfying the following slightly weaker conditions.

(i) $c \in N(r) \setminus N(r')$ and $c' \in N(r') \setminus N(r)$.
(ii) $N(r) \cap N(r') \neq \varnothing$.
(iii) $N(r) \cap C = N(r') \cap C$ where $C = \{c'' \neq c, c' : |N(c'') \setminus \{r, r'\}| \geqslant \frac{1}{2}np\}$.
(iv) $N(c) \cap R = N(c') \cap R$ where $R = \{r'' \neq r, r' : |N(r'') \cap C| \geqslant \frac{1}{2}np\}$.

Note that all columns in C and all rows in R are heavy, so assumed already placed. Fixing (r, r', c, c'), we note that by a similar calculation as in Lemma 14, any column $c'' \neq c, c'$ lies in C with probability at least $1 - n^{-1/25}$ independently of (i) and (ii). Indeed, even conditioned on $N(r)$ and $N(r')$, for any $c'' \neq c, c'$,

$$\mathbb{P}(c'' \notin C) = \mathbb{P}\big(\operatorname{Bin}(n-2, p) < \tfrac{1}{2}np\big) \leqslant e^{-(1-o(1))np/8} \leqslant n^{-1/25},$$

independently for each c''. Let E be the event that $|C| < n - 2 - 2n^{24/25} = (1 - o(1))n$. Then as the number of $c'' \neq c, c'$ not in C is stochastically dominated by a $\operatorname{Bin}(n-2, n^{-1/25})$ random variable, by Lemma 4,

$$\mathbb{P}(E) \leqslant e^{-(1/3)(n-2)n^{-1/25}} = n^{-\omega(1)}.$$

Now conditioned on $N(r)$, $N(r')$ and C, and assuming E does not hold, any row $r'' \neq r, r'$ lies in R with probability at least $1 - n^{-1/25}$. Indeed,

$$\mathbb{P}(r'' \notin R) \leqslant \mathbb{P}\big(\operatorname{Bin}(|C|, p) < \tfrac{1}{2}np\big) \leqslant e^{-(1-o(1))np/8} \leqslant n^{-1/25},$$

as an entry being 1 in row r'' is positively correlated with the condition that its column is in C.

Now, given (r, r', c, c'), the probability that (i) holds is $p^2(1-p)^2$. Conditioned on this, (ii) holds with probability at most np^2. Conditioned on (i) and (ii), (iii) holds with probability at most

$$\big((1-p)^2 + p^2 + 2p(1-p)n^{-1/25}\big)^{n-2} = e^{-2pn+o(1)} \leqslant n^{-2/3-2\varepsilon+o(1)}.$$

Conditioned on this the probability that (iv) holds but E does not occur is then at most

$$\big((1-p)^2 + p^2 + 2p(1-p)n^{-1/25}\big)^{n-2} = e^{-2pn+o(1)} \leqslant n^{-2/3-2\varepsilon+o(1)}.$$

Thus the expected number of such 4-tuples is at most

$$n^4 \cdot \big(\mathbb{P}(E) + p^2(1-p)^2 \cdot np^2 \cdot n^{-2/3-2\varepsilon+o(1)} \cdot n^{-2/3-2\varepsilon+o(1)}\big) = n^{-1/3-4\varepsilon+o(1)} = o(1).$$

Hence we may assume $N(r) \cap \mathcal{Y} = \varnothing$ for all unplaced rows r and similarly $N(c) \cap \mathcal{X} = \varnothing$ for all unplaced columns c.

Now consider the graph G restricted to the unplaced rows and columns. The above argument shows that we can assume this forms a union of components in G of total cardinality at most $o(n)$. Isolated vertices correspond to zero rows or zero columns. Isolated edges correspond to isolated 1s in M. Thus is is enough to show that G contains no components with between 3 and $o(n)$ vertices. We count the expected number of such components by counting the number of possible choices of spanning trees for such components. We get that the expected number of these components is then at most

$$\begin{aligned}\sum_{k=3}^{o(n)} \binom{2n}{k} p^{k-1} k^{k-2} (1-p)^{k(n-k)} &\leqslant \sum_{k=3}^{o(n)} \left(2ne \cdot e^{-p(n-k)}\right)^k p^{k-1} \\ &\leqslant \sum_{k=3}^{o(n)} n^{1-(1/3+\varepsilon+o(1))k} = o(1). \qquad \square\end{aligned}$$

Remark 17. For $p < \frac{1}{3n} \log n$ another obstruction to reconstructibility appears, namely pairs of rows (or columns) each of which contains two 1s which themselves are the unique 1 in their column (or row). In graph terms these consist of at least two components that are isomorphic to a path on 3 vertices (with the central vertices both in the same bipartite class). In general more complex tree components on k vertices appear for $p < \frac{1}{kn} \log n$ and if two isomorphic copies of a tree T exist (with the isomorphism respecting the bipartition of G) then the matrix M fails to be reconstructible as we can interchange the vertices in one bipartite class of T with their counterparts in another copy of T without affecting the multisets of rows and columns. This does however affect the matrix M for $k \geqslant 2$. For $k = 1$ pairs of isolated vertices in the same class correspond to pairs of zero rows or zero columns which is the main obstacle to *strong* reconstructibility, but do not prevent reconstructing M as no edges are changed when they are swapped. It should be noted that isolated tree components are not the only obstacle to reconstructibilty. For example, for $p < \frac{1}{4n} \log n$ can have paths on 5 vertices with only the middle vertex possibly joined to other vertices. This corresponds to say two rows r, r' and three columns c, c', c'' with $N(c) = \{r\}$, $N(c') = \{r'\}$, $N(r) = \{c, c''\}$, $N(r') = \{c', c''\}$. In this case r and r' can be swapped giving a different matrix with the same multisets of rows and columns.

Proof of Theorem 1. If $(\frac{1}{3} + \varepsilon)\frac{1}{n} \log n \leqslant p \leqslant \frac{1}{2}$ then by Lemma 10 or Lemma 16 we can w.h.p. reconstruct M up to rows and columns with isolated 1s. If there is at most one isolated 1 then clearly we can reconstruct the whole of M. If there are two or more isolated 1s then we can't reconstruct M as permuting the rows containing these isolated 1s will give a different matrix with the same row and column multisets. Hence the result follows from Lemma 8, with the explicit constant in part *(c)* being as in Lemma 8.

If $p < (\frac{1}{3} + \varepsilon)\frac{1}{n} \log n$ then again by Lemma 8 we have that w.h.p. there are either two isolated 1s or the matrix has fewer than two 1s in total. $\square$

5 A fast algorithm for reconstruction

Lemma 18. *There exists an algorithm that w.h.p. either identifies all rows and columns, or finds a pair of isolated* 1*s and takes* $O(n^2)$ *time.*

Proof. We will restrict our attention to the case when $p = O(\frac{1}{n}\log^2 n)$ as for larger p the result follows from [ADV23].

Constructing a list of all neighbours of every row and column index and every row and column value in G_R and G_C takes $O(n^2)$ time as we have to scan every vector in $\mathcal{R} \cup \mathcal{C}$. As w.h.p. there are only $O(\log^2 n)$ neighbours of any index or value, constructing $\mathcal{D}_2(v)$ for every v then takes only $O(n \operatorname{polylog} n)$ time. Sorting and finding matches between indices and values then again takes $O(n \operatorname{polylog} n)$ time. Identifying any isolated 1s also takes $O(n \operatorname{polylog} n)$ time. So unless $p \geqslant (\frac{1}{3} + \varepsilon)\frac{1}{n}\log n$ by Lemma 8 we will w.h.p. have terminated with either a pair of isolated 1s which demonstrates non-reconstructibility or a reconstructed matrix with at most one non-zero entry. We may thus assume $p \geqslant (\frac{1}{3} + \varepsilon)\frac{1}{n}\log n$ from now on.

The final part of the algorithm relies on calculating the multisets $[N(c) \cap \mathcal{X} : c \in N(r)]$ for rows r and similarly for columns and identifying rows or columns that are uniquely determined. Again, calculating these multisets takes only $O(n \operatorname{polylog} n)$ time. If this fails to identify all rows and columns then as in the proof of Lemma 16 w.h.p. the remaining rows and columns all contain isolated 1s, which we can easily check.

Thus overall the algorithm takes $O(n^2)$ time (with most of the time taken up in the initial scanning of rows and columns to find their neighbours) and correctly identifies M w.h.p. or finds a pair of isolated 1s showing that reconstruction is impossible. □

6 Conclusion

We have shown that there is a sharp threshold for reconstructibility at $p \sim \frac{1}{2n}\log n$, and a sharp threshold for strong reconstructibility at $p \sim \frac{1}{n}\log n$. These results both assume that we know *all* the rows and columns in $\mathcal{R}$ and $\mathcal{C}$. But what if we are missing some of the rows and columns? For example, suppose we are given $\mathcal{R}$, but only a (random) subset of cn elements from $\mathcal{C}$: for what range of c and p can we reconstruct M with high probability?

It would also be very interesting to investigate the problem when there are errors in our data. For example, suppose every entry in each row from $\mathcal{R}$ is given incorrectly with probability q: when can we reconstruct M with high probability? This seems to be interesting even when $p = 1/2$ and q is a small constant.

A similar question arises when we have missing data for both rows and columns: when can we reconstruct almost all of M? And, in a different direction, what happens if the data is corrupted adversarially?

References

[AC22a] KARTICK ADHIKARI and SUKRIT CHAKRABORTY (2022). Shotgun assembly of Linial-Meshulam model. *arXiv preprint arXiv:2209.10942* .

[AC22b] KARTICK ADHIKARI and SUKRIT CHAKRABORTY (2022). Shotgun assembly of random geometric graphs. *arXiv preprint arXiv:2202.02968* .

[ADV23] CAELAN ATAMANCHUK, LUC DEVROYE, and MASSIMO VICENZO (2023). An algorithm to recover shredded random matrices.

[AFLM10] K. J. ASCIAK, M. A. FRANCALANZA, J. LAURI, and W. MYRVOLD (2010). A survey of some open questions in reconstruction numbers. *Ars Combinatoria* **97**, 443–456.

[AMRW96] RICHARD ARRATIA, DANIELA MARTIN, GESINE REINERT, and MICHAEL S WATERMAN (1996). Poisson process approximation for sequence repeats, and sequencing by hybridization. *Journal of Computational Biology* **3**(3), 425–463.

[BH77] J. A. BONDY and R. L. HEMMINGER (1977). Graph reconstruction—a survey. *Journal of Graph Theory* **1**(3), 227–268.

[Bol90] BÉLA BOLLOBÁS (1990). Almost every graph has reconstruction number three. *Journal of Graph Theory* **14**(1), 1–4.

[Bol01] BÉLA BOLLOBÁS (2001). Random graphs, *Cambridge Studies in Advanced Mathematics*, vol. 73. Second edn. (Cambridge University Press, Cambridge).

[Bon91] J. A. BONDY (1991). A graph reconstructor's manual. *Surveys in combinatorics, 1991 (Guildford, 1991), London Math. Soc. Lecture Note Ser.*, vol. 166, 221–252.

[DFS94] MARTIN DYER, ALAN FRIEZE, and STEPHEN SUEN (1994). The probability of unique solutions of sequencing by hybridization. *Journal of Computational Biology* **1**(2), 105–110.

[DYM22] JIAN DING, JIANGYI YANG, and HENG MA (2022). Shotgun threshold for sparse Erdős-Rényi graphs. *arXiv preprint arXiv:2208.09876* .

[GM22] JULIA GAUDIO and ELCHANAN MOSSEL (2022). Shotgun assembly of Erdős-Rényi random graphs. *Electronic Communications in Probability* **27**, Paper No. 5, 14.

[HT21] HAN HUANG and KONSTANTIN TIKHOMIROV (2021). Shotgun assembly of unlabeled Erdős-Rényi graphs. *arXiv preprint arXiv:2108.09636* .

[JKRS23] TOM JOHNSTON, GAL KRONENBERG, ALEXANDER ROBERTS, and ALEX SCOTT (2023). Shotgun assembly of random graphs.

[Kel42] PAUL JOSEPH KELLY (1942). On isometric transformations. Ph.D. thesis, University of Wisconsin.

[LS16] JOSEF LAURI and RAFFAELE SCAPELLATO (2016). Topics in graph automorphisms and reconstruction, *London Mathematical Society Lecture Note Series*, vol. 432. Second edn. (Cambridge University Press, Cambridge).

[MBT13] ABOLFAZL S. MOTAHARI, GUY BRESLER, and DAVID N. C. TSE (2013). Information theory of DNA shotgun sequencing. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory* **59**(10), 6273–6289.

[MR19] ELCHANAN MOSSEL and NATHAN ROSS (2019). Shotgun assembly of labeled graphs. *IEEE Transactions on Network Science and Engineering* **6**(2), 145–157.

[MS15] ELCHANAN MOSSEL and NIKE SUN (2015). Shotgun assembly of random regular graphs. *arXiv preprint arXiv:1512.08473* .

[MU17] MICHAEL MITZENMACHER and ELI UPFAL (2017). Probability and computing. Second edn. (Cambridge University Press, Cambridge). Randomization and probabilistic techniques in algorithms and data analysis.

[Mül76] VLADIMIR MÜLLER (1976). Probabilistic reconstruction from subgraphs. *Commentationes Mathematicae Universitatis Carolinae* **17**(4), 709–719.

[Rob55] HERBERT ROBBINS (1955). A remark on Stirling's formula. *American Mathematical Monthly* **62**(1), 26–29.

[Ula60] S. M. ULAM (1960). A collection of mathematical problems. Interscience Tracts in Pure and Applied Mathematics, no. 8 (Interscience Publishers, New York-London).