Random Intersection Graphs and Missing Data

Dror Salti, Yakir Berchenko
Dept. of Industrial Engineering and Management
Ben-Gurion University of the Negev
Beersheba, Israel
saltidr@post.bgu.ac.il, berchenk@bgu.ac.il

Abstract

Random-graphs and statistical inference with missing data are two separate topics that have been widely explored each in its field. In this paper we demonstrate the relationship between these two different topics and take a novel view of the data matrix as a random intersection graph. We use graph properties and theoretical results from random-graph theory, such as connectivity and the emergence of the giant component, to identify two threshold phenomena in statistical inference with missing data: loss of identifiability and slower convergence of algorithms that are pertinent to statistical inference such as expectation-maximization (EM). We provide two examples corresponding to these threshold phenomena and illustrate the theoretical predictions with simulations that are consistent with our reduction.

1 Introduction

Missing data (or incomplete data) are data in which the values of one or more variables are missing. This constitutes a frequent problem in quantitative research studies and today there are many common techniques for handling incomplete data. These techniques range from the most simple complete case (CC) analysis to more sophisticated and often preferable methods such as Expectation-Maximization (EM) based maximum likelihood estimation (MLE), multiple imputation methods to fill in the missing values, and inverse-probability-weighting (IPW) of the missingness pattern (Little and Rubin 2019; Sun and Tchetgen Tchetgen 2018).

The missing data process is said to be missing completely-at-random (MCAR) if it is independent of both observed and unobserved variables in the full data, and missing at-random (MAR) if, conditional on the observed variables, the process is independent of the unobserved ones (Rubin 1976). A missing data process which is neither MCAR nor MAR is said to be missing-not-at-random (MNAR) (see also the recent and more comprehensive taxonomy (Doretti, Geneletti, and Stanghellini 2018)).

A key element in the missing-data literature is the missingness matrix, R, which is a $0/1$ matrix indicating whether a data-point in the related data-matrix is observed or missing. Since such $0/1$ matrices can be interpreted naturally as an incidence matrix of a bipartite graph, when the missingness pattern is random it can be treated as a random graph; here we point to the relationships between key results and phenomena in random-graph theory and the capacity of statistical analysis with missing data.

In this paper we focus on Missing Completely At Random (MCAR) problems in linear systems. The goal of this work is to present a different view on these problems and to address some theoretical and simulations results with an emphasis on the fraction of missingness in the data-set. In section 2 we review the theory of random intersection graphs, and present the relevant results; in section 3 we cast the missingness matrix as a random intersection graph and then derive some conclusions regarding model identifiability (section 4.1) and convergence rates of statistical algorithms (section 4.2). We focus mainly on MCAR for ease of presentation; however, at the end of section 4.2 (and much more extensively in section 6) we address non-MCAR missingness. Simulations demonstrating these phenomena are presented in section 5.

2 Random Intersection Graphs

The classic random graph model of Erdős and Rényi - $G(n,p)$ considers a fixed set of n vertices and edges that exist with a certain probability $p = p(n)$, independently from each other, (Erdős and Rényi 1960); typically, various features of $G(n,p)$ are studied while $n \to \infty$. Similarly, a *random intersection graph* (RIG) deals with randomly connecting vertices from two different sets (Karoński, Scheinerman, and Singer-Cohen 1999). In its simplest form, the model is defined as follows: given a set V of n vertices and a set A of m auxiliary vertices, construct a bipartite graph $B(n,m,p)$ by letting each edge between vertices $v \in V$ and $a \in A$ exist independently with probability $p = p(n,m)$. The random intersection graph $G(n,m,p)$ with vertex set V is obtained by connecting two vertices $v, w \in V$ if and only if there is a vertex $a \in A$ such that a is linked to both v and w in $B(n,m,p)$. (Karo'nski et al. 1999).

Several key phenomena considered with the classic Erdős-Rényi random graph model $G(n,p)$, are also present

in the random intersection graph model $G(n, m, p)$; in particular, complete connectivity, and the existence of the giant component.

2.1 Connectivity

Recall the following definition (see (Bondy, Murty, and others 1976) for fundamental graph theoretical definitions, and (Bollobás 2001) for random-graph definitions)

Definition 1 (Connected Graph)**.** *A graph $G = (V, E)$ is said to be connected if for every pair of vertices $v, w \in V$ there is a path joining them. The maximal connected subgraphs are called components.*

The connectivity of $G(n, m, p)$ is a threshold phenomenon described by the following result by Singer-Cohen

Theorem 1. *Let $G(n, m, p)$ a random intersection graph with $m = n^\alpha$, $\alpha > 0$ and*

$$p^{**} = \begin{cases} \frac{\ln n + \omega}{m}, & \text{for } \alpha \leq 1 \\ \sqrt{\frac{\ln n + \omega}{nm}}, & \text{for } \alpha > 1 \end{cases}$$

where ω is a slowly varying function of n

(i) If $\omega \to -\infty$, then with high probability $G(n, m, p)$ is disconnected and does not contains a perfect matching.

(ii) If $\omega \to \infty$, then with high probability $G(n, m, p)$ is connected and contain a perfect matching.

See proof and additional details in section 3 of (Rybarczyk 2011), and (Karoński, Scheinerman, and Singer-Cohen 1999).

2.2 The Giant Component

A more subtle and interesting phenomenon than complete connectivity is the existence of the giant component

Definition 2 (Giant Component)**.** *A giant component is a connected component of a given random graph $G = (V, E)$ that contains a finite fraction of the entire graph's vertices - $O(|V|)$ (Bollobás 2001).*

Specifically, the number of vertices in the giant component scales as $O(n)$, and if the number of vertices in the largest component scales only as $o(n)$ then there is no giant component.

There are numerous work and rigorous results on the behavior of the largest component in $G(n, p)$, the classic Erdős-Rényi random graph (Erdős and Rényi 1960; Bollobás 1984; Janson et al. 1993). Molloy and Reed argue that: "Perhaps the most studied phenomenon in the field of random graphs is the behavior of the size of the largest component in $G(n, p)$ when $p = \frac{c}{n}$ and c is near 1". (Molloy and Reed 1998).

In the random intersection graph - $G(n, m, p)$, the component evolution was analyzed by Behrisch for the case where the scaling of vertices and attributes is $m = n^\alpha$. In his work, Behrisch showed that if p is small enough, $\mathcal{N}(G)$ the order (number of vertices) of the largest component is asymptotically almost surely $O(log(n))$. On the other hand, if p is large enough the largest component is actually much larger:

Theorem 2. *Let $G(n, m, p)$ be a random intersection graph with $m = n^\alpha$ and $p^2 m = \frac{c}{n}$. Furthermore let p be the single solution to $p = exp(c(p-1))$ in the interval $(0, 1)$ for $c > 1$. Then we have asymptotically almost surely (a.a.s)*

(i) $\mathcal{N}(G) \leq \frac{9}{(1-c)^2} ln(n)$ *for* $\alpha > 1, c < 1$

(ii) $\mathcal{N}(G) = (1+o(1))(1-p)n$ *for* $\alpha > 1, c > 1$

(iii) $\mathcal{N}(G) \leq \frac{10\sqrt{c}}{(1-c)^2} \sqrt{\frac{n}{m}} ln(m)$ *for* $\alpha < 1, c < 1$

(iv) $\mathcal{N}(G) = (1+o(1))(1-p)\sqrt{cmn}$ *for* $\alpha < 1, c > 1$

See section 4.2 of (Behrisch 2007) for a proof.

Additionally, Britton et al. (Britton et al. 2008) studied the $\alpha = 1$ regime, as well as a Reed-Frost process on the graph (i.e., edge percolation) for complementary results; Berchenko et al. (Berchenko et al. 2009), found the $(1 - C_\Delta)^{-1}$ scaling relation between the critical point in RIGs and that of the (triangle-free) configuration model random graph (Molloy and Reed 1998) (where C_Δ is the so-called clustering coefficient).

Remark 1. As pointed out by Behrisch, note the surprising result in (iv) above, where the size of the largest component is sub-linear compared to n, in contrast to the case for the classic Erdős-Rényi random graph $G(n, p)$, although it is still super-logarithmic.

For our purposes, the above result can be summarized concisely as

Corollary 1. *$p^* = \sqrt{\frac{1}{mn}}$, is the critical value in $G(n, m, p)$ for the existence of large components (asymptotically almost surely).*

which follows immediately from Theorem 2.

3 From Data Matrices to Random Intersection Graphs

For the first stage in connecting RIG to statistical theory and (missingness) data, we now observe the connection between the representation matrix $R(n, m, p)$ of a RIG and the missingness matrix $R_{n \times m}$ of an $n \times m$ partially-observed data-matrix with a fraction p of its entries missing.

The representation matrix $R(n, m, p)$ of a RIG is an alternative view of the graph. This matrix is a $n \times m$ matrix whose rows represent the vertices of $G(n, m, p)$ and whose columns represent the elements of the universal set $A = \{1, ..., m\}$. The entries in $R(n, m, p)$ are 0's and 1's. Each entry is independently 1 with probability p (and 0 with probability $1 - p$). From the random representation matrix $R(n, m, p)$ we derive the graph $G(n, m, p)$ by deeming two vertices to be adjacent if and only if the corresponding rows have a 1 in a common column (Karoński, Scheinerman, and Singer-Cohen 1999). The random representation matrix $R(n, m, p)$ can derive a dual random intersection graph to the one reviewed above - $G(m, n, p)$. The dual graph has m vertices that correspond to the columns in $R(n, m, p)$. Two vertices are adjacent if and only if the corresponding columns have a 1 in a common row.

In the realm of missing data methodology, the most basic setting is the following: n independent m-dimensional vectors are measured, for example from n different subjects;

stacking these vectors creates the data-matrix $X_{n\times m}$, where each column depicts a certain variable measured; finally, however, only a fraction of $1-p$ of the entries of X are observed, with the rest missing (for example, each $x_{i,j}$ is missing independently with probability p). Consider the missingness matrix $R_{n\times m}$ where $r_{i,j}=1$ if $x_{i,j}$ is missing, and 0 otherwise[1]. Clearly, each missingness matrix correspond to an intersection graph (see Figure 1 which exemplifies the reduction we suggest from a data matrix with missing values to a random intersection graph). Note as well that there is a one-to-one correspondence between $R_{n\times m}$ and the support of the RIG's representation $R(n,m,p)$. However, it is worth inquiring how they compare in distribution. The following proposition addresses this issue.

Propositions 1. *Let K denote a binomial random variable, $K\sim B(nm,p)$. If the locations of missingnes are exchangeable (i.e., the $r_{i,j}$'s are exchangeable[2]) and $\sum_{i,j} r_{i,j}\sim K$, then $R_{n\times m}$ and $R(n,m,p)$ are identical in distribution.*

Remark 2. Note that MCAR missingness implies the condition above that "missingness-**locations** are completely at random", while the converse is not true. The conditions of proposition 1 are met even for non-MCAR missingness; see section 6.

Using proposition 1 combined with the results reviewed in section 2 makes it possible to address key issues in statistics and inference.

4 Combinatorial Aspects of Missingness

We begin with two relatively crude example applications, before continuing to two more subtle ones.

4.1 Connectivity and Identification

There is yet another notion of "connectivity" in statistics (Van Buuren 2018).

Definition 3 (Pattern-connectivity)**.** *A missing data pattern $R_{n\times m}$ is said to be connected if any observed data point can be reached from any other observed data point through a sequence of horizontal or vertical moves $r_{i,j}$ to $r_{i,k}$, or $r_{i,j}$ to $r_{l,j}$ (like the rook in chess) where every move in the sequence lands in an observed data point.*

Connected patterns are needed to identify unknown parameters. For example, to be able to estimate a correlation coefficient between two variables, they need to be connected, either directly by a set of cases that have scores for both, or indirectly through their relationship with a third set of connected data. Unconnected patterns often arise in particular data collection designs, such as data combination of different variables and samples or potential outcomes; however, for MCAR data we have

Theorem 3. *Let p^{**} be the critical value for graph-connectivity given by theorem 1. For $p<1-p^{**}$ a.a.s. $R(n,m,p)$ is pattern-connected.*

Proof. Consider $G(n,m,q)$ where $q=1-p$, which is equivalent to the alternative coding of $G(n,m,p)$, with $r_{i,j}=1$ if $x_{i,j}$ is *observed*, and 0 otherwise. When $q>p^{**}$ a.a.s. $G(n,m,q)$ is graph-connected. Note that every pair of vertices, v and w are thus connected by a path which must take a sequence of horizontal or vertical moves on $R(n,m,q)$. Thus, every pair of half-edges, $r_{v,j}=1$ and $r_{w,i}=1$ incident at v and w, are also connected by a rook's path and $R(n,m,q)$ is pattern-connected. □

Seguing now to the linear regression model

$$y_i=\langle x_i,\beta^*\rangle+\varepsilon_i, i=1,2,\ldots,n$$

and $p>p^{**}$ with standard graph-connectivity, the picture is even simpler. There is data missing in every row of $R_{n\times m}$ and thus a complete-case analysis cannot be conducted; furthermore, even sequential imputation via chained equations or EM (Van Buuren et al. 2006; Dempster, Laird, and Rubin 1977) are likely to perform poorly due to the difficulty of finding a good initial starting point; see figure 3 in section 5.

4.2 Large Missingness Components and Statistical Algorithms

Statistical procedures such as EM and inverse-probability-weighting (IPW) (Robins and Gill 1997; Sun and Tchetgen Tchetgen 2018) can perform poorly even for small values of $p<<p^{**}$. When $p<p^{*}$, and very little data is missing, $G(n,m,p)$ is a block-graph (i.e., a clique tree in which every bi-connected component is a clique; *cf.* decomposable graphs in Gaussian graphical models (Buhl 1993), and perfect elimination ordering (Chandran et al. 2003)) without any large cycles. However, when $p^{*}<p<<p^{**}$, even though only a little data is missing, $G(n,m,p)$ contains large components with large cycles which may harm the compatibility of sequential algorithms.

EM Algorithm. The difficulties of computing directly the maximum likelihood estimator (MLE) for statistical models prompted the development of the EM algorithm. A few algorithms of the EM-type were analyzed in early work before Dempster, Laird and Rubin (1977) (Dempster, Laird, and Rubin 1977) introduced the EM algorithm in its modern general form, and many extensions and variants have been suggested since (see section 2 in (Balakrishnan et al. 2017)). Briefly, EM is an iterative procedure that repeatedly cycles through two steps: the E (expectation) step imputes missing values and the M (maximization) step estimates the model's parameters based on the "filled-in" data from the previous E step (See (Allison 2003) for a more thorough overview of the steps of the algorithm).

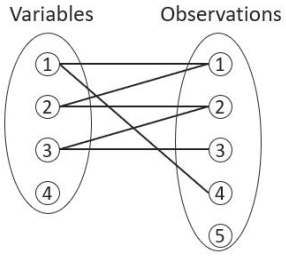
Although this iterative process continues until convergence, a *global* convergence is more difficult to establish (early works (Wu 1983; Dempster, Laird, and Rubin 1977) examined the population-EM operator, under some tacit

[1]The alternative coding, with $r_{i,j}=1$ if $x_{i,j}$ is *observed*, and 0 otherwise, is also common, but we find it less convenient here.

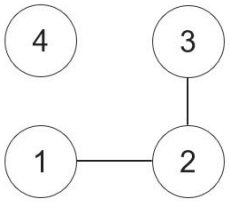
[2]Recall that a sequence of random variables $X_1,X_2,X_3,\ldots$ is said to be *exchangeable* if the joint probability distribution of the sequence does not change when the positions in the sequence in which finitely many of them appear are altered. i.e., permutations of the indices does not change the probability.

	X1	X2	X3	X4
Obs. 1	NA	NA	X13	X14
Obs. 2	X21	NA	NA	X24
Obs. 3	X31	X32	NA	X34
Obs. 4	NA	X42	X43	X44
Obs. 5	X51	X52	X53	X54

(a)



(b)



(c)

Figure 1: A simple example of the derivation of a random intersection graph from a 5×4 design matrix. The steps from (a) a design matrix with missing values, through (b) a random bipartite graph to (c) a random intersection graph.

assumptions, while only recently more nuanced work examined the random sample-based EM operator more rigorously, although in a more limited setting (Balakrishnan et al. 2017; Dwivedi et al. 2018; Xu, Hsu, and Maleki 2016)). Nevertheless, the EM algorithm is widely applied to incomplete data problems, and there is now a very rich literature on its behavior. In practical terms, the EM algorithm has proven itself in effectiveness and correctness (See experimental results in (Allison 2001; Friedman 1998; Graham 2003)).

EM Convergence. Here, in light of previous sections, instead of focusing on the end-point of the EM iterations, we discuss the number of iterations themselves.

Suppose again we observe a response variable $y_i \in \mathbb{R}$ that is linked to a covariate vector $x_i \in \mathbb{R}^m$ via the linear model

$$y_i = \langle x_i, \beta^* \rangle + \varepsilon_i, i = 1, 2, \ldots, n$$

Here, the coefficients vector $\beta^* \in \mathbb{R}^m$ is unknown, and $\varepsilon_i \in \mathbb{R}$ is the observation noise, independent of x_i. In addition, let $\{x_i\}_{i=1}^n$ be independent and identically distributed (i.i.d) from the multivariate Normal distribution. Furthermore, we assume that the missing mechanism is MCAR and each entry in the design matrix X is missing at probability p independently of the other entries.

The reduction presented in proposition 1 and theorems 1-2 suggest the following regarding convergence of the EM algorithm:

- When the missing fraction p in the design matrix X is too big, i.e., $p > p^{**}$, the associated random intersection graph is a.a.s. connected (Theorem 1). Therefore, although the missing mechanism is ignorable (Rubin 1976), the performance of the EM algorithm in these cases will be poor, which means that the estimated parameters produced by the EM algorithm will be far from the real parameters.
- When the missing fraction p in the design matrix is small enough, i.e., $p < \sqrt{\frac{1}{mn}}$, the EM algorithm converges with a small number of iterations, i.e., in linear time (in m). The largest component in the respective random intersection graph is of size $O(\ln m)$ which means that the graph contains only small cycles. Therefore, there exists a consistent and simple order for the imputation (E step of the algorithm).

  However, when $p > \sqrt{\frac{1}{mn}}$, a.a.s. a giant component exists in the respective random intersection graph (corollary 1), so the structure of the graph is complex and there are large cycles which make it difficult for the algorithm in the E step. Therefore, in these cases, there is massive growth in the number of iterations of the EM algorithm before convergence.

  In other words, the missing fraction $p = \sqrt{\frac{1}{mn}}$ defines a threshold phenomenon for the EM algorithm convergence running time.

Remark 3. Note that this is a purely "structural" consideration, that ignores the model parameters altogether. These, and in particular the signal-to-noise ratio $\frac{||\beta||}{\sigma_\varepsilon}$ (where $||\beta||$ is the norm of the coefficients vector of the linear regression model and σ_ε is the variance of the observation noise), were found to play a role as well (Balakrishnan et al. 2017), as indeed we also noticed in our simulations (paradoxically, when the SNR is large enough

$$\sqrt{\frac{\sigma_\varepsilon}{||\beta||}} < p < \sqrt{\frac{1}{mn}}$$

the EM algorithm required more cycles until convergence; data not shown).

Figure 2 provides a graphical representation of the cases above, and figures 3 and 4 present the results of the associated simulations described below.

In addition to EM and sequential imputation, and when the data is MAR (see section 6) we expect similar behavior from IPW methods as well (Sun and Tchetgen Tchetgen 2018), due to the emergence of a non-block-tree with incompatible missingness patterns. However, this is not pursued further here.

5 Simulations

In order to investigate and examine the issues described in the previous section, series of simulation studies were performed.

The data were simulated in the form of a linear regression model:

$$y_i = \langle x_i, \beta^* \rangle + \varepsilon_i, i = 1, 2, \ldots, n$$

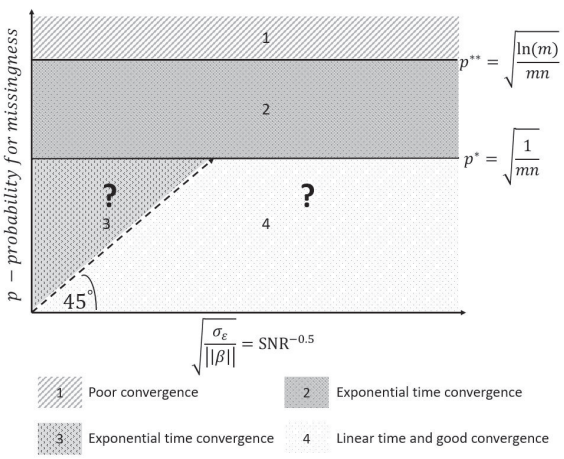


Figure 2: Convergence properties according to p and $\mathrm{SNR}^{-0.5}$.

The first step was to generate the coefficients vector β^*, the data (design) matrix X and the noise vector ε. We chose to generate β^* as a normal random vector, X as a normal random matrix according to the assumptions described in section 4.2 and ε as a normal random vector with a mean equal to 0. According to the linear regression model, the response vector Y was calculated. The next step was to omit some of the values from X to create a missing data set; each entry in the matrix was omitted at probability p independently.

When the data set with missing values was ready, we applied the EM algorithm to produce an estimate $\hat{\beta}$. We used the `norm` package in the `R` program with a sensitive update that counts the number of cycles until convergence and calculated the following measures from the results:

1. $||\beta^* - \hat{\beta}||$ - Distance between the estimated coefficients vector by the EM algorithm and the original coefficients vector.
2. Number of cycles until convergence of the algorithm.

In Figure 3, we plot the results of simulations using

$$n = 100, m = 30, X \sim N(0, I), \varepsilon_i \sim N(0, 1)\ \forall i.$$

We provide a plot of the scaled distance (the error) between the real coefficients vector - β^* and its estimated one - $\hat{\beta}$ versus the probability for missingness, for five different coefficients vectors with different norms $||\beta|| \in \{50, 100, 500, 1000, 5000\}$. For all five choices of coefficients vectors, there was a massive growth in the error around the same value of p, thus agreeing with the first prediction that corresponds to Theorem 1 and 3.

We also verified the results of the second issue empirically. Figure 4 shows the results for a different set of simulations. In this set, the probability for missingness p was constant and the number of coefficients varied. We provide two plots of the number of cycles per coefficient[3] ($\frac{cycles}{m}$) versus (a) the number of coefficients and (b) the normalized (standardized) number of coefficients according to Corollary 1, p^2mn. We used five different probabilities $p \in \{0.02, 0.022, 0.024, 0.026, 0.028\}$ and $n = 100$.

For more simulations results, see Section 7.

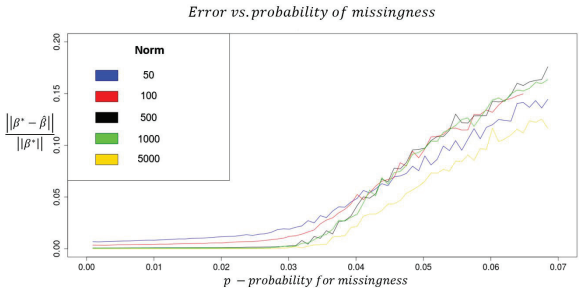


Figure 3: Plot of the scaled distance between the real coefficients vector - β^* and its estimated one - $\hat{\beta}$ versus the probability for missingness - p. Each point represents the average over 100,000 runs. In this simulation $n = 100, m = 30$. As predicted by Theorem 1, $p^{**} = \sqrt{\frac{\ln n}{nm}} = 0.033$ $(\alpha > 1)$ defines the threshold phenomenon.

6 Discussion

The work presented here introduced an interesting and useful connection between random graph theory and statistical applications with missing data. This novel view of a data matrix with missingness as a random intersection graph enables a surprising use of classic combinatorial results, in addition to the more prevalent spectral methodology (Tropp 2015). In particular, as far as we are aware, this is the first time that the giant component emerges naturally in practical algorithmic concerns[4]. Nevertheless, this work leaves much to explore.

First, we focused here on MCAR and did not deal with non-MCAR missingness. However, MCAR missingness is not required, as the following example demonstrates.

Example 1. Let the entries of $X_{n\times m}$ be independent standard normal random variables. Draw K, a binomial random variable, $K \sim B(nm, p)$, and remove the K largest entries of X. Clearly, the locations of missingness are exchangeable, whereas the missingness mechanism is not MCAR.

However, in this paper, for simplicity, we eschewed adding to the various types of missingness and its rich taxonomy (Doretti, Geneletti, and Stanghellini 2018) and aimed to avoid the debate regarding the scientific plausibility of certain missingness patterns (see (Robins and Gill 1997) and the discussion in (Sun and Tchetgen Tchetgen 2018)). Moreover, when focusing on MCAR, our motivation was not

[3]The computational load increases also with the number of coefficients and the size of the data; thus, in order to isolate the increase due to the giant component we normalised by m.

[4]The giant component does indeed play a major role in modeling, in epidemiology and sociology for example, as well as in the analysis of algorithms where the input is a random Erdős-Rényi graph, $G(n, p)$. However, the former does not concern computational issues, whereas in the latter the random graph (and giant component) does not emerge naturally, but rather is "hard coded".

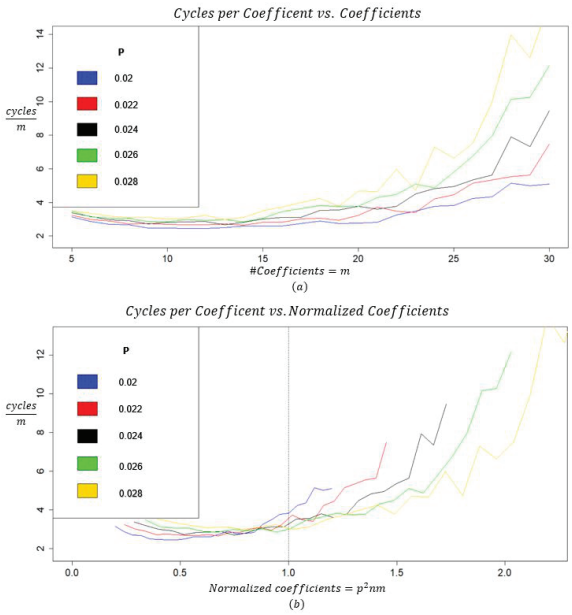


Figure 4: The number of cycles (per coefficient) vs. the number of coefficients and size of the data. Here the number of observations, n, was kept fixed while the number of coefficients, m, was varied $m = \{5, 6, \ldots, 30\}$. (a) Finite size effect interfere with pinpointing where a large increase in the computational load began (top panel); therefore a re-scaled x-axis was used (bottom panel). In (b) the critical point is the same for every set of simulations according to Corollary 1 (the vertical line at 1). Notice the reversal of the ordering of the lines (from largest to smallest) near the critical point $p^2nm = 1$. Each line represents the average over 100,000 runs.

only ease of presentation, but rather also our impression that MCAR missingness is actually sufficiently important on its own.

Second, even though spectral methods are often less refined, the graph-theoretic approach cannot easily address parameter-dependent behavior; in particular, the unique effect of the SNR we observed in our simulations (and noted in (Balakrishnan et al. 2017)). Currently, there is very little we are able to say about the impact of the SNR, and a combined approach might be more fruitful (Drton 2018).

Finally, the threshold phenomena we examined here and particularly the giant component, could be useful to practical applications such as confidentiality and privacy preservation (Henle, Matthews, and Harel 2018). There, it might be the case that the random removal of only a small additional fraction of the data (for privacy preservation) might impose substantial difficulties on the adversary.

7 Appendix

In addition to the set of simulations we presented in Section 5, we also performed a set of simulations with data matrix X as a uniform random matrix, i.e. each entry x_{ij} in the data matrix X is uniformly distributed. The conclusions from the results are presented in Figures 5 and 6 are similar to those presented for the normal case.

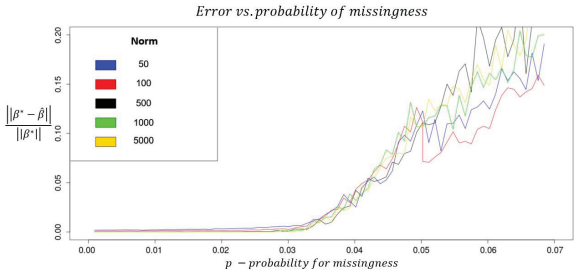


Figure 5: This plot is the same as Figure 3, except the distribution of X. In this case the distribution is uniform, $\forall i, j\ x_{ij} \sim U(0, 50)$. As expected, $p^{**} = \sqrt{\frac{\ln n}{nm}} = 0.033$ defines the threshold phenomenon.

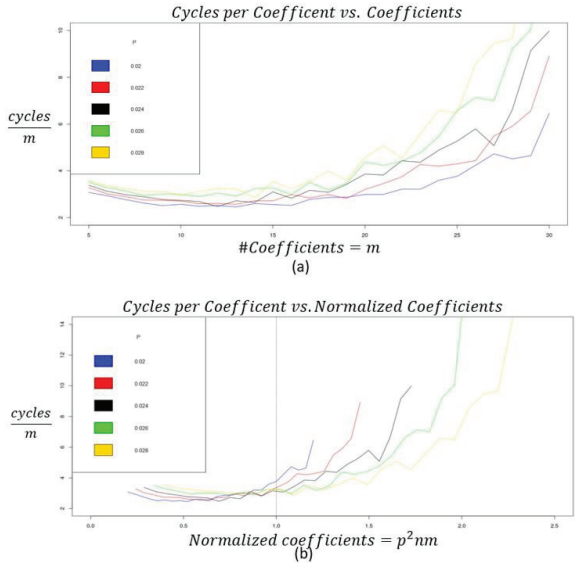


Figure 6: This pair of plots is the same as Figure 4, except the distribution of X. In this case the distribution is uniform, $\forall i, j\ x_{ij} \sim U(0, 50)$.

References

Allison, P. D. 2001. *Missing data*, volume 136. Sage publications.

Allison, P. D. 2003. Missing data techniques for structural equation modeling. *Journal of abnormal psychology* 112(4):545.

Balakrishnan, S.; Wainwright, M. J.; Yu, B.; et al. 2017. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics* 45(1):77–120.

Behrisch, M. 2007. Component evolution in random intersection graphs. *the electronic journal of combinatorics* 14(1):17.

Berchenko, Y.; Artzy-Randrup, Y.; Teicher, M.; and Stone, L. 2009. Emergence and size of the giant component in clustered random graphs with a given degree distribution. *Physical review letters* 102(13):138701.

Bollobás, B. 1984. The evolution of random graphs. *Transactions of the American Mathematical Society* 286(1):257–274.

Bollobás, B. 2001. *Random graphs*. Cambridge university press.

Bondy, J. A.; Murty, U. S. R.; et al. 1976. *Graph theory with applications*, volume 290. Macmillan London.

Britton, T.; Deijfen, M.; Lagerås, A. N.; and Lindholm, M. 2008. Epidemics on random graphs with tunable clustering. *Journal of Applied Probability* 45(3):743–756.

Buhl, S. L. 1993. On the existence of maximum likelihood estimators for graphical gaussian models. *Scandinavian Journal of Statistics* 263–270.

Chandran, L. S.; Ibarra, L.; Ruskey, F.; and Sawada, J. 2003. Generating and characterizing the perfect elimination orderings of a chordal graph. *Theoretical Computer Science* 307(2):303–317.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* 1–38.

Doretti, M.; Geneletti, S.; and Stanghellini, E. 2018. Missing data: a unified taxonomy guided by conditional independence. *International Statistical Review* 86(2):189–204.

Drton, M. 2018. Algebraic problems in structural equation modeling. In *The 50th Anniversary of Gröbner Bases*, 35–86. Mathematical Society of Japan.

Dwivedi, R.; Ho, N.; Khamaru, K.; Jordan, M. I.; Wainwright, M. J.; and Yu, B. 2018. Singularity, misspecification, and the convergence rate of em. *arXiv preprint arXiv:1810.00828*.

Erdős, P., and Rényi, A. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* 5(1):17–60.

Friedman, N. 1998. The bayesian structural em algorithm. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, 129–138. Morgan Kaufmann Publishers Inc.

Graham, J. W. 2003. Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling* 10(1):80–100.

Henle, T.; Matthews, G. J.; and Harel, O. 2018. Data confidentiality. *Methods in Health Services Research* 1–15.

Janson, S.; Knuth, D. E.; Łuczak, T.; and Pittel, B. 1993. The birth of the giant component. *Random Structures & Algorithms* 4(3):233–358.

Karoński, M.; Scheinerman, E. R.; and Singer-Cohen, K. B. 1999. On random intersection graphs: The subgraph problem. *Combinatorics, Probability and Computing* 8(1-2):131–159.

Little, R. J., and Rubin, D. B. 2019. *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Molloy, M., and Reed, B. 1998. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, probability and computing* 7(3):295–305.

Robins, J. M., and Gill, R. D. 1997. Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in medicine* 16(1):39–56.

Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63(3):581–592.

Rybarczyk, K. 2011. Sharp threshold functions for random intersection graphs via a coupling method. *the electronic journal of combinatorics* 18(1):36.

Sun, B., and Tchetgen Tchetgen, E. J. 2018. On inverse probability weighting for nonmonotone missing at random data. *Journal of the American Statistical Association* 113(521):369–379.

Tropp, J. A. 2015. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning* 8(1-2):1–230.

Van Buuren, S.; Brand, J. P.; Groothuis-Oudshoorn, C. G.; and Rubin, D. B. 2006. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation* 76(12):1049–1064.

Van Buuren, S. 2018. *Flexible imputation of missing data*. Chapman and Hall/CRC.

Wu, C. J. 1983. On the convergence properties of the em algorithm. *The Annals of statistics* 11(1):95–103.

Xu, J.; Hsu, D. J.; and Maleki, A. 2016. Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*, 2676–2684.