# Sketching Sparse Covariance Matrices and Graphs

**Gautam Dasarathy**[1], **Pariskhit Shah**[2], **Badri Narayan Bhaskar**[1], **and Robert Nowak**[1]

[1]Department of Electrical and Computer Engineering, University of Wisconsin - Madison
[2]Department of Computer Sciences, University of Wisconsin - Madison

## Abstract

This paper considers the problem of recovering a sparse $p \times p$ matrix X given an $m \times m$ matrix $Y = AXB^T$, where $A$ and $B$ are known $m \times p$ matrices with $m \ll p$. The main result shows that there exist constructions of the "sketching" matrices $A$ and $B$ so that even if $X$ has $\mathcal{O}(p)$ non-zeros, it can be recovered exactly and efficiently using convex optimization as long as these non-zeros are not concentrated in any single row/column of $X$. Furthermore, it suffices for the size of $Y$ (the sketch dimension) to scale as $m = \mathcal{O}\left(\sqrt{\#\text{ nonzeros in } X} \times \log p\right)$. Our approach relies on a novel result concerning tensor products of random bipartite graphs, which may be of independent interest.

We also describe two interesting applications of our results: (a) estimating sparse covariance matrices from compressed realizations and (b) a novel paradigm of lossless sketching/compression of sparse graphs.

## 1 Introduction

An important feature of many modern data analysis problems is the presence of a large number of variables relative to the number of observations. Such high dimensionality occurs in a range of applications in bioinformatics, climate studies, and economics. Accordingly, a fruitful and active research agenda in the recent years has been the development of sampling, estimation and learning methods that take into account *structure* in the underlying model and thereby making these problems tractable. A notion of structure that has gained popularity is that of *sparsity* and estimating sparse signals has been the subject of intense research in the past few years [6, 5, 9]. In this paper, we will study a more nuanced notion of structure that arises naturally in problems in involving sparse matrices; we call this *distributed sparsity*. For what follows, it will be convenient to think of the unknown high-dimensional signal of interest as being represented as a matrix $X$. Roughly speaking, the signal is said to be distributed sparse if every row and every column of $X$ has only a few non-zeros. We will see that it is possible to design efficient and effective acquisition and estimation mechanisms for such signals.

We will begin with two brief examples where distributed sparsity might arise naturally. Our first example is that of **covariance matrices**. Consider natural phenomena where it is reasonable to expect that each covariate is correlated with only a few other covariates. For instance, it is observed that protein signaling networks are such that there are only a few significant correlations [14] and hence the discovery of the such networks from experimental data naturally leads to the estimation of a covariance matrix (where the covariates are proteins) which is (approximately) distributed sparse. Similarly, the covariance structure corresponding to longitudinal data is distributed sparse [8]. As our next example, we will consider **graphs**. Graphs offer a powerful modeling framework to capture many real world

situations where small components interact with and affect each other. It is often natural to assume that each entity under consideration interacts significantly with only a few other entities in the system. The adjacency matrices representing these graphs are distributed sparse and therefore, learning such interaction systems can be thought of as learning bounded degree graphs. Of course, vertices of sparse random graphs (like Erdös- Renyi random graphs $\mathcal{G}(p,q)$ with $pq = \mathcal{O}(\log p)$) as well as those of many real-world graphs tend to have small degrees [4]. We elaborate on these examples in Sections 3.1 and 3.2 respectively and we refer the reader to an extended version of this paper [7] for more examples and a more detailed treatment.

## 2 Problem Setup and Main Results

Our goal is to invert an underdetermined linear system of the form

$$Y = AXB^T, \tag{1}$$

where $A = [a_{ij}] \in \mathbb{R}^{m\times p}, B = [b_{ij}] \in \mathbb{R}^{m\times p}$, with $m \ll p$ and $X \in \mathbb{R}^{p\times p}$[1]. Since the matrix $X \in \mathbb{R}^{p\times p}$ is linearly transformed to obtain the smaller dimensional matrix $Y \in \mathbb{R}^{m\times m}$, we will refer to $Y$ as the *sketch* of $X$ and to the quantity $m$ as the *sketching dimension*. Since the value of $m$ signifies the amount of compression achieved, it is desirable to have as small a value of $m$ as possible. Rewriting the above using tensor product notation, with $y = \text{vec}(Y)$ and $x = \text{vec}(X)$, we equivalently have

$$y = \begin{bmatrix} b_{11}A & b_{12}A & \cdots & b_{1p}A \\ b_{21}A & b_{22}A & \cdots & b_{2p}A \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1}A & b_{m2}A & \cdots & b_{mp}A \end{bmatrix} x = (B \otimes A)x, \tag{2}$$

where $\text{vec}(X)$ simply *vectorizes* the matrix $X$, i.e., produces a long column vector by stacking the columns of the matrix and $B \otimes A$ is the tensor (or Kronecker) product of $B$ and $A$.

While it is not possible to invert such underdetermined systems of equations in general, the rapidly growing literature on what has come to be known as *compressed sensing* [9] suggests that this can be done under certain conditions if $x$ (or equivalently $X$) has only a few non-zeros. As hinted earlier, it will turn out however that one cannot handle arbitrary sparsity patterns and that the non-zero pattern of $X$ needs to be *distributed*, i.e., each row/column of $X$ cannot have more than a few, say $d$, non-zeros. We will call such matrices *$d-$distributed sparse*.

The main part of this paper is devoted to showing that if $X$ is $d-$distributed sparse and if $A$ and $B$ are chosen appropriately, then the following convex optimization program has a unique solution which equals $X$.

$$\underset{\tilde{X}}{\text{minimize}} \ \left\|\tilde{X}\right\|_1 \ \text{subject to} \quad A\tilde{X}B^T = Y. \tag{P$_1$}$$

Here, by $\|\tilde{X}\|_1$ we mean $\sum_{i,j}\left|\tilde{X}_{i,j}\right|$. In particular, we prove the following result.

**Theorem 1.** *Suppose that $X$ is $d-$distributed sparse. Also, suppose that $A, B \in \{0,1\}^{m\times p}$ are drawn independently and uniformly from the $\delta-$random bipartite ensemble*[2]*. Then as long as*

$$m = \mathcal{O}(\sqrt{dp}\log p) \quad \textit{and} \quad \delta = \mathcal{O}(\log p),$$

*there exists a $c > 0$ such that the optimal solution $X^*$ of (P$_1$) equals $X$ with probability exceeding $1 - p^{-c}$. Furthermore, this holds even if $B$ equals $A$.*

Let us pause here and consider some implications of this theorem:

[1] The techniques developed here extend to accomodate "moderately" rectangular matrices. We refer the reader to [7] for more details.

[2] Roughly speaking, the $\delta-$random bipartite ensemble consists of the set of all 0-1 matrices that have almost exactly $\delta$ ones per column. We discuss this briefly in Section 4 and the reader is referred to [7] for more details.

1. Ideally one might try to find the sparsest matrix $\tilde{X}$ that satisfies $Y = A\tilde{X}B^T$. But this is computationally intractable and ($P_1$) is a natural convex relaxation of this objective. However, note that ($P_1$) does not impose any structural restrictions on $X^*$. In other words, even though $X$ is assumed to be distributed sparse, this (highly non-convex) constraint need not be factored in to the optimization problem. This ensures that ($P_1$) is a Linear Program (see e.g., [3]) and can thus be solved efficiently.

2. Even if an oracle were to reveal the exact locations of the non-zeros of $x$, we would require $\mathcal{O}(dp)$ measurements to recover $x$. Comparing this to Theorem 1, we see that the performance of ($P_1$) is near optimal since the number of measurements ($m^2$) is only a logarithm away from this trivial lower bound[3].

3. Finally, inversion of under-determined linear systems where the linear operator assumes a tensor product structure has been studied earlier [12, 10]. However, those methods are relevant only in the regime where the sparsity of the signal to be recovered is much smaller than the dimension $p$. The proof techniques they employ will unfortunately not allow one to handle the more demanding situation the sparsity scales linearly in $p$.

We refer the reader to Section 4 for a sketch of the theoretical ideas behind the proof and to the extended manuscript [7] for the actual proof and intuitions.

Next we will state the following "approximation" result that shows that the solution of ($P_1$) is close to the optimal $d-$ distributed sparse approximation for any matrix $X$. The proof is similar to the proof of Theorem 3 in [2].

Given $p \in \mathbb{N}$, let $[p] \triangleq \{1, 2, \ldots, p\}$ and let $\mathfrak{W}_{d,p}$ denote the set of all $\Omega \subset [p] \times [p]$ such that $\Omega$ could be the support of a $d-$distributed sparse matrix $X \in \mathbb{R}^{p \times p}$. Given $\Omega \in \mathfrak{W}_{d,p}$ and a matrix $X \in \mathbb{R}^{p \times p}$, we write $X_\Omega$ to denote the projection of $X$ onto the set of all matrices supported on $\Omega$. That is, $[X_\Omega]_{i,j} = X_{i,j}$ if $(i,j) \in \Omega$ and 0 otherwise.

**Theorem 2.** *Suppose that $X$ is an arbitrary $p \times p$ matrix and that the hypotheses of Theorem 1 hold. Let $X^*$ denote the solution to the optimization program ($P_1$). Then, there exist constants $C_1, C_2 > 0$ such that the following holds with probability exceeding $1 - p^{-C_1}$.*

$$\|X^* - X\|_1 \leq C_2 \left( \min_{\Omega \in \mathfrak{W}_{d,p}} \|X - X_\Omega\|_1 \right). \tag{3}$$

The above theorem tells us that even if $X$ is not sparse or specially structured, the solution of the optimization program ($P_1$) approximates $X$ as well as the best possible $d-$distributed sparse approximation of $X$ (up to a constant factor). This has interesting implications, for instance, to situations where a $d-$distributed sparse $X$ is corrupted by a "noise" matrix $E$ as shown in the following corollary.

**Corollary 1.** *Suppose $X \in \mathbb{R}^{p \times p}$ is $d-$distributed sparse and suppose that $\widehat{X} = X + E$. Then, there exists a constant $C_3$ such that the solution $X^*$ to the optimization program*

$$\min_{\tilde{X}} \left\| \tilde{X} \right\|_1 \text{ subject to } A\tilde{X}B^T = A\widehat{X}B^T$$

*satisfies $\|X^* - X\|_1 \leq C_3 \|E\|_1$ with high probability.*

We refer the interested reader to the extended manuscript [7] for the proofs of Theorem 2 and Corollary 1.

## 3 Applications

It is instructive at this stage to consider a few examples of the framework we set up in this paper. These applications demonstrate that the modeling assumptions we make viz.,

[3] This logarithmic factor also makes an appearance in the measurement bounds in the compressed sensing literature [5].

tensor product sensing and distributed sparsity are important and arise naturally in a wide variety of contexts. We refer the interested reader to [7] for a more detailed treatment of the applications of this framework.

### 3.1 Covariance Sketching

One particular application that will be of interest to us is the estimation of covariance matrices from sketches of the sample vectors. We call this *covariance sketching*.

Consider a scenario in which the goal is to estimate the covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ of a high-dimensional zero-mean random vector $\xi = (\xi_1, \ldots, \xi_p)^T$. Towards this end, we may obtain $n$ independent realizations of the statistical process $\xi^{(1)}, \ldots, \xi^{(n)} \in \mathbb{R}^p$ and when $p$ is large, it may be infeasible or undesirable to sample and store the entire realizations. Thus we propose an alternative acquisition mechanism: *pool covariates* together to form a collection of new variables $Z_1, \ldots, Z_m$, where typically $m \ll p$. In particular, we may have measurements of the form $Z = A\xi$ where $A \in \{0,1\}^{m\times p}$. The goal therefore is to recover the covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ using only the samples of the $m-$dimensional *sketch vector* $Z^{(i)} = A\xi^{(i)}, i = 1, \ldots, n$, where the $\xi^{(i)} \in \mathbb{R}^p$ are independent random realizations of the underlying process.

Notice that the sample covariance computed using the vectors $\left\{Z^{(i)}\right\}_{i=1}^n$ satisfies $\widehat{\Sigma}_Z^{(n)} = \frac{1}{n}\sum_{i=1}^n Z^{(i)}(Z^{(i)})^T = A\widehat{\Sigma}^{(n)}A^T$, where $\widehat{\Sigma}^{(n)} := \frac{1}{n}\sum_{i=1}^n \xi^{(i)}(\xi^{(i)})^T$ is the (maximum likelihood) estimate of $\Sigma$ from the samples $\xi^{(1)} \ldots, \xi^{(n)}$. The theory developed in this paper tells us that if one supposes that the underlying random vector $\xi$ has the property that each $\xi_i$ depends on only a few (say, $d$) of the other $\xi_j$'s (i.e., the true covariance matrix $\Sigma$ is $d-$distributed sparse), then $\Sigma$ can be recovered from $A\Sigma A^T$ as long as $A \in \{0,1\}^{m\times p}$ is chosen appropriately and $m \geq \mathcal{O}\left(\sqrt{dp}\log p\right)$. This is of course just a stylized version of the above covariance sketching problem.

In fact, observe that $\widehat{\Sigma}^{(n)} := \frac{1}{n}\sum_{i=1}^n \xi^{(i)}(\xi^{(i)})^T$ can be considered to be a "noise corrupted version" of $\Sigma$ where under certain assumptions the noise, $\widehat{\Sigma}^{(n)} - \Sigma$, is such that $\left\|\widehat{\Sigma}^{(n)} - \Sigma\right\|_1 \to 0$ almost surely as $n \to \infty$ by the strong law of large of numbers. Therefore, an application of Theorem 2 tells us that solving ($\mathrm{P}_1$) with the observation matrix $\widehat{\Sigma}_Z^{(n)}$ gives us an asymptotically consistent procedure to estimate the covariance matrix $\Sigma$ from sketched realizations.

We anticipate that these results will be of interest in many areas such as quantitative biology where it maybe possible to naturally pool together covariates and measure interactions at this pool level.

### 3.2 Sketching Sparse Graphs

Large graphs play an important role in many prominent problems of current interest; two such examples are graphs associated to communication networks (such as the internet) and social networks. One is often interested in learning the structure of large graphs, finding motifs like cliques in such graphs or just compression of these graphs. The problem of sketching graphs has recently gained attention in the literature [1, 11].

The framework of our paper suggests a new and natural notion of sketching a given graph $G = (V, E)$. The result is a weighted *sketch* graph $\widehat{G} = (\widehat{V}, \widehat{E})$ whose vertex set is of much smaller cardinality than $V$. To obtain this, one would partition the vertex set $V = V_1 \cup V_2 \cup \ldots \cup V_m$; in the compressed graph $\widehat{G}$, each partition $V_i$ is represented by a node. (We note that this need not necessarily be a disjoint partition.) For each pair $V_i, V_j \in \widehat{V}$, the associated edge weight is the total number of edges crossing from nodes in $V_i$ to the nodes in $V_j$ in the original graph $G$. Note that if an index $k \in V_i \cap V_j$, the self edge $(k, k)$ must be included when counting the total number of edges between $V_i$ and $V_j$. (We point out that the edge $(V_k, V_k) \in \widehat{E}$ also carries a non-zero weight and is precisely equal to the

number of edges in $G$ that have both endpoints in $V_k$.) If one defines an $m \times p$ matrix $A$ such that row $A_i$ is the indicator vector for the set $V_i$, then $Y := AXA^T$ is precisely the adjacency matrix representation of $\widehat{G}$.

Therefore, our results show that when the graph $G$ has $p$ vertices and a maximum degree $d$, then exists a suitable random partitioning scheme such that (a) the proposed method of encoding the graph is lossless and (b) the resulting graph has only $\mathcal{O}(\sqrt{dp}\log p)$ vertices. Moreover, the original graph $G$ can be unravelled from the smaller sketched graph $\widehat{G}$ efficiently using the linear program ($P_1$).

## 4 Theory

We will now provide a brief outline of the theory behind these results and we refer the reader to [7] for complete proofs and further details.

**A. Distributed Sparsity.**
Consider Figure 1 which shows two matrices with $\mathcal{O}(p)$ non-zeros. Suppose that the non-zero pattern in $X$ looks like that of the matrix on the left (which we dub as the "arrow" matrix). It is clear that it is impossible to recover this $X$ from $AXB^T$ *even if* we know the non-zero pattern in advance. For instance, if $v \in \ker(A)$, then the matrix $\tilde{X}$, with $v$ added to the first column of $X$ is such that $AXB^T = A\tilde{X}B^T$ and $\tilde{X}$ is also an arrow matrix and hence indistinguishable from $X$. Similarly, one can "hide" a kernel vector of $B$ in the first row of the arrow matrix. In other words, it is impossible to uniquely recover $X$ from $AXB^T$.

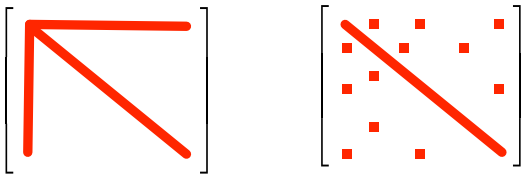


Figure 1: Two matrices with $\mathcal{O}(p)$ non-zeros. The "arrow" matrix is impossible to recover by covariance sketching while the distributed sparse matrix is.

**B. Tensor Products of Random Bipartite Graphs.**
As alluded to earlier, our theory recommends that the sketching matrices be chosen as the adjacency matrices of certain random bipartite graphs. In particular, we pick $A$ and $B$ to be the adjacency matrices of randomly generated graphs $G = ([p], [m], E)$ such that for each $i \in [p]$, we pick $\delta$ neighbors $j_1, \ldots, j_\delta$ uniformly at random from $[m]$ (with replacement[4]).

In past work [2, 13], the authors show that a random graph generated as above is, for suitable values of $\epsilon$ and $\delta$, a $(k, \delta, \epsilon)-$expander. That is, for all sets $S \subset [p]$ such that $|S| \leq k$, the size of the neighborhood $|N(S)|$ is no less than $(1-\epsilon)\delta\,|S|$. If $A$ is the adjacency matrix of such a graph, then it can then be shown that this implies that $\ell_1$ minimization would recover a $k-$sparse vector $x$ if one observes the sketch $Ax$ (actually, [2] shows that these two properties are equivalent). Notice that in our context, the vector that we need to recover is $\mathcal{O}(p)$ sparse and therefore, our random graph needs to be a $(\mathcal{O}(p), \delta, \epsilon)-$expander. Unfortunately, this turns out to not be true of the tensor product graph $G_1 \otimes G_2$[5] when $G_1$ and $G_2$ are randomly chosen as directed above.

However, we show that if $G_1$ and $G_2$ are picked as above, then the graph $G_1 \otimes G_2$ satisfies what can be considered a *weak distributed expansion* property. This roughly says that the

[4] While it is possible to work with sampling without replacement, we chose this to aid in clear presentation of the ideas

[5] We use $G_1 \otimes G_2$ to denote the graph whose adjacency matrix corresponds to the tensor product of the adjacency matrices of $G_1$ and $G_2$

neighborhood of a $d-$distributed $\Omega \subset [p] \times [p]$ is large enough. Moreover, we show that this is in fact sufficient to prove that with high probability $X$ can be recovered from $AXB^T$ efficiently. This combinatorial result is a key technical contribution of our work and we expect that it might be of independent interest.

## 5 Conclusions

In this paper we have introduced the notion of distributed sparsity for matrices. We have shown that when a matrix is $X$ distributed sparse, and $A, B$ are suitable random binary matrices, then it is possible to recover $X$ from under-determined linear measurements of the form $Y = AXB^T$ via $\ell_1$ minimization. We have also shown that this recovery procedure is robust in the sense that if $X$ is equal to a distributed sparse matrix plus a perturbation, then our procedure returns an approximation with accuracy proportional to the size of the perturbation. Our results follow from a new lemma about the properties of tensor products of random bipartite graphs. We also describe two interesting applications where our results would be directly applicable.

In future work, we plan to investigate the statistical behavior and sample complexity of estimating a distributed sparse matrix (and its exact support) in the presence of various sources of noise (such as additive Gaussian noise, and Wishart noise). We expect an interesting trade-off between the sketching dimension and the sample complexity.

**References**


[1] Kook Jin Ahn, Sudipto Guha, and Andrew McGregor. Graph sketches: sparsification, spanners, and subgraphs. In *Proceedings of the 31st symposium on Principles of Database Systems*, pages 5–14. ACM, 2012.

[2] R. Berinde, A.C. Gilbert, P. Indyk, H. Karloff, and M.J. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. In *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, pages 798 –805, sept. 2008.

[3] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*. Athena Scientific Belmont, MA, 1997.

[4] B. Bollobás. *Random graphs*, volume 73. Cambridge university press, 2001.

[5] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

[6] E.J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[7] Gautam Dasarathy, Parikshit Shah, Badri Narayan Bhaskar, and Robert Nowak. Sketching sparse matrices. *CoRR*, arxiv:1303.6544[cs.IT], 2013.

[8] Peter J Diggle and Arūnas P Verbyla. Nonparametric estimation of covariance structure in longitudinal data. *Biometrics*, pages 401–415, 1998.

[9] David Leigh Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[10] M.F. Duarte and R.G. Baraniuk. Kronecker product matrices for compressive sensing. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 3650 –3653, march 2010.

[11] Anna C Gilbert and Kirill Levchenko. Compressing network graphs. In *Proceedings of the LinkKDD workshop at the 10th ACM Conference on KDD*. Citeseer, 2004.

[12] S. Jokar. Sparse recovery and kronecker products. In *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*, pages 1 –4, march 2010.

[13] M. Amin Khajehnejad, Alexandros G. Dimakis, Weiyu Xu, and Babak Hassibi. Sparse recovery of positive signals with minimal expansion. *CoRR*, abs/0902.4045, 2009.

[14] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science Signalling*, 308(5721):523, 2005.