Sparse random structures: Analysis and Computations

Emmanuel Candes (Caltech),
Alan Edelman (MIT),
John Gilbert (University of California, Santa Barbara),
Raj Rao Nadakuditi (University of Michigan),
Roland Speicher (Queen's University),
Balint Virag (University of Toronto)

January 24 - January 29, 2010

The workshop was organized with an intent to extend to sparse and combinatorial structures the benefits that random matrix theory has had on continuous and dense systems. We brought together researchers working across the breadth of science and academia on applications involving random structures as they arose in the context of random graphs, networks, compressed sensing, sparse matrices, and low rank approximation theory. Many of the participants initially remarked that they were "the only one studying this at the workshop" only to find that others at the workshop, coming from different research communities were studying similar objects with an entirely different perspective. As the workshop progressed this led to increasingly robust conversations on mutual interests and broadened the community of researchers that each participant could now access for future efforts in this exciting and increasingly relevant area.

1 Presentation Highlightss

The underlying theme of this work was sparsity and randomness in all its computational and theoretical manifestations. As organizers, we were inspired by the observation that more often than noticed, a computational trick can also be a theoretical trick. Thus researchers working at the interface of computational mathematics with an eye towards exploiting sparsity (or the concept of few amongst many) have much in common with theoreticians who exploit sparsity to distill a complex problem into its analytical essence.

While there are a number of outstanding references on random structures available to practitioners wanting to address a new problem they might encounter, the goal of this workshop was to making this rich body of knowledge accessible to the non-specialist so that they might jump-start the discovery of new applications of this theory. Introducing mathematicians across many research communities to a set of related applied problems can lead to the development of newer and more powerful techniques. We believe that bringing together the practitioners and the mathematicians will jump start research in the area of sparse random structures.

To that end, the workshop was a resounding successes. We were able to involve many young researchers so that they might benefit from the interplay of disciplines represented so as to obtain breakthroughs that are so valuable in this area.

Since the workshop brought together researchers from many different traditional areas under the umbrella of a still-nascent and developing area of sparse random structures, we will jump past a formal description of the "area". Instead we provide a fuller description of the presentations with the goal of exposing the intertwining lines of inquiry. Then we will point out some of the more promising future directions.

2 Presentation Highlights

Since the workshop brought together researchers from quite different communities there were a wide variety of talks. The speakers brought into sharp focus the different notions of sparsity encountered in their respective areas and the role of random matrices in it. There was an explicit effort made to point out open problems and possible future directions. We now highlight some of the presentations with an emphasis on identifying connections between different areas made during the meeting and opportunities for creative application of the theory of random matrices for solving, as yet, unsolved problems.

2.1 Combinatorial techniques

Combinatorial techniques play an important role in the theory of random matrices in the enumeration of quantities such as the moments and free cumulants. The presentations of McKay, Mingo and Novak focused on new techniques that could facilitate the analysis of increasingly complicated sparse random structures and the computation of precise bounds for combinatorial quantities of interest.

McKay gave a tutorial overview of the method of switchings [14], and showed how an estimate of the ratio of the sizes of a family of sets may be estimated from the relations between the sets. Such an analysis can yield useful precise estimates in the settings where the digraph of sets and relations is a path; in a more complicated setting useful tail bounds can be obtained. In the context of the theme of the workshop, McKay showed how this method is a particularly powerful tool for obtaining the best known results for sparse enumeration.

Mingo build on this general theme by discussing how sharp bounds for sums associated with graphs of matrices [19] may be obtained. Novak exposed the connection between combinatorial identities arising out of the study of restricted permutation matrices and Mehta's integral in random matrix theory [20].

McKay's presentation triggered a lot of informal discussion since the techniques he presented were not well known to researchers in random matrix theory. What emerged was a hope that these techniques could be employed to yield tighter bounds in random matrix theoretic settings.

2.2 Random graph models and analysis

One of the areas emphasized in the workshop was the study of real-world networks and graphs using techniques from random matrix theory. To that end, the presentations of Davis and Lescovec highlighted the progress made and the significant challenges that remain in the bridging the gap between the development of models for random graph that are analytically tractable and that capture the characteristics observed in real-world graphs and networks. Davis' talk described the University of Florida Sparse Matrix Collection which is a large, widely available and actively grow set of sparse matrices that arise in real applications as diverse as electromagnetics, semiconductor devices, thermodynamics, computer graphics/vision and networks to name a few. The collection [9] is widely used by the sparse matrix algorithms community for the development and performance evaluation of sparse matrix algorithm and Davis' talk led to discussions on how such an accessible database may be used to develop analytical models.

Lescovec's talk picked up on this theme by surveying some of the analytical models in the literature that attempt to capture various aspects of network structure and highlighted their advantages in terms of analytical tractability and their shortcomings in terms of their ability to predict the sorts of features that arise when studying real-world data sets. The breakthrough work by Lescovec was the development of a generative analytical sparse random network/graph model called the Kronecker graph model that is analytical tractable and is able to accurately capture features that arise in real-world data sets. Lescovec described the class of binomial attribute graphs, which are an extension of the Kronecker graph model, that accounts for the heterogeneity in the population of nodes [18, 17].

The presentations by Sen and Dumitriu highlighted new analytical results for classical random graph models. Sen considered the spectral distribution of the adjacency matrix for a wide variety of random trees such as preferential attachment trees, random recursive trees, random binary trees, uniform random trees and provided analytical results for the same [5]. Dumitriu discussed the asymptotic behavior of the spectrum and eigenvectors of adjacency matrices of d-regular random n-graphs in various scaling regimes [12]. Rogers surveyed the cavity approach [23] for analyzing sparse random matrices and showed how it can be used to rederive McKay's law for d-regular random graphs. Biane described operator algebraic techniques for analyzing non-Hermitaian techniques [6] while Bordenave emphasized analytical techniques that facilitate the analysis of heavy-tailed random matrices [7].

An important outcome of the discussion that followed was the recognition that the prediction the eigenspectrum of the Kronecker graph and binomial attribute graph models is an important open question that might be solved using the techniques from random matrix theory employed in the analysis of sparse and non-Hermitian random matrices. This triggered discussions on the extensions needed to be able to analyze the (complex) eigen and signular-spectrum Kronecker graph model.

2.3 Discovering structure in graphs

The sequence of talks by Djidvjev, Harding and Mahoney addressed the practically important question of discovering structure in graphs. One way to analyze and understand the structure and the functioning of large networks is to divide their nodes into communities/clusters (maximal groups of nodes with denser in-cluster links and fewer links connecting nodes from different clusters). Djidjev showed how the problem of finding a partition maximizing the modularity of a given network could be reduced to solving a number of minimum weighted cut problems on a complete graph with the same vertices as the original network and appropriately defined edge weights [11]. The resulting minimum cut problem could then be efficiently solved using multi-level graph partitioning methods. Harding presented an alternate way of tackling this problem by leveraging results on the moments of adjacency matrices from random matrix theory and by the employing generalized method of moments (GMM) to uncover the underlying deterministic structure of linking within various communities. Preciado extended this approach by considering the spectral moments of the Laplacian matrix of the network and provided a complete characterization of what measurements are most relevant to characterize the Laplacian spectrum from the point of view of community detection [22].

Mahoney presented a survey style talk where he discussed recent empirical results on the structural properties of large social and information networks and argued these networks are particularly ill-suited for analysis with many traditional machine learning and data analysis tools. Mahoney's insight was that this has to do with the fact that the relationship between structures that may be interpreted as geometric and structures that exhibit empirical signatures of quasirandomness is substantially more complex in large social and information networks than it is in many more traditional classes of data.

2.4 Techniques for recovery with missing or corrupted data

Compressed sensing is an important new technique for acquiring and reconstructing a signal utilizing the prior knowledge that it is sparse or compressible. Plan presented several novel theoretical results regarding the recovery of a low-rank matrix from a sparse set of measurements consisting of linear combinations of the matrix entries using sparse approximation techniques [8]. Kolda looked at the problem of producing a factorization of a tensor structured data set in the setting where the data set has missing values. Kolda's presented an algorithm and showed results from a numerical simulations for settings showing that even when a lot of data is discarded, the recovered factorization is still very accurate [1]. The discussions identified an opportunity to utilize techniques from the analysis of matrix completion with missing data to analyze the tensor factorization completion in the missing data setting.

Ward extends results in the literature about recovery of sparse trigonometric polynomials from few point samples to the recovery of polynomials having a sparse expansion in Legendre basis. Vavasis showed how NP-hard problems such as clique and biclique can be solved by nuclear norm minimization [2].

2.5 Computational aspects

Randomization has emerged as a powerful tool for speeding up computations. Gunnar-Martinsson spoke of techniques for enabling very large-scale computations - random matrix theory plays a critical role here in providing performance guarantees in the form of accuracy bounds. The underlying techniques use randomized sampling to reduce the effective dimensionality of the data while loosening communication constraints, and maintaining, or even improving, the accuracy and robustness of existing deterministic techniques [13].

Gilbert's presentation dwelled on the role of linear algebra as a high-level algebraic primitive for computing on matrices or matrix discretizations of operators. Gilbert's talk triggered a spirited discussion on the challenge of finding equivalent algebraic primitives for computation on large graphs.

2.6 Statistical applications

The notion of sparsity arises in statistical applications in the notion that the signal occupies a low rank subspace relative to the noise. The presentations of Hero, Nadakuditi and Perry dealt with issues arising from the discriminating between low rank signal in full rank noise and the fundamental limits of these techniques. Hero spoke of the problem of screening a large number of variables for pairwise correlations; Nadakuditi described random matrix theory based predictions for when principal component analysis fails while Perry discussed issues related to identification of the signal subspace with cross-validation [21]. Virag presented an operator theoretic derivation of the phase transition in distributions for the spiked Wishart model.

The discussions that followed identified an opportunity to use Nadakuditi's linear algebraic to provide an alternate derivation of the phase transition of spiked Wishart models. Kritchevskii discussed potential extensions of the theory to finite rank perturbations of infinite dimensional stochastic operators.

2.7 Condition number estimation

The presentations by Blake, Rudelson and Loh focused on aspects related to condition number estimation of random matrices. Rudelson focused on random conjunction matrices [16] and highlighted how techniques, borrowed from the analysis of high-dimensional convex bodies, were brought to bear on this problem. Blake discussed aspects related to the condition number distribution of random matrices that arise in the context of coding theory. Blake made several conjectures on the behavior of random matrices over finite fields. The discussions that ensued highlighted an opportunity to extend techniques from matrix theory to verify these conjectures.

2.8 Quantum information theory

An interesting connection that emerged in the workshop was the connection between random matrix theory, quantum information theory and compressed sensing. Hayden presented a survey talk on applications and desiderata for random matrices in quantum information theory. The role of random matrices arises naturally because quantum mechanics is noncommutative and the random objects of study are invariably matrices. Quantum information theory therefore provides a rich source of random matrix problems with applications ranging from the best known codes for sending quantum data through noisy media to subroutines in quantum algorithms and new encryption procedures.

In many of these applications random matrix theory plays an integral role in establishing existence theorems. In that respect, there are strong parallels to compressed sensing where random sampling matrices satisfying the RIP facilitate exact reconstruction. In quantum information theory Haar-distributed unitary matrices play an analogous role [15].

An important open problem identified in the workshop that can benefit from increased interaction between the random matrix theory, compressed sensing [10] and quantum information theory community is to construction of efficient, constructible deterministic (i.e. non-random) of matrices that achieve the same performance. Curran spoke about connections between quantum groups and free probability theory [3].

3 Outcome of the Meeting

Spirited discussions were held throughout the workshop. Since many of the attendees were from different communities, the breakfast, lunch and dinner sessions at the Banff Centre led to many opportunities for informal discussions in which open problems, opportunities for collaborations, areas of mutual interest and possible future directions were discussed. In the following we list some of the major relevant problems which were pointed out in this discussion meeting and in numerous other discussions between the participants throughout the workshop.

1. McKay's presentation triggered a lot of informal discussion since the techniques he presented were not well known to researchers in random matrix theory. What emerged was a hope that these techniques could be employed to yield tighter bounds in random matrix theoretic settings.

2. An important outcome of the discussion that followed was the recognition that the prediction the eigenspectrum of the Kronecker graph and binomial attribute graph models is an important open question that might be within reach due to recent developments in operator valued free probability theory. This triggered discussions on the extensions in operator valued free probability theory needed to be able to analyze the Kronecker graph model.

3. Biane's presentation on using tools from free probability to compute the Brown Measure led to a discussion on the (close) relationship between the Brown Measure and the (complex) eigenvalue distribution of non-Hermitian - random matrices. Bordenave, Rogers and Biane led discussions on how to make the connection more precise in a rigorous manner that takes advantages of the breakthroughs of Tao and Vu in proving the universality of Girko's circular law.

4. Hayden identified an important open problem in the workshop that can benefit from increased interaction between the random matrix theory, compressed sensing and quantum information theory community. The challenge is to construct efficient, deterministic (i.e. non-random) matrices that achieve (nearly the) same performance as their well-known random counterparts.

5. Blake's presentation brought into sharp focus the open problem of characterizing the condition number distributions of random matrices whose elements are drawn at random from a finite field.

6. Edelman revisited the connection between random matrices and random polytopes and alluded to the impressive recent work in this area by Veryshynin, Rudelson, et al. An open problem is to find the counterparts (if any) of the famous distributions arising in random matrix theory for random problems.

7. Loh's presented a numerical linear algebraic perspective of Tao-Vu smallest singular value theorem. He presented numerical evidence that indicated that the Tao-Vu theorem [24] is conservative in the sense that it dramatically underestimates the rate of convergence of the smallest singular value of a Gaussian random matrix.

8. Nadakuditi's work [4] brought into sharp focus the connection between free probability and the thresholds at which the phase transition in the eigenvalues and eigenvectors of low rank perturbations of large random matrices sets in. An open problem is to rigorously extend this theory to the infinite dimensional stochastic operator setting.

9. Dumitriu's work provides new insights on the behavior of d-random regular graphs in different scaling regimes. An open problem is to provide a complete analysis and characterization of any associated localization/delocalization phenomenon in the eigenvectors.

10. Mahoney and Gilbert's talks spelt out the challenge for finding the right algebraic primitives for computational and statistical inference on graphs and networks.

11. Plan and Kolda's work on matrix and tensor completion with missing data (respectively) brought into focus the open problem of identifying the breakdown point of such completion analogous to the Donoho-Tanner boundary in sparse approximation theory for the matrix and tensor setting.

12. Vavasis' work on clique discovery using nuclear norm optimization revealed that a lot is less is known on the robustness properties of these algorithms to adversarial attacks (such as edge deletions, etc.). There are conjectured connections with Nadakuditi's work [4] on the breakdown of subspace methods that merit deeper understanding.

References

[1] E. Acar, T.G. Kolda, D.M. Dunlavy, and M. Morup. Scalable Tensor Factorizations for Incomplete Data. *Arxiv preprint arXiv:1005.2197*, 2010.

[2] B. Ames and S.A. Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Arxiv preprint arXiv:0901.3348*, 2009.

[3] T. Banica, S. Curran, and R. Speicher. Classification results for easy quantum groups. *arXiv*, 906, 2009.

[4] F. Benaych-Georges and R. Rao. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Preprint*, 910, 2009.

[5] S. Bhamidi, S.N. Evans, and A. Sen. Spectra of large random trees. *Arxiv preprint arXiv:0903.3589*, 2009.

[6] P. Biane and F. Lehner. Computation of some examples of Brown's spectral measure in free probability. *Arxiv preprint math/9912242*, 1999.

[7] C. Bordenave, P. Caputo, and D. Chafai. Spectrum of non-Hermitian heavy tailed random matrices. 2010.

[8] E.J. Candes and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *Arxiv preprint arXiv:1001.0339*, 2010.

[9] T. Davis. University of Florida sparse matrix collection. *NA Digest*, 97(23):7, 1997.

[10] R.A. DeVore. Deterministic constructions of compressed sensing matrices. *Journal of Complexity*, 23(4-6):918–925, 2007.

[11] H. Djidjev. A scalable multilevel algorithm for graph clustering and community structure detection. *Algorithms and Models for the Web-Graph*, pages 117–128, 2008.

[12] I. Dumitriu and S. Pal. Sparse regular random graphs: spectral density and eigenvectors. *Arxiv preprint arXiv:0910.5306*, 2009.

[13] N. Halko, P.G. Martinsson, and J.A. Tropp. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. *arXiv*, 909, 2009.

[14] M. Hasheminezhad and B.D. McKay. Combinatorial estimates by the switching method.

[15] P. Hayden, D. Leung, P.W. Shor, and A. Winter. Randomizing quantum states: Constructions and applications. *Communications in Mathematical Physics*, 250(2):371–391, 2004.

[16] S.P. Kasiviswanathan, M. Rudelson, A. Smith, and J. Ullman. The price of privately releasing contingency tables and the spectra of random matrices with correlated rows. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 775–784. ACM, 2010.

[17] J. Leskovec. Modeling Network Structure using Kronecker Multiplication. *Network*, 100:101–102.

[18] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: an approach to modeling networks. *The Journal of Machine Learning Research*, 11:985–1042, 2010.

[19] J.A. Mingo and R. Speicher. Sharp Bounds for Sums Associated to Graphs of Matrices. *Arxiv preprint arXiv:0909.4277*, 2009.

[20] J. Novak. An asymptotic version of a theorem of Knuth. *Advances in Applied Mathematics*, 2010.

[21] P.O. Perry and P.J. Wolfe. Minimax rank estimation for subspace tracking. *Submitted, January*, 2009.

[22] V.M. Preciado and A. Jadbabaie. From Local Measurements to Network Spectral Properties: Beyond Degree Distributions. *Arxiv preprint arXiv:1004.3524*, 2010.

[23] T. Rogers, I.P. Castillo, R. Kühn, and K. Takeda. Cavity approach to the spectral density of sparse symmetric random matrices. *Physical Review E*, 78(3):31116, 2008.

[24] T. Tao and V. Vu. Random matrices: The distribution of the smallest singular values. *Geometric And Functional Analysis*, 20(1):260–297, 2010.