Learning in Networks: Performance Limits and Algorithms

Bruce Hajek (University of Illinois at Urbana-Champaign)
Yihong Wu (Yale University)
Jiaming Xu (Duke University)

November 13-18, 2022

1 Overview of the Field

The workshop focuses on statistical learning in large-scale networks, typically represented by graphs with vertices and edges. The research involves modeling the networks and observations, devising learning algorithms, analyzing the performance of the algorithms, deriving bounds on the possible performance of best algorithms, and deploying theoretically-grounded algorithms to real network data. Many machine learning problems deal with networks that encode similarities or relationships among different objects, for which observational data may be limited in extent and noisy. Thus learning the desired information requires highly efficient algorithms that can process large-scale network data and detect tenuous statistical signatures.

The aim of the workshop is to bring together the leading researchers in this area to discuss recent results and open problems, as well as to explore new mathematical techniques and models to study these problems. In addition, our goal is to give graduate students and post-doctoral researchers an opportunity to learn about recent results and important open problems in this field, as well as to present their own research.

Research on learning in networks combines techniques from probability theory, graph theory and combinatorics, statistical physics, optimization, and information theory. The researchers attending the workshop will span the disciplines of mathematics, applied probability, computer science, physics, information theory, and statistics. Besides tutorials and research presentations, the workshop will encourage and provide time for attendees to form study groups to have focused discussions on research problems and directions that arise during the workshop. The goal of the workshop is to act as a catalyst for new research directions and approaches in the emerging research area of statistical learning for large scale networks.

2 Objectives

The overall objective is to better understand what is information theoretically and computationally possible to learn from large-scale network data, and to identify the algorithms to do it. The topics are organized into three interrelated areas, ranging from inference problems for single graphs, to inference involving two graphs, to classification of graphs from general families.

2.1 Recovering structure from single graphs

Over the past ten years, there has been much focus on a particular problem of recovering structure from a graph, namely, the community detection problem. See recent expositions in [22, 1]. Community detection

has found numerous applications across various disciplines. Most work assumes that networks can be partitioned into groups of nodes with denser connections internally and sparser connections between groups, and considers random graph models with some underlying cluster structure such as the stochastic block model (SBM) or planted partition model. In its simplest form, nodes are partitioned into clusters, and any two nodes are connected independently at random with probability p if they are in the same cluster and with probability q otherwise. The problem of cluster recovery under SBM has been extensively studied and many efficient algorithms with provable performance guarantees have been developed. An intriguing computational barrier has been identified when community sizes are smaller, showing that an "informationally possible but computationally hard" regime exists under the planted clique hardness hypothesis. Promising recent developments are given in [4], and further developments in that direction will be explored at the workshop.

The stochastic block model fails to capture two basic properties of networks that are seen in practice. Firstly, it does not model networks that grow over time, such as citation networks or social networks. Secondly, it does not model graphs with heavy-tailed degree distributions, and therefore does not explain how real networks, such as the political blog network, have a few nodes with very high degrees (hubs). Existing community detection algorithms, developed for the, may not perform well on real networks, even though they achieve remarkable performance in theory. Recently, Johnathan Jordan [19] formulated a general model of preferential attachment in which the attachment weights can depend on the labels of both existing and the arriving vertices. Antunovic, Mossel, and Racz [2] proposed instead that the community membership of a new vertex be determined endogenously, based on the membership of the vertices to which the new vertex is attached. An open area of research to be addressed at the workshop is to derive community detection and recovery algorithms in the face of dynamics as modeled in these two papers.

Applications in DNA assembly and particle tracking have brought to the fore two other problems of discovering structure in graphs, namely, recovery of hidden Hamiltonian cycles and recovery of hidden bipartite matchings. For the recovery of hidden bipartite matchings, a challenging problem is to prove the information-theoretic limit, which was conjectured and empirically computed by the statistical physicists [8]. Researchers working this area from the physics side, statistics side, and mathematical analysis side will be able to exchange ideas to push for breakthroughs on this conjecture at the workshop. The hidden Hamiltonian cycle recovery problem is motivated from de novo genome assembly, the reconstruction of an organism's long sequence of A,G,C,T nucleotides from high-throughput, fragmented DNA sequencing data. The high-throughput DNA sequencing method generates billions of short fragment reads of DNA with low cost; however, these short reads are of only about 100 or 200 base pairs in length and c prone to errors; thus it is challenging to assemble those noisy, short reads to reconstruct a long and contiguous genome sequence both accurately and efficiently. Despite the significant algorithmic progress for genome assembly from short reads [5], we are still far from obtaining a high-quality and highly contiguous genome sequence. This will be an interesting topic of focus at the workshop.

2.2 Learning from multiple graphs

Graph matching, or learning the vertex correspondence between two edge-correlated graphs, is an interesting prototype problem with numerous applications in network privacy, system biology, computer vision, and natural language processing [9]. Both seeded graph matching wherein an initial seed set of correctly matched vertex pairs is revealed as side information and the seedless graph matching wherein such an initial seed set is unavailable are of interest. Driven by applications in network de-anonymization, a recent line of work initiated the statistical analysis of graph matching by assuming that G_a and G_b are randomly generated according to the following correlated Erdos-Renyi random graphs model $\mathcal{G}(n, q; s)$. Specifically, $G_a \sim \mathcal{G}(n, q)$ and, viewing G_a and G_b as adjacency matrices, G_b is obtained from G_a by flipping $1 \to 0$ with probability $1-s$ and $0 \to 1$ with probability $q(1-s)/(1-q)$ so that $G_b \sim \mathcal{G}(n, q)$. Equivalently, imagine there is a parent Erdos-Renyi graph $\mathcal{G}(n, p)$, where $p = q/s$, and G_a and G_b are each obtained by independently removing each edge from the parent graph with probability s. One can imagine G_a is a facebook friendship network of a group of people with identities removed, and G_b is the twitter network of the same group of people with known identities; the task is to de-anonymize the vertex identities in facebook network by finding the underlying mapping between the vertex sets of G_a and G_b.

In the noiseless case $s = 1$, graph matching under $\mathcal{G}(n, q; 1)$ reduces to the graph automorphism problem for an Erdos-Renyi graph $\mathcal{G}(n, q)$. In this case, a celebrated result [26] shows that exact recovery of

the underlying permutation is information-theoretically impossible if and only if $nq \geq \log n + \omega(1)$ for $q \leq 1/2$. Recent work [11] has extended this result to noisy case where $s < 1$, showing that the maximum likelihood estimator, or equivalently the optimum of quadratic assignment problem, coincides with the ground truth π^* with probability $1 - o(1)$, provided that $nqs \geq \log n + \omega(1)$ under additional assumptions $q \leq O(\log^{-1} n)$ and $q(1-s)^2/s \leq O(\log^{-3}(n))$; on the contrary, any estimator is correct with probability $o(1)$, if $nqs \leq \log n - \omega(1)$. From a computational perspective, in the noiseless case with $s = 1$, there exist linear-time algorithms which attain the recovery threshold whenever $np = \log n + \omega(1)$. However, in the noisy case, very little is known about the performance guarantees of efficient graph matching algorithms. A focus of the workshop will be on the computational - informational gap, and tools from quadratic assignment algorithms and stochastic analysis should be brought together to bear on this problem. Experts on the quadratic assignment problem will be able to exchange ideas with physicists and statisticians on this problem at the workshop.

2.3 Learning graphical properties efficiently via graph sampling

Learning properties of large graphs from samples is an important problem in statistical network analysis, dating back to the early work of Goodman [17] and Frank [14]. In many big-data analytic applications such as social networks and Internet service provider networks which typically involve billions of vertices, the full network is either unobserved (due to the underlying experimental design and data collection mechanism) or too expensive to be stored (due to constraints on computational resources) [10]. What is observed is a partial snapshot of the network, e.g., a subset of the vertices or edges, which can be viewed as sampled from a statistical model. Furthermore, even when the full network can be observed, sampling is a powerful summarization technique that significantly reduces the amount of data while retaining the ability to reconstruct the important features of the original network. In dealing with sampled network data, one must rely on carefully designed statistical estimators that take into account the bias induced by sampling in order to conduct sound inference.

Motivated by applications in social networks, econometrics, and Internet tomography, one of the most useful graph properties to learn from samples is the number of various features in a network, including basic local structures such as motifs or graphlets (e.g. edges, triangles, wedges, stars, cycles, cliques, clustering coefficients), or global features such as the number of connected components. Various algorithms based on edge and degree queries have been proposed in the computer science literature to estimate the average degree, triangle counts, and more general subgraph counts or connected components [3] in sublinear time; however, these results largely focus on time complexity and use adaptive queries that might not be possible in weaker sampling models. In the statistics literature, estimating graph properties has a long history dating back to the early work of Frank, Capobianco, and Goodman; however, these results are mainly confined to unbiased estimators and little is known about their optimality. Building on recent work on estimating distributional properties on large domains, the workshop will focus on a modern study of learning graph properties under various sampling models under the new light of high-dimensional statistics.

3 Presentation Highlights

3.1 Algorithmic Decorrelation and Planted Clique in Dependent Random Graphs

[Guy Bresler] There is a growing collection of average-case reductions starting from Planted Clique (or Planted Dense Subgraph) and mapping to a variety of statistics problems, sharply characterizing their computational phase transitions. These reductions transform an instance of Planted Clique, a highly structured problem with its simple clique signal and independent noise, to problems with richer structure. In this talk we aim to make progress in the other direction: to what extent can these problems, which often have complicated dependent noise, be transformed back to Planted Clique? Such a bidirectional reduction between Planted Clique and another problem shows a strong computational equivalence between the two problems. As a concrete instance of a more general result, we consider the planted clique (or dense subgraph) problem in an ambient graph that has dependent edges induced by randomly adding triangles to the Erdos-Renyi graph G(n,p), and show how to successfully eliminate dependence by carefully removing the triangles while approximately preserving the clique (or dense subgraph). In order to analyze our reduction we develop new

methods for bounding the total variation distance between dependent distributions. Joint work with Chenghao Guo and Yury Polyanskiy.

3.2 Spectral algorithms for community detection

[Julia Gaudio] Many networks exhibit community structure, meaning that there are two or more groups of nodes which are densely connected. Identifying these communities gives valuable insights about the latent features of the nodes. Community detection has been used in a wide array of applications including online advertising, recommender systems (e.g., Netflix), webpage sorting, fraud detection, and neurobiology. I will present my work on algorithms for community detection in two contexts, each with an underlying probabilistic generative model.

(1) Censored networks: How can we identify communities when some connectivity information is missing? Here we consider recovery from the Censored Stochastic Block Model. (Joint work with Souvik Dhara, Elchanan Mossel, and Colin Sandon [12])

(2) Higher-order networks: Beyond pairwise relationships. Here we consider recovery from the Hypergraph SBM, where we are given access to the "similarity matrix" of the hypergraph. (Joint work with Nirmit Joshi [15])

We show that simple spectral algorithms achieve the information-theoretic thresholds of both exact recovery problems.

3.3 The (symmetric) Ising perceptron: progress and problems

[Will Perkins] The Perceptron model was proposed as early as the 1950's as a toy model of a one-layer neural network. The basic model consists of a set of solutions (either the Hamming cube or the sphere of dimension n) and a set of constraints given by n-dimensional Gaussian vectors. The constraints are that the inner product of a solution vector with each constraint vector scaled by sqrtn must lie in some interval on the real line. Probabilistic questions about the model include the satisfiability threshold (or the "storage capacity") and questions about the typical structure of the solution space. Algorithmic questions include the tractability of finding a solution (the learning problem in the neural network interpretation). I will survey the model, the main problems, and recent progress.

3.4 Stochastic Bin Packing with Time-Varying Item Sizes

[Weina Wang] In today's computing systems, there is a strong contention between achieving high server utilization and accommodating the time-varying resource requirements of jobs. Motivated by this problem, we study a stochastic bin packing formulation with time-varying item sizes, where bins and items correspond to servers and jobs, respectively. Our goal is to answer the following fundamental question: How can we minimize the number of active servers (servers running at least one job) given a budget for the cost associated with resource overcommitment on servers? We propose a novel framework for designing job dispatching policies, which reduces the problem to a policy design problem in a single-server system through policy conversions. Through this framework, we develop a policy that is asymptotically optimal as the job arrival rate increases. This is a joint work with Yige Hong at Carnegie Mellon University and Qiaomin Xie at the University of Wisconsin–Madison.

3.5 Spectral pseudorandomness and the clique number of the Paley graph

[Dmitry Kunisky] The Paley graph is a number-theoretic construction of a graph on the vertex set of a finite field of prime order p that in many ways behaves "pseudorandomly." One manifestation of pseudorandomness is that the clique number of the Paley graph is widely believed to be polylogarithmic in p. In contrast, the best known upper bounds are only of order square root of p; it is a long-standing open problem in number theory to improve on this scaling.

I will present several pieces of recent and ongoing work studying approaches to this question based on convex optimization and spectral graph theory, which involve understanding the extent to which the Paley graph is "spectrally pseudorandom" in various senses. First, I will show that the degree 4 sum-of-squares

relaxation of the clique number of the Paley graph has value at least the cube root of p, derandomizing an analogous result for Erdos-Renyi random graphs due to Deshpande and Montanari (2015). On the other hand, I will offer some evidence that this relaxation may in fact yield bounds of polynomial scaling between the square and cube roots of p, thanks to the spectrum of the Paley graph being sufficiently different from that of an Erdos-Renyi graph. Second, I will show that certain deterministic induced subgraphs of the Paley graph have the same limiting spectrum as induced subgraphs on random sets of vertices of the same size. I will outline how stronger results of this form would also lead to clique number bounds improving on the state of the art.

Based partly on joint work with Xifan Yu.

3.6 Correlated stochastic block models: graph matching and community recovery

[Miklos Racz] I will discuss statistical inference problems on edge-correlated stochastic block models. We determine the information-theoretic threshold for exact recovery of the latent vertex correspondence between two correlated block models, a task known as graph matching. As an application, we show how one can exactly recover the latent communities using multiple correlated graphs in parameter regimes where it is information-theoretically impossible to do so using just a single graph. Furthermore, we obtain the precise threshold for exact community recovery using multiple correlated graphs, which captures the interplay between the community recovery and graph matching tasks. This is based on joint work with Julia Gaudio and Anirudh Sridhar [16].

3.7 Uniqueness of BP fixed point for Ising models

[Yury Polyanskkiy] In the study of Ising models on large locally tree-like graphs, in both rigorous and non-rigorous methods one is often led to understanding the so-called belief propagation distributional recursions and its fixed point (also known as Bethe fixed point, cavity equation, 1RSB etc). In this work we prove there is at most one non-trivial fixed point for Ising models for both zero and certain random external fields.

As a concrete example, consider a sample A of Ising model on a rooted tree (regular, Galton-Watson, etc). Let B be a noisy version of A obtained by independently perturbing each spin as follows: B_v equals to A_v with some small probability δ and otherwise taken to be a uniform +-1 (alternatively, 0). We show that the distribution of the root spin A_p conditioned on values B_v of all vertices v at a large distance from the root is independent of δ and coincides with δ=0. Previously this was only known for sufficiently "low-temperature" models. Our proof consists of constructing a metric under which the BP operator is a contraction (albeit non-multiplicative). I hope to convince you our proof is technically rather simple.

This simultaneously closes the following 5 conjectures in the literature: uselessness of global information for a labeled 2-community stochastic block model, or 2-SBM (Kanade-Mossel-Schramm'2014); optimality of local algorithms for 2-SBM under noisy side information (Mossel-Xu'2015); independence of robust reconstruction accuracy to leaf noise in broadcasting on trees (Mossel-Neeman-Sly'2016); boundary irrelevance in broadcasting on trees (Abbe-Cornacchia-Gu-P.'2021); characterization of entropy of community labels given the graph in 2-SBM (ibid).

Joint work with Qian Yu (Princeton) [27].

3.8 Finite-sample lower bounds on information requirements for causal network inference

[Xiaohan Kang] Recovery of the causal structure of dynamic networks from noisy measurements has long been a problem of intense interest across many areas of science and engineering. Many algorithms have been proposed, but there is no work that compares the performance of the algorithms to converse bounds in a non-asymptotic setting. As a step to address this problem, this talk discusses lower bounds on the error probability for causal network support recovery in a linear Gaussian setting [20]. The bounds are based on the use of the Bhattacharyya coefficient for binary hypothesis testing problems with mixture probability distributions. Comparison of the bounds and the performance achieved by two representative recovery algorithms are given for sparse random networks based on the Erdős–Rényi model. A related problem of estimating the error probabilities for a binary hypothesis testing problem from likelihood ratio samples is also discussed.

3.9 Attributed Graph Alignment: Fundamental Limits and Efficient Algorithms

[Lele Wang] We consider the graph alignment problem, where the goal is to identify the vertex/user correspondence between two correlated graphs. Existing work mostly recovers the correspondence by exploiting the user-user connections. However, in many real-world applications, additional information about the users, such as user profiles, might be publicly available. In this talk, we introduce the attributed graph alignment problem, where additional user information, referred to as attributes, is incorporated to assist graph alignment. We establish both the information-theoretic limits and the feasible region by polynomial-time algorithms for the attributed graph alignment. Our results span the full spectrum between models that only consider user-user connections and models where only attribute information is available [24].

3.10 Average-Case Computational Complexity of Tensor Decomposition

[Alex Wein] Tensor decomposition is an algorithmic primitive with applications in many machine learning tasks, including community detection and its mixed-membership or multi-layer variants.

We consider a simple model for tensor decomposition: suppose we are given a random rank-r order-3 tensor—that is, an n-by-n-by-n array of numbers that is the sum of r random rank-1 terms—and our goal is to recover the individual rank-1 terms. In principle, this decomposition task is possible when $r < cn^2$ for a constant c, but all known polynomial-time algorithms require $r \ll n^{3/2}$. Is this a fundamental barrier for efficient algorithms?

In recent years, the average-case complexity of various high-dimensional statistical tasks has been resolved in restricted-but-powerful models of computation such as statistical queries, sum-of-squares, or low-degree polynomials. However, tensor decomposition has remained elusive, largely because its hardness is not explained by a "planted versus null" testing problem. We show the first formal hardness for average-case tensor decomposition: when $r \gg n^{3/2}$, the decomposition task is hard for algorithms that can be expressed as low-degree polynomials in the tensor entries [25].

3.11 Random graph matching at Otter's threshold via counting chandeliers

[Jiaming Xu] We propose an efficient algorithm for graph matching based on similarity scores constructed from counting a certain family of weighted trees rooted at each vertex. For two ER graphs $\mathcal{G}(n, q)$ whose edges are correlated through a latent vertex correspondence, we show that this algorithm correctly matches all but a vanishing fraction of the vertices with high probability, provided that $nq \to \infty$ and the edge correlation coefficient ρ satisfies $\rho^2 > \alpha \approx 0.338$, where α is Otter's tree-counting constant. Moreover, this almost exact matching can be made exact under an extra condition that is information-theoretically necessary. This is the first polynomial-time graph matching algorithm that succeeds at an explicit constant correlation and applies to both sparse and dense graphs. In comparison, previous methods either require $\rho = 1 - o(1)$ or are restricted to sparse graphs.

The crux of the algorithm is a carefully curated family of rooted trees called chandeliers, which allows effective extraction of the graph correlation from the counts of the same tree while suppressing the undesirable correlation between those of different trees.

Based on joint work with Cheng Mao (Gatech), Yihong Wu (Yale), and Sophie H. Yu (Duke) [21].

3.12 Local and global expansion in random geometric graphs

[Tselil Schramm] Consider a random geometric 2-dimensional simplicial complex X sampled as follows: first, sample n vectors $u_1, \ldots, u_n$ uniformly at random on S^{d-1} ; then, for each triple $i, j, k \in [n]$, add $\{i, j, k\}$ and all of its subsets to X if and only if $\langle u_i, u_j \rangle \geq \tau$, $\langle u_j, u_k \rangle \geq \tau$, $\langle u_i, u_k \rangle \geq \tau$. We prove that for every $\epsilon > 0$, there exists a choice of $d = \Theta(\log n)$ and $\tau = \tau(\epsilon, d)$ so that with high probability, X is a high-dimensional expander of average degree n^ϵ in which each 1-link has spectral gap bounded away from $1/2$.

3.13 Optimal Full Ranking from Pairwise Comparisons

[Chao Gao] We consider the problem of ranking n players from partial pairwise comparison data under the Bradley-Terry-Luce model. For the first time in the literature, the minimax rate of this ranking problem is derived with respect to the Kendall's tau distance that measures the difference between two rank vectors by counting the number of inversions. The minimax rate of ranking exhibits a transition between an exponential rate and a polynomial rate depending on the magnitude of the signal-to-noise ratio of the problem. To the best of our knowledge, this phenomenon is unique to full ranking and has not been seen in any other statistical estimation problem. To achieve the minimax rate, we propose a divide-and-conquer ranking algorithm that first divides the n players into groups of similar skills and then computes local MLE within each group. The optimality of the proposed algorithm is established by a careful approximate independence argument between the two steps [6].

3.14 Universality of Approximate Message Passing algorithms and tensor networks

[Zhou Fan] Approximate Message Passing (AMP) algorithms provide a valuable tool for studying mean-field approximations and dynamics in a variety of applications. Although usually derived for matrices having independent Gaussian entries or satisfying rotational invariance in law, their state evolution characterizations are expected to hold over larger universality classes of random matrix ensembles [23].

3.15 Revisiting Jerrum's Metropolis Process for the Planted Clique Problem

[Ilias Zadik] Jerrum in 1992 (co-)introduced the planted clique model by proving the (worst-case initialization) failure of the Metropolis process to recover any o(sqrt(n))-sized clique planted in the Erdos-Renyi graph G(n,1/2). This result is classically cited in the literature of the problem, as the "first evidence" the o(sqrt(n))-sized planted clique recovery task is "algorithmically hard".

In this work, we show that the Metropolis process actually fails to work (under worst-case initialization) for any o(n)-sized planted clique, that is the failure applies well beyond the sqrt(n) "conjectured algorithmic threshold". Moreover we also prove, for a large number of temperature values, that the Metropolis process fails also under "natural initialization", resolving an open question posed by Jerrum in 1992. This is joint work with Zongchen Chen and Elchanan Mossel [7].

3.16 Detection-Recovery Gap for Planted Dense Cycles

[Cheng Mao] Planted dense cycles are a type of latent structure that appears in many applications, such as small-world networks in social sciences and sequence assembly in computational biology. We consider a model where a dense cycle with expected bandwidth $n\tau$ and edge density p is planted in an Erdős–Rényi graph $G(n, q)$. We characterize the computational thresholds for the associated detection and recovery problems for the class of low-degree polynomial algorithms. In particular, a gap exists between the two thresholds in a certain regime of parameters. For example, if $n^{-3/4} \ll \tau \ll n^{-1/2}$ and $p = Cq = \tilde{\Theta}(1)$ for a constant $C > 1$, the detection problem is computationally easy while the recovery problem is hard.

3.17 On community detection in preferential attachment networks

[Bruce Hajek] A message passing algorithm is derived for recovering communities within a graph generated by a variation of the Barabási-Albert preferential attachment model [18]. The estimator is assumed to know the arrival times, or order of attachment, of the vertices. The derivation of the algorithm is based on belief propagation under an independence assumption. Two precursors to the message passing algorithm are analyzed: the first is a degree thresholding (DT) algorithm and the second is an algorithm based on the arrival times of the children (C) of a given vertex, where the children of a given vertex are the vertices that attached to it. Comparison of the performance of the algorithms shows it is beneficial to know the arrival times, not just the number, of the children. The probability of correct classification of a vertex is asymptotically determined by the fraction of vertices arriving before it. Two extensions of Algorithm C are given: the first is based on joint likelihood of the children of a fixed set of vertices; it can sometimes be used to seed the message passing algorithm. The second is the message passing algorithm. Simulation results are given.

4 Outcome of the meeting

There were many interesting connections among the problems and results discussed at the meeting. It demonstrated a vibrant major research effort underway to understand the performance limits, from both an information and computational viewpoint, for many mostly unsupervised statistical learning problems. The workshop brings together many young researchers and promotes research from members of underrepresented groups.

The planted clique problem, the planted dense subgraph problem, and the problem of finding the largest clique in an Erdös-Rényi random graph were the primary examples of problems believed to be computationally difficult in the average case. Zadik presented work on the limitations of the Metropolis algorithm for such problems. Bresler focused on expanding the set of equivalently difficult problems. Wein focused on proving the performance limits of low-degree polynomial algorithms. Mao presented new information-theoretic and computational thresholds for the planted dense cycles model.

Graph matching and community detection problems were topics of many of the talks. Wang presented results of graph matching with attribute information. Xu presented results on efficient algorithms for graph matching based on counting subgraphs and identified a possible boundary in parameter space where a gap emerges between information limits and computationally feasible limits. Gaudia presented results showing the power of spectral methods, with careful weighting, can provide exact recovery up to information-theoretic limits in a censored block model. Racz presented the sharp information-theoretic threshold for community recovery under two correlated stochastic block models.

Progress continues to be made in the special case of sparse random graphs, related to random trees and message-passing algorithms. Polyanski presented a general method based on information theory to prove in great generality that there is at most one solution to a key fixed point equation for probability distributions that have naturally arisen in many works in this area such as the BP iteration. Fan presented new tools for analyzing the approximate message-passing algorithms in dense regimes.

Kunisky presented work showing an interplay between statistical methods and number theoretic methods. Perkins presented an overview and recent results on the storage capacity and typical complexity for the perceptron storage problem. Schramm presented work showing the high-dimensional expansion and spectral property of random geometric graphs.

5 Open problems

Through talks, discussions offline, and an open problem session, many interesting open problems and conjectures were raised. Some of the open problems discussed are listed here.

1. (Raised by Mikos Racz) Given an Erdos-Renyi graph with parameter ½, consider the problem of identifying a large clique by an adaptive algorithm that queries pairs of vertices and for each pair learns whether the pair is an edge in the graph. Suppose computation power is not constrained but only $O(n)$ queries are permitted. A simple two-stage algorithm can find a clique of size $\frac{3}{2}\log_2 n$ with high probability. An open question is whether cliques larger than that can be found with $O(n)$ queries. There is a published upper bound on the largest size that can be found of roughly $(1.8)\log_2 n$.

2. (Attributed to Karp) For Erdos-Renyi graphs with parameter ½, determine the largest clique that can be found using a polynomial complexity algorithm. In particular, can such algorithms achieve size $\alpha \log_2 n$ for some $\alpha > 1$ as $n \to \infty$?

3. (Raised by Yihong Wu) Computational gap for learning mixture of Gaussians? Given $X_1, \ldots, X_n$ drawn iid from a density f which is k-component Gaussian mixture (k-GM), namely, $\sum_{i=1}^{k} w_i N(\mu_i, I_d)$ with unknown weights and centers, the goal is to learn the density f by producing a proper (also a k-GM) estimator $\hat{f}$, which is close to the true density in the sense that, say, the Hellinger distance satisfies $E[H(\hat{f}, f)] \leq \epsilon$. Assume that k is a constant. It is known that [13] (a) Information theoretically, the optimal sample complexity is $n = \Theta(d/\epsilon^2)$. (b) Computationally, for $k \leq 2$, there are polynomial-time (in d and n) algorithms that achieve the optimal sample complexity; however, for $k \geq 3$, the best polynomial-time algorithms require a sample size $n = \Omega(d/\epsilon^4)$. Is there a computational barrier that emerges for learning a 3-component Gaussian mixture model?

4. (Raised by Dmitriy Kunisky) Suppose $X_1, \dots, X_n$ are i.i.d. standard Gaussian vectors in R^d. If $d \to \infty$, how quickly can n grow with d such that there is an ellipsoid passing through all of the X_i? This problem was posed by Saunderson, Parrilo, and Willsky (2013), who conjectured a sharp threshold on the scale $n \sim d^2$: if $n \leq (\frac{1}{4} - \epsilon)d^2$ then with high probability (w.h.p.) the X_i can be interpolated by an ellipsoid, while if $n \geq (\frac{1}{4} + \epsilon)d^2$ then w.h.p. they cannot (in both cases for $\epsilon > 0$ fixed as $d \to \infty$. However, the best known rigorous results do not achieve either threshold. The best known negative result follows by a simple linear algebra argument, showing that if $n \geq \frac{1}{2}d^2$ then w.h.p. the X_i cannot be interpolated by an ellipsoid. The best known positive result, due to Venkat, Turner, and Wein (2022), shows that if $n \leq d^2/\mathrm{polylog}(d)$ then w.h.p. the X_i can be interpolated by an ellipsoid. Improving either of these results to approach closer to the conjectured sharp threshold is an open problem.

5. (Raised by Jiaming Xu) Planted minimum spanning tree model. Consider a complete weighted graph with n vertices and edge weights i.i.d. $\exp(1)$. Frieze (1985) proves that the expected total weight of the minimum spanning tree converges to $\zeta(3) = \sum_{i=1}^{\infty} i^{-3}$ as $n \to \infty$. Now, let's consider a planted model, where T^* is the random minimum spanning tree described as above. Then we observe a new complete weighted graph W such that W_e is independently distributed according to $\exp(\lambda)$ if the edges e is contained in T^* and $\exp(1/n)$ otherwise. The goal is to estimate T^* based on the observation of W. Note that the maximum likelihood estimator reduces to the minimum spanning tree $T_{\min}$ on W. An open question is to determine the asymptotic overlap between $T_{\min}$ and T^* as a function of λ. Note that the random minimum spanning tree T^* is different from the uniform spanning tree in the complete graph. Moreover, T^* can be recursively constructed using the Kruskal's algorithm.

6 Acknowledgement

The organizers of the workshop wish to thank the BIRS CMO organization for sponsoring the workshop, and for the staff to provide a wonderful environment for the exchange of ideas and collaboration. In addition, they appreciate that all fifteen in-person participants stayed engaged throughout the week.

References

[1] E. Abbe. Community detection and stochastic block models: recent developments. *J. Machine Learning Research*, 18:1–86, 2018. arXiv 1703.10146.

[2] T. Antunović, E. Mossel, and M.Z. Racz. Coexistence in preferential attachment networks. *Combinator. Probab. Comp.*, 25:797–822, 2016.

[3] P. Berenbrink, B. Krayenhoff, and F. Mallmann-Trenn. Estimating the number of connected components in sublinear time. *Inform. Process. Lett.*, 114(11):639–642, 2014.

[4] M. Brennan, G. Bresler, and W. Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. *J. Machine Learning Research (COLT)*, 2018. arXiv 1806.07508.

[5] J.A. Chapman, I. Ho, S. Sunkara, S. Luo, G.P. Schroth, and D.S. Rokhsar. Meraculous: de novo genome assembly with short paired-end reads. *PloS one*, 6(8):e23501, 2011.

[6] Pinhan Chen, Chao Gao, and Anderson Y Zhang. Optimal full ranking from pairwise comparisons. *The Annals of Statistics*, 50(3):1775–1805, 2022.

[7] Zongchen Chen, Elchanan Mossel, and Ilias Zadik. Almost-linear planted cliques elude the metropolis process. *arXiv preprint arXiv:2204.01911*, 2022.

[8] M. Chertkov, L. Kroc, F. Krzakala, M. Vergassola, and L. Zdeborová. Inference in particle tracking experiments by passing messages between images. *Proceedings of the National Academy of Sciences*, 107(17):7663–7668, 2010.

[9] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(03):265–298, 2004.

[10] G. Cormode and N. Duffield. Sampling for big data: a tutorial. In *Proc. 20th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 1975–1975. ACM, 2014.

[11] D. Cullina and N. Kiyavash. Exact alignment recovery for correlated Erdos Renyi graphs. *arXiv preprint arXiv:1711.06783*, 2017.

[12] Souvik Dhara, Julia Gaudio, Elchanan Mossel, and Colin Sandon. Spectral recovery of binary censored block models*. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3389–3416. SIAM, 2022.

[13] Natalie Doss, Yihong Wu, Pengkun Yang, and Harrison H. Zhou. Optimal estimation of high-dimensional Gaussian mixtures. *to appear in* The Annals of Statistics, Feb 2020. arXiv:2002.05818.

[14] O. Frank. Estimation of the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*, 5(4):177–188, 1978.

[15] Julia Gaudio and Nirmit Joshi. Community detection in the hypergraph sbm: Optimal recovery given the similarity matrix. *arXiv preprint arXiv:2208.12227*, 2022.

[16] Julia Gaudio, Miklos Z Racz, and Anirudh Sridhar. Exact community recovery in correlated stochastic block models. *arXiv preprint arXiv:2203.15736*, 2022.

[17] L.A. Goodman. On the estimation of the number of classes in a population. *Ann. Math. Statistics*, 20:572–579, 1949.

[18] Bruce Hajek and Suryanarayana Sankagiri. Community recovery in a preferential attachment graph. *IEEE Transactions on Information Theory*, 65(11):6853–6874, 2019.

[19] J. Jordan. Geometric preferential attachment in non-uniform metric spaces. *Electronic Journal Probability*, 18(8):1–15, 2013.

[20] Xiaohan Kang and Bruce Hajek. Lower bounds on information requirements for causal network inference. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 754–759. IEEE, 2021.

[21] Cheng Mao, Yihong Wu, Jiaming Xu, and Sophie H Yu. Random graph matching at otter's threshold via counting chandeliers. *arXiv preprint arXiv:2209.12313*, 2022.

[22] C. Moore. The computer science and physics of community detection: Landscapes, phase transitions, and hardness. *Bulletin European Association for Theoretical Computer Science*, 121, Feb 2017. arXiv 1702.00467.

[23] Tianhao Wang, Xinyi Zhong, and Zhou Fan. Universality of approximate message passing algorithms and tensor networks. *arXiv preprint arXiv:2206.13037*, 2022.

[24] Ziao Wang, Ning Zhang, Weina Wang, and Lele Wang. On the feasible region of efficient algorithms for attributed graph alignment. *arXiv preprint arXiv:2201.10106*, 2022.

[25] Alexander S Wein. Average-case complexity of tensor decomposition for low-degree polynomials. *arXiv preprint arXiv:2211.05274*, 2022.

[26] E. M. Wright. Graphs on unlabelled nodes with a given number of edges. *Acta Mathematica*, 126(1):1–9, 1971.

[27] Qian Yu and Yury Polyanskiy. Ising model on locally tree-like graphs: Uniqueness of solutions to cavity equations. *arXiv preprint arXiv:2211.15242*, 2022.