On Thursday afternoon, four problems were posed, and this led to four active discussion groups who worked through the last two days of the workshop and presented their findings on Friday. Here we list the problems and discussion outcomes.

1. Statistical procedure to compare different rankings

Posed by D. Gleich and A. Langville

Problem description
Suppose two different methods applied to a graph $G = (V, E)$ give rankings σ_1, σ_2 of the nodes. For instance, σ_1 is the output produced by a new ranking method, and σ_2 is the output produced by PageRank. In the current literature, one present some statistics (for instance, Kendall's tau) to measure the similarity of σ_1 and σ_2, and thereby make claims on the similarity between the new method and PageRank. But this is just a measure of similarity of the output on one graph.

Question
Can we find rigorous statistical tests to compare f_1 and f_2?

Discussions
The question can be phrased more concretely as follows: given a graph $G = (V, E)$, let $f : V \to \mathbb{R}$ be a centrality measure, that is, a function which is invariant under graph isomorphisms. This function induces the ranking on the node. (For example, the PageRank function, the node degree function). Consider the behavior of this function over all possible graph: this gives us a distribution over all possible rankings $\mathbb{S}_n$ where $n = |V|$ is the number of nodes. That is, f induces a ranking function $f^R : G \to \mathbb{S}_n$. Now the question is equivalent to finding statistical tests to compare the distributions of the two functions f_1^R, f_2^R over all possible graphs G on n nodes.

We identified two issues: firstly, one may want to look at a subset of these graphs only. That is, given an instance of the graph (eg: the graph G), we only compare f_1^R, f_2^R over a subset of graphs on n nodes which are "similar" to G. Then the question is what is a reasonable subset of graphs to look at? We suggested, as an example, that one can look at random graphs of the same degree distribution. This can be obtained by simulations.

Secondly, one may want to model the distribution of f^R as one of the location-scale family, where a scale parameter measures the spread, and the location parameter corresponds to the "median permutation", that is, the permutation that minimizes the sum of the distances to all the other permutations on the range of f^R. This problem is NP-hard in general, but if we make assumptions about the classes of the permutations to be summarized, it's potentially tractable. Clearly the choice of metric on $\mathbb{S}_n$ is important. For example, if $f_1^R(G) = \sigma$ and $f_2^R(G) = \sigma'$ are distinctively different rankings (for example, one is a reverse of the other), then we want to say that f_1^R and f_2^R are very different (on G), while, if σ and σ' only differ by one transposition, then we want to say that f_1^R and f_2^R are very similar. Anne Shiu and Jason Morton pointed out that they encountered a similar problem of detecting permutations with cycles from random permutation in genetics, in which themselves and their coauthors have solved using techniques from algebraic geometry.

Continuing with a geometric approach, we noted that one can inscribe the permutahedron in the smallest sphere containing it. Then we can view the distribution over the permutahedron as distribution over the sphere, and one can measure the distance between two permutations by their geodesic distance on the sphere. This gives a natural metric on $\mathbb{S}_n$. Then, suppose that f_1^R and f_2^R have the same scale parameter. To compare their location parameters, note that if we assume rotational symmetry of the induced distributions, the permutation closest to the intrinsic mean approximates the minimizer of average Kendall tau distance. Thus an intrinsic or directional statistics based test for equality of the means can be used to decide if two distributions, and hence ranking procedures, are indistinguishable.

Outcome summary
Through the discussions, we learned of related work of Morton and Shiu. The problem was broken into smaller questions in which we made progress in solving, and naturally raised open questions one can consider and follow up.

References
Problem write-up by D. Gleich and A. Langville.

J. Morton, L. Pachter, A. Shiu, B. Sturmfels, O. Wienand, Convex rank tests and semigraphoids, SIAM J. Discrete Math., 23 (2009), no. 3, pp. 11171134

Analyzing and modeling rank data, John I Marden, Chapman & Hall, 1995.

2. Recovering distribution over $\mathbb{S}_n$ given partial information on λ-coefficients

Posed by Srikanth Jagabathula and Devavrat Shah

Let $\mathcal{F} := \{f : \mathbb{S}_n \to \mathbb{R}_+\}$ be the space of functions with non-negative values on $\mathbb{S}_n$. Let $f \in \mathcal{F}$ be the (unknown) function we are interested in learning. Suppose that we know the Fourier transform of f: $\hat{f}_\lambda = \sum_{\sigma \in \mathbb{S}_n} f(\sigma)\rho_\lambda(\sigma)$ (note that ρ_λ denotes everything up to λ, not just one single irreducible representation). Let $|f|_0$ denote the sparsity of f, that is, the support of f on $\mathbb{S}_n$:

$$|f|_0 := \mathsf{supp}(f) = \{\sigma \in \mathbb{S}_n : f(\sigma) \neq 0\}.$$

Question
Suppose that $|f|_0 = K$. For what values of K is the recovery of f from $\hat{f}_\lambda$ not possible?

Background
The authors who posed this problem have proved the following bound on K

Theorem 1. *With respect to random model $R(K,T)$, the probability of error is uniformly bounded away from 0 for all n large enough and any λ, if*

$$K \geq \frac{3D_\lambda^2}{n\log n}\left[\log\left(\frac{D_\lambda^2}{n\log n} \vee T\right)\right]. \tag{1}$$

However, they conjectured that a sharper bound $K \approx D_\lambda$ can be obtained.

Below is a detailed write-up by one of the authors, Srikanth Jagabathula, starting with some necessary information theory preliminaries and ending with a proof of the above theorem.

2.1. Information Theory Preliminaries.

Here we recall some necessary Information Theory preliminaries. Further details can be found in the book by Cover and Thomas.

Consider a discrete random variable X that is uniformly distributed over a finite set $\mathscr{X}$. Let X be *transmitted* over a *noisy* channel to a receiver; suppose the receiver receives a random variable Y, which takes values in a finite set $\mathscr{Y}$. Essentially, such "transmission over noisy channel" setup describes any two random variables X, Y defined through a joint probability distribution over a common probability space.

Now let $\hat{X} = g(Y)$ be an estimation of the transmitted information that the receiver produces based on the observation Y using some function $g : \mathscr{Y} \to \mathscr{X}$. Define probability of error as $p_{\text{err}} = \Pr(X \neq \hat{X})$. Since X is uniformly distributed over $\mathscr{X}$, it follows that

$$p_{\text{err}} = \frac{1}{|\mathscr{X}|}\sum_{x \in \mathscr{X}} \Pr(g(Y) \neq x | x). \tag{2}$$

Recovery of X is called asymptotically reliable if $p_{\text{err}} \to 0$ as $|\mathscr{X}| \to \infty$. Therefore, in order to show that recovery is not asymptotically reliable, it is sufficient to prove that p_{err} is bounded away from 0 as $|\mathscr{X}| \to \infty$. In order to obtain a lower bound on p_{err}, we use Fano's inequality:

$$H(X|\hat{X}) \leq 1 + p_{\text{err}} \log |\mathscr{X}|. \tag{3}$$

Using (3), we can write

$$\begin{aligned} H(X) &= I(X;\hat{X}) + H(X|\hat{X}) \\ &\leq I(X;\hat{X}) + p_{\text{err}} \log|\mathscr{X}| + 1 \\ &\overset{(a)}{\leq} I(X;Y) + p_{\text{err}} \log|\mathscr{X}| + 1 \\ &= H(Y) - H(Y|X) + p_{\text{err}} \log|\mathscr{X}| + 1 \\ &\leq H(Y) + p_{\text{err}} \log|\mathscr{X}| + 1, \end{aligned} \tag{4}$$

where we used $H(Y|X) \geq 0$ for a discrete[1] valued random variable. The inequality (a) follows from the data processing inequality: if we have Markov chain $X \to Y \to \hat{X}$, then $I(X;\hat{X}) \leq I(X;Y)$. Since $H(X) = \log|\mathscr{X}|$, from (4) we obtain

$$p_{\text{err}} \geq 1 - \frac{H(Y)+1}{\log|\mathscr{X}|}. \tag{5}$$

Therefore, to establish that probability of error is bounded away from zero, it is sufficient to show that

$$\frac{H(Y)+1}{\log|\mathscr{X}|} \leq 1-\delta, \tag{6}$$

for any fixed constant $\delta > 0$.

2.2. **Proof of the theorem.**

Proof. Our goal is to show that when K is large enough (in particular, as claimed in the statement of Theorem 1), the probability of error of *any* recovery algorithm is uniformly bounded away from 0. For that, we first fix a recovery algorithm, and then utilize the above setup to show that recovery is not asymptotically reliable when K is large. Specifically, we use (6), for which we need to identify random variables X and Y.

To this end, for a given K and T, let f be generated as per the random model $R(K,T)$. Let random variable X represent the support of function f i.e., X takes values in $\mathscr{X} = S_n^K$. Given λ, let $\hat{f}(\lambda)$ be the partial information that the recovery algorithm uses to recover f. Let random variable Y represent $\hat{f}(\lambda)$, the $D_\lambda \times D_\lambda$ matrix. Let $h = h(Y)$ denote the estimate of f, and $g = g(Y) = \mathsf{supp}h$ denote the estimate of the support of f produced by the given recovery algorithm. Then,

$$\begin{aligned} \Pr(h \neq f) &\geq \Pr(\mathsf{supp}h \neq \mathsf{supp}f) \\ &= \Pr(g(Y) \neq X). \end{aligned} \tag{7}$$

Therefore, in order to uniformly lower bound the probability of error of the recovery algorithm, it is sufficient to lower bound its probability of making an error in recovering the support of f. Therefore, we focus on

$$p_{\text{err}} = \Pr(g(Y) \neq X).$$

It follows from the discussion in Section 2.1 that in order to show that p_{err} is uniformly bounded away from 0, it is sufficient to show that for some constant $\delta > 0$

$$\frac{H(Y)+1}{\log|\mathscr{X}|} \leq 1-\delta. \tag{8}$$

[1]The counterpart of this inequality for a continuous valued random variable is not true. This led us to study the limitation of recovery algorithm over model $R(K,T)$ rather than $R(K,\mathscr{C})$.

Observe that $|\mathscr{X}| = (n!)^K$. Therefore, using Stirling's approximation, it follows that

$$\log|\mathscr{X}| \;=\; (1+o(1))Kn\log n. \tag{9}$$

Now $Y = \hat{f}(\lambda)$ is a $D_\lambda \times D_\lambda$ matrix. Let $Y = [Y_{ij}]$ with $Y_{ij}, 1 \le i,j \le D_\lambda$, taking values in $\{1,\ldots,KT\}$; it is easy to see that $H(Y_{ij}) \le \log KT$. Therefore, it follows that

$$\begin{aligned} H(Y) \;&\le\; \sum_{i,j=1}^{D_\lambda} H(Y_{ij}) \\ &\le\; D_\lambda^2 \log KT \;=\; D_\lambda^2\left(\log K + \log T\right). \end{aligned} \tag{10}$$

For small enough constant $\delta > 0$, it is easy to see that the condition of (8) will follow if K satisfies the following two inequalities:

$$\frac{D_\lambda^2 \log K}{Kn\log n} \le \frac{1}{3}(1+\delta) \quad\Leftarrow\quad \frac{K}{\log K} \ge \frac{3(1-\delta/2)D_\lambda^2}{n\log n}, \tag{11}$$

$$\frac{D_\lambda^2 \log T}{Kn\log n} \le \frac{1}{3}(1+\delta) \quad\Leftarrow\quad K \ge \frac{3(1-\delta/2)D_\lambda^2 \log T}{n\log n}. \tag{12}$$

In order to obtain a bound on K from (11), consider the following: for large numbers x, y, let $y = (c+\varepsilon)x\log x$, for some constants $c, \varepsilon > 0$. Then, $\log y = \log x + \log\log x + \log(c+\varepsilon)$ which is $(1+o(1))\log x$. Therefore,

$$\frac{y}{\log y} \;=\; \frac{c+\varepsilon}{1+o(1)}x \;\ge\; cx, \tag{13}$$

for $x \to \infty$ and constants $c, \varepsilon > 0$. Also, observe that $y/\log y$ is a non-decreasing function; hence, it follows that for $y \ge (c+\varepsilon)x\log x$, $y/\log y \ge cx$ for large x. Now take $x = \frac{D_\lambda^2}{n\log n}$, $c = 3$, $\varepsilon = 1$ and $y = K$. Note that $D_\lambda \ge n$ for all λ of interest; therefore, $x \to \infty$ as $n \to \infty$. Hence, (11) is satisfied for the choice of

$$K \;\ge\; \frac{4D_\lambda^2}{n\log n}\left(\log\frac{D_\lambda^2}{n\log n}\right). \tag{14}$$

From (8), (11), (12), and (14) it follows that the probability of error of any algorithm is at least $\delta > 0$ for n large enough and any λ if

$$K \;\ge\; \frac{4D_\lambda^2}{n\log n}\left[\log\left(\frac{D_\lambda^2}{n\log n} \vee T\right)\right]. \tag{15}$$

□

Outcome summary

There were some discussions on this problem amongst Konder, Orrison and Jagabathula, however, there was no definitive outcome. The problem remains opened. Further description of the problem can be found at the AIM workshop website.

References

Problem write-up by Srikanth Jagabathula and Devavrat Shah

3. Rademacher complexity of kernel classes on $\mathbb{S}_n$

Posed by Risi Kondor

Question
What is the Rademacher complexity of kernel classes on $\mathbb{S}_n$?

Background
A kernel K on $\mathbb{S}_n$ is a symmetric positive semidefinite function $K : \mathbb{S}_n \times \mathbb{S}_n \to \mathbb{R}$. For example, $K(\sigma_1, \sigma_2)$ is the diffusion kernel induced by adjacent transpositions on the Caley Graph of S_n. Given a set $F := \{f : X \to \mathbb{R}\}$ consisting real-valued functions on some space X, the Rademacher complexity of F is defined as

$$R_m(F) := E_{X_i,\epsilon_i}[\sup_{f\in F} \frac{1}{m}\sum_{i=1}^{m} \epsilon_i f(X_i)]$$

where ϵ_i are i.i.d Bernoulli. Reproducing kernel Hilbert spaces (RKHS) induced by K is the set of functions $F_K = \{f : S_n \to \mathbb{R} : f(x) = \sum_{i=1}^{m} \alpha_i K(x, x_i)$ for some m, x_i, α_i. It is known that if $F_K^B = \{f \in F_K : \|f\|_K \leq B\}$, then

$$E[R_m(F_K^B)] \leq \frac{2B}{m} E(\sqrt{\sum_i K(X_i, X_i)})$$

Discussions
There were discussions on the role of kernel in ranking. Risi Kondor gave some motivations in his talk, and also in the following small example: suppose we have n partial rankings: $\sigma_1, ...\sigma_n$ of n different people ranking on movies. Now suppose we have a new user, who told us that she prefers A to B to C. Then, taking $A > B > C$ as the input, should we recommend $D > E$ or $E > D$? This problem leads to the problem of learning on permutations with sparse data, in which one approach is regularization. Specifically, a kernel induce a RKHS, which gives a sense of smoothness, and has been used for regularization for other set of functions in machine learning. In this particular instance, the space of functions we want to regularize is the space of functions on the symmetric group $\{f : S_n \to \mathbb{R}\}$. By regularization, we mean replacing the original loss function $\sum_i L(f(\sigma_i))$ by its regularized version $R(f) := \sum_i L(f(\sigma_i)) + \|f\|_{\mathcal{H}}^2$ where $\|f\|_{\mathcal{H}}^2$ is the norm of f induced by the norm on the RKHS corresponding to a chosen kernel K.

For example, $L(f(\sigma_i))$ can be the multi class hinge loss. This is a multi class support vector machine (SVM). Since f is in the RKHS, the solution is of the form $f(\sigma) = \sum_i \alpha_i K(\sigma_i, \sigma)$. This approach has the advantage that the kernel ignores the symmetric group $\mathbb{S}_n$, and coverts the problem to a usual SVM problem.

In order to get a bound on the convergence rate of the estimator, one needs a VC or Rademacher bound on the RKHS. Here, Lester Mackay pointed out that from lecture notes of Peter Bartlett, the problem $\min_{f\in\mathcal{H}} \frac{1}{n}\sum_i L(f(\sigma_i)) + \|f\|_{\mathcal{H}}^2$ is equivalent to $\min_{\|f\|_H^2 \leq B} \frac{1}{n}\sum_i L(f(\sigma_i))$. Then one has

$$\hat{R}_n(F_B) \leq \frac{B}{\sqrt{n}}\sqrt{1/n\sum_i K(\sigma_i, \sigma_i)}, \quad \text{and } R_n(F_B) \leq B\sqrt{\sum_{i=1}^{\infty} \lambda_i/n}.$$

This result, together with the bound

$$E \sup_{f \in \mathcal{F}} [E(L(f)) - 1/m \sum_i L(f, \sigma_i)] \leq 2R_n(\mathcal{F}_B)$$

in Konder's notes leads to a solution to the problem.

Outcome summary
With the discovery of a result from Peter Bartlett's lecture notes coupled with Konder's own results, the problem has been solved. Konder said this will appear in his next paper.

References
Stat281B Lecture Notes, Spring 2009, Peter Bartlett

4. Non convex surrogates for ranking (and resulting information patterns)

Posed by Yoram Singer

The goal is to improve on the bound appeared in the joint work of Peter Bartlett with Sylvain Arlot, Mike Jordan, Jon McAuliffe and Mikhail Traskin.

Background
Suppose we have a loss function we are interested in minimizing. For example:

$$\sum_{i,j} a_{ij} 1_{f(x_i)-f(x_j)<0}.$$

Finding f^* which optimizes the loss is a NP hard problem. So we consider minimizing a surrogate (a relaxation of the loss), for example

$$\sum_{i,j} a_{ij} 1_{f(x_i)-f(x_j)<0}.$$

Suppose that the minimizer of the surrogate function is $\hat{f}$. The aforementioned work of Bartlett and his coauthors gave conditions for choices of the surrogate function in which one have $\hat{f} = f^*$ (in some asymptotic sense).

Outcome summary
Peter Bartlett gave a talk on his result in classification. Lester Mackay gave a talk on his work in a similar problem in ranking context. However, due to time constraint, there was little further discussions on this topic. The problem remains open.

References
Sylvain Arlot, Peter Bartlett, Mike Jordan, Jon McAuliffe and Mikhail Traskin
Convex methods for classification. IMS Medallion Lecture. June 2008.

5. Connection between AHP (analytic hierarchical process) and Borda Count

Posed and commented on by Donald Saari

Question
What is the relationship between AHP and pairwise voting?

Discussions
Recall that AHP works with matrices with weights

$$\begin{bmatrix} 1 & a_{12} & a_{13} & \dots \\ \frac{1}{a_{12}} & 1 & a_{23} & \dots \\ \frac{1}{a_{13}} & \frac{1}{a_{23}} & 1 & \dots \end{bmatrix}$$

Can we connect these weights in AHP and pairwise voting? The answer is *yes*. Suppose, that i, j are candidates, and candidate i gets one million votes, candidate j gets 500 thousand votes. Then by defining a_{ij} to be is the exponential of the difference in tally

$$a_{ij} = \exp(1000000 - 500000) = \exp(500000),$$

we can form a matrix of weights in AHP. The consistency condition in AHP ($a_{ij}a_{jk} = a_{ik}$) naturally corresponds to the consistency in voting by taking log:

((tally for i) - (tally for j)) $\times$ (tally for j - tally for k) = (tally for i - tally for k).

So this means a question on AHP can be translated into a question on social choice (by log) and then mapped back (by exp).

Outcome
This is an important revelation, and Donald Saari plans to publish a paper on this topic.