Submitted to the Annals of Applied Statistics

DON'T MIND THE (EIGEN) GAP

By Song Wang* and Karl Rohe*

Department of Statistics, University of Wisconsin - Madison

Pengsheng Ji and Jiashun Jin have collected and analyzed a fun and fascinating dataset that we are eager to use as an example in a course on Statistical Network Analysis. In this comment, we partition the core of the paper citation graph and interpret the clusters by analyzing the paper abstracts using bag-of-words. Under the Stochastic Blockmodel (SBM), the eigengap reveals the number of clusters. We find several eigengaps and that there are still clusters beyond the largest eigengap. Through this illustration, we argue against a simplistic interpretation of model selection results from the Stochastic Blockmodel (SBM) literature. In short, don't mind the gap.

Pengsheng Ji and Jiashun Jin [2] have collected and analyzed three networks that we are eager to use in classes on statistical network analysis. As statisticians, we all have a contextual understanding of the processes that these networks describe, often down to individualized knowledge about the nodes and their relationships. The individuals are our colleagues, mentors, and friends; some of the papers we have studied for exams and for research; these papers motivate our own work and the work of our colleagues. As such, we claim that the contributions of this paper come not just from a deeper understanding of citations and co-authorship, but rather from providing a canonical example for young researchers to begin studying network analysis. The future of statistical network analysis is not merely about predicting node labels or identifying missing edges. There are many other, potentially more interesting questions and this data set provides a playground to explore. For example, how do ideas spread through a social network? Or, what is the relationship between theory and practice? Because of our relationship to the pieces of these networks, these networks provide a way for students to start thinking about these complex problems. As such, this network provides a reality check. For those that pursue these issues, One must be careful to draw inferences too wide from this data; there are biases induced by the "boundary effects" of this network due to sampling, as discussed in the paper.

*This research is supported by NSF grant DMS-1309998 and ARO grant W911NF-15-1-0423.

Keywords and phrases: Networks, Spectral Clustering, Text Analysis, Eigen Gap.

The following sentence from Ji and Jin is a starting point for this comment:

> The elbow point of the scree-plot [of Figure 2] may be at the 3rd, 5th, or 8th largest eigenvalue, suggesting that there may be 2, 4, or 7 communities.

In particular, we are troubled by the implication that we must choose the number of communities, or that there is one right answer.

In this comment, we study two different clusterings of the paper citation network; Here, the nodes are papers (not authors). We interpret the clustering via a *post hoc* bag-of-word analysis of the abstracts. The abstracts are not used to detect the clusters, but rather to interpret the clusters. Similar to the findings in Ji and Jin [2] that many communities of statistician networks consists of authors sharing the research fields, we find that in both clusterings, the papers are divided by research topics. We present the partition for $K = 11$ and $K = 20$ clusters and argue that neither of these choices should be interpreted as "the correct" choice of K. For both choices of K, each cluster has:

1. more connections within the cluster than to all other clusters combined (Tables 1 and 3) and
2. a coherent description from the bag-of-word analysis (Tables 2 and 4).

Moreover, just because we find a partition by research topic does not preclude the possibility of other good partitions. For example, perhaps authors are more likely to cite authors in their own department. Partitioning by department could be unrelated to the partition by research topic. Such a partition would not be wrong, but perhaps it is not the strongest partition in the data. We must disabuse ourselves of the notion of "the correct partition." Instead, there are several "reasonable partitions"; some of these clusterings might be consistent with one another (as might be imagined in a hierarchical clustering), others might not be consistent. Our code and the bag-of-words representation of the abstracts will be made available at `https://github.com/orgs/RoheLab/`.

1. Partitioning the core of the citation graph.

A set of four R libraries dramatically facilitate the data analysis below. `igraph` for handling networks [4], `Matrix` for handling sparse matrices [5], `tm` for text processing [7], and `rARPACK` for fast eigen computations of sparse matrices [6].

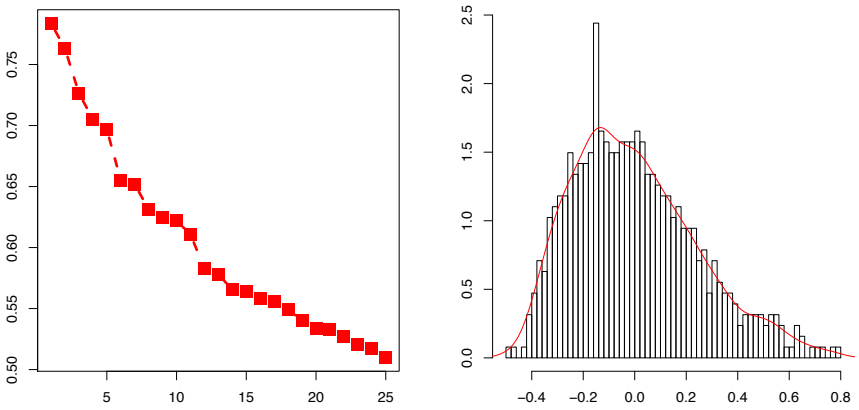
1.1. *Processing the graph.*

Citations are directed connections. For simplicity, these edges were symmetrized. The resulting network has 3248 papers and 5712 edges. Many large networks have a core-periphery structure; the

core contains a subset of the nodes which are highly connected and the periphery contains low degree nodes that are weakly connected to the core. In our analysis below, we focus on understanding the core of the graph. The computations below are performed on the 4-core of the graph.[1] This reduces the number of papers from 3248 to 635.

Using `Matrix`, we constructed $\tilde{A}_\tau = D_\tau^{-1/2} A D_\tau^{-1/2}$, where $[D_\tau]_{ii} = \tau + \sum_\ell A_{i\ell}$ and $\tau = \sqrt{\frac{1}{n}\sum_{ij} A_{ij}}$. Then, we computed the leading 30 eigenvalues and eigenvectors of $\tilde{A}_\tau$ with `rARPACK`.[2] These eigenvalues are displayed in a screeplot in the left panel of Figure 1. All of the gaps in this screeplot are small, suggesting that there is not a clear choice for K, the number of clusters. We first explore the choice of $K = 11$ below. Because the dimension of $\tilde{A}_\tau$ is not too large, we can also compute the full eigendecomposition; the right panel of Figure 1 gives a histogram of all 635 eigenvalues. Notice that there is not a clear separation of the leading eigenvalues.

Fig 1: Display of the top 25 singular values (left) and the histogram of all the eigenvalues (right) of the degree weighted adjacency matrix $\tilde{A}_\tau$.



Let $X \in R^{635\times K}$ be the matrix made up of leading K eigenvectors. Define $X^* \in R^{635\times K}$ to contain the row normalized version of X; $X_i^* \leftarrow X_i/\sqrt{\sum_j X_{ij}^2}$ where X_i and X_i^* are the ith rows of the respective matrices.[3] Run k-means on the rows of X^*. This algorithm is called RSC as in [1].

[1] A basic algorithm for finding the 4-core removes all nodes with degree less than four (and any edges connected to these nodes). Then, this step is iterated until convergence.

[2] When using a sparse eigen solver like ARPACK, it is a good idea to compute more eigenvectors than you plan to use. This makes the computations more stable.

[3] SCORE uses a normalization step that is slightly different. Without any normalization step, the largest cluster often contains more than 95% of the nodes in the graph. Both the

1.2. *Processing the abstracts.* To interpret these clusters, we represented the abstracts in their bag-of-word representation using a text mining package called `tm` in R. We did some initial cleaning by removing the stopwords, numbers, and punctuations through setting certain parameters; and we also combined some plural words with ending 's' and past time verbs with ending 'ed' by writing some regular expressions. After this, there were 5529 unique words in the abstracts of the 635 papers in the 4-core. Eliminating words that appear in fewer than 10 papers leaves 793 unique words.

In the end, we have $M \in \{0,1\}^{635\times 793}$ with $M_{ij} = 1$ if and only if paper i contains word j in the abstract and otherwise 0.Using the 11 clusters of papers from RSC, define $p \in R^{11\times 793}$, where $p_{u\ell}$ is the proportion of abstracts in cluster u that contain word ℓ. Define $\tilde{p} \in R^{11\times 793}$ so that $p_{u\ell}$ is the proportion of abstracts *outside* of cluster u that contain word ℓ. For each cluster, Table 2 reports the words that have the largest values in

$$vst(p) - vst(\tilde{p}), \quad \text{where } vst(p) = arcsin\sqrt{p}$$

is a variance stabilizing transformation for the proportions.

2. Interpreting the results. A summary of the clusters found from Section 1.1 are shown in Table 1.

TABLE 1

Summary of $K = 11$ Clusters discovered by RSC on the 4-core of the Paper Citation Network. `Size` *gives the number of papers in each cluster. The sums of degrees for nodes in each cluster are divided into* `In` *and* `Out` *two parts.*

id	Size	In	Out	id	Size	In	Out
1	140	1350	287	7	44	222	41
2	84	788	57	8	41	220	68
3	80	426	136	9	40	290	29
4	65	446	75	10	23	114	36
5	57	372	123	11	15	64	8
6	46	340	34				

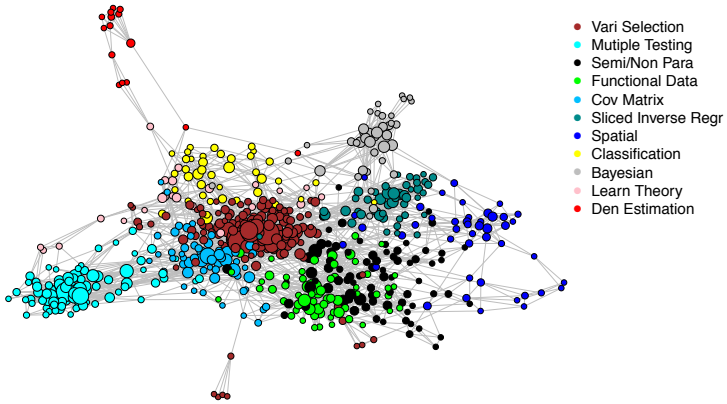
The words from the abstracts facilitate the interpretations here. Based on the largest elements in $vst(p) - vst(\tilde{p})$, we have named the clusters *variable selection, multiple testing, semi-/non-parametric* etc. in the second column of Table 2. Figure 2 gives a visualization of the communities in the 4-core network, where the nodes are colored by the estimated community labels. This figure was generated in `igraph` with layout as fruchterman.reingold.

normalization here and the normalization in SCORE provide a substantial improvement in the balance of the clusters.

TABLE 2
Summary of the 11 clusters discovered by RSC in paper citation network (635 nodes). The representative words are chosen by the criteria in Equation (1.2).

id	name	top five representative words for each cluster
1	Vari Selection	lasso, selection, variable, penalty, oracle
2	Mutiple Testing	false, discovery, testing, hypotheses, rate
3	Semi/Non Para	asymptotic, semiparametric, nonparametric, additive, quantile
4	Functional Data	functional, principal, scalar, data, component
5	Cov Matrix	matrix, covariance, matrices, graphical, definite
6	Sliced Inverse Regr	reduction, dimension, sliced, inverse, central
7	Spatial	spatial, computational, predictive, maximum, likelihood
8	Classification	classification, learning, machine, minimization
9	Bayesian	dirichlet, process, posterior, prior, computation
10	Learn Theory	confidence, coverage, wavelet, construct, mean
11	Den Estimation	nonparametric, density, error, measurement, kernel

Fig 2: Display of the 11 communities found by RSC in the 4-core part of Paper Citation Network. Nodes from different communities are colored diferently, and the size of a node reflects its relative degree.



We chose $K = 11$ by looking at the screeplot in the left panel of Figure 1. This choice of K leads to interpretable clusters. However, the rest of the eigenvalues are not merely noise. The next table repeats the analysis with $K = 20$ (for which there is no eigengap). Notice that for every cluster, `In` > `Out`, suggesting that these clusters are real. Moreover, the representative words show how these clusters are still meaningful. In particular, several clusters from $K = 11$ have been divided into two sub-clusters (e.g. Lasso, Spatial, Learning Theory, Spatial, Non-parametric) and new clusters have emerged (e.g. Design, Quantile regression).

The histogram of the eigenvalues in the right panel of Figure 1 shows

no clear gap that defines the "leading eigenvalues." Don't mind the small eigengaps in plot like the left panel of Figure 1. Just because there is a gap, it doesn't mean that the rest of the eigenvectors are noise.

TABLE 3
*Count of edges staying in and that going out for each of the **20** clusters are discovered by RSC in 4-core of the paper citation network.* `Size`, `In` *and* `Out` *are defined in Table 1.*

id	Name	Size	In	Out	id	Name	Size	In	Out
1	Multiple Testing	77	754	48	11	Bayes	29	130	66
2	Lasso I	62	546	310	12	Spatial I	23	130	23
3	FDA	51	364	74	13	Quantile regression	23	94	34
4	Cov Estimation	46	312	122	14	Learning Theory I	20	112	44
5	Dim Reduction	45	336	32	15	Learning Theory II	20	104	29
6	Lasso II	44	292	262	16	Classification	15	64	40
7	Longitudinal	37	202	102	17	Non-parametric II	14	62	6
8	Forecast	36	130	84	18	Spatial II	11	46	9
9	Bayesian non-para	32	252	27	19	Designs	11	42	8
10	Non-parametric I	29	124	50	20	Semiparametric	10	36	24

TABLE 4
*Summary of the **20** clusters discovered by RSC in the 4-core of the citation network.*

	name	top five representative words (some 10, for interpretation)
1	Multiple Testing	false, discovery, testing, hypotheses, rate
2	Lasso I	selection, variable, lasso, oracle, penalty
3	FDA	functional, principal, scalar, observed, data
4	Cov Estimation	matrix, covariance, matrices, graphical, norm
5	Dim Reduction	reduction, dimension, sliced, inverse, central
6	Lasso II	lasso, high-dimensional, p, sparse, larger
7	Longitudinal	longitudinal, semiparametric, asymptotic, working, data
8	Forecast (in other fields)	differential, article, statistical, dynamic, equation ordinary, compared, modeling, classification, cross-validation
9	Bayesian non-para	dirichlet, process, posterior, prior, computation
10	Non-parametric I	additive, smoothing, spline, backfitting, smooth
11	Bayes	bayesian, prior, posterior, mixture, scale
12	Spatial I(bayes)	spatial, gaussian, covariance, computational, process
13	Quantile regression	quantile, model, regression, resampling, future
14	Learning Theory I	minimization, risk, inequalities, classification, empirical
15	Learning Theory II	confidence, coverage, mean, construct, unknown
16	Classification	data, analysis, classification, discriminant, population
17	Non-parametric II	nonparametric, error, measurement, kernel, setting
18	Spatial II(frequentist)	spatial, marginal, dependence, likelihood, multivariate
19	Designs	orthogonal, constructing, frequentist, construction, empirical likelihood, design, enjoy, seen, flexible
20	Semiparametric	semiparametric, inference, parameter, nuisance, yield

References.

[1] Qin, T. and Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel , *Advances in Neural Information Processing Systems,* pp. 3120–3128.
[2] Ji, P. and Jin, J. (2014). Coauthorship and citation networks for statisticians. *arXiv preprint arXiv:1410.2840.*.
[3] Jin, J. (2015). Fast community detection by SCORE. *The Annals of Statistics*, **43(1)**, 57–89.
[4] Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org
[5] Bates D. and Maechler M. (2016). Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.2-6. https://CRAN.R-project.org/package=Matrix
[6] Qiu Y., Mei J. and authors of the ARPACK library. See file AUTHORS for details. (2016). rARPACK: Solvers for Large Scale Eigenvalue and SVD Problems. R package version 0.11-0. https://CRAN.R-project.org/package=rARPACK
[7] Feinerer I. and Hornik K. (2015). tm: Text Mining Package. R package version 0.6-2. https://CRAN.R-project.org/package=tm

E-mail: songwang@stat.wisc.edu E-mail: karlrohe@stat.wisc.edu