Comment: *Coauthorship and citation networks for statisticians*

Mladen Kolar – University of Chicago Booth School of Business
Matt Taddy – Microsoft Research and Chicago Booth

This article by Ji and Jin (JJ throughout) provides a network analysis, by two experts in the field, on the connections between statisticians and statistics research papers. This is not just an exercise in navel-gazing, but also an opportunity to compare results obtained by different methods in an area we know well: our profession. We think that the article leads to lessons for how we use network models and what data we choose to analyze. First, one can gain insight into a network by considering meta-information for the nodes – in this case, the research paper abstracts. Second, since summary statistics like closeness and betweenness centrality are sensitive to partial network observation, one needs to take care in defining the universe of nodes.

Topic analysis

In our first study, we consider decomposing the abstracts of the articles into latent 'topics', e.g., as in the LDA of Blei et al. (2003). The properties of the citation network can then be considered in light of the topical *content* of the articles. We use the `maptpx` R package (Taddy, 2012) to obtain posterior maximizing point estimates for LDA topics. The `maptpx` package applies the Bayes factors of Taddy (2012) in model selection, and for this data we find that a 15 topic decomposition is optimal. The full code to fit this model (and all of our other analyses) and to generate summaries is at `https://github.com/TaddyLab/statsArticles`.

We focus on topics that have seen their usage change over time – their mean proportions within documents during the first and last five years differ by more than 0.01. Most topics were stable over time, so that only three meet this threshold. These three topics are shown in the left panel of Figure 1. The list of words given for each topic are those with the highest *lift*: within-topic probability over the average corpus-wide occurrence rate.

Topic 1 seems to contain traditional mathematical statistics content, especially for non and semi parametric analysis. The three articles most representative of this topic (i.e., have the highest estimated usage) are

Backfitting and smooth backfitting for additive quantile models (Lee et al., 2010)

Depth weighted scatter estimators (Zuo and Cui, 2005), and

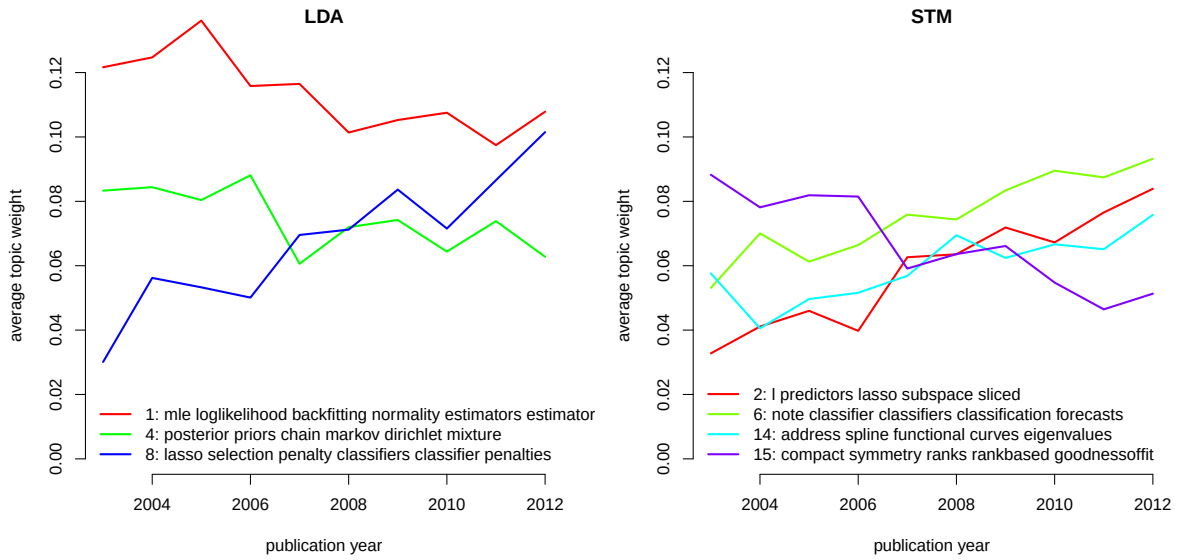

Figure 1: Annual average document-topic weights (that proportion of document devoted to each topic) for topics that saw their usage change by more than 0.1 between the first and last five years of our sample. The left panel shows topics from LDA via `maptpx` and the right shows topics from STM via `stm`.

Estimating invariant laws of linear processes by U-statistics (Schick and Wefelmeyer, 2004). This topic remains relatively popular, but its usage is decreasing over time. The other topic that is decreasing in popularity, topic 4, appears to represent Bayesian analysis. Its three most representative articles are

A conjugate prior for discrete hierarchical log-linear models (Massam et al., 2009)

Bayesian analysis of variable-order, reversible Markov chains (Bacallado, 2011)

Recapture models under equality constraints for the conditional capture probabilities (Farcomeni, 2011)

Over the same time period, we see a dramatic rise in the prevalence of topic 8, which appears to represent the material related to the popular penalized-deviance estimation framework. Its three most representative articles are

The adaptive lasso and its oracle properties (Zou, 2006)

Regularization and variable selection via the elastic net (Zou and Hastie, 2005)

On the adaptive elastic-net with a diverging number of parameters (Zou and Zhang, 2009)

Thus, our topic model shows an increased interest in penalized deviance estimation (and techniques popular in both statistics and machine learning). This rise in popularity for topic 8 has come at the expense of Bayesian analysis and another topic that is roughly interpretable as a different flavor of flexible analysis. Certainly, our personal experience supports the loss of

market share for fully Bayesian analysis; as datasets have grown in size we are often maximizing posteriors (which yields penalized deviance estimators) rather than averaging over them (Taddy, 2012, 2015, are personal examples).

This topic decomposition can be related back to the document networks. For example, we might be interested in which topics tend to generate higher citations. A linear regression of citations per year (the total count divided by the years since publication) onto a year indicator (i.e, a publication-date-specific intercept) and the document-topic weights yields an estimated extra 0.1 citations for per extra 0.1 weight on topic 8, our penalized deviance topic. This has a z-value of around 100, and it is the *only* topic that has a significant effect on citations. Note that this is only the 8th most common topic in our sample, so that the higher cite count is not explainable by prevalence. It is impossible to tell whether the topic is gaining popularity because it tends to lead to citations, or (what seems more likely) if these papers are cited often because the topic is becoming popular.

Our analysis above follows a simple two-stage procedure: we first estimate the topics and then relate them to network statistics (cite counts, i.e., degree in the citation network). A more complex procedure, such as the recent work by Tan et al. (2015), could be used to jointly estimate network communities and text topics (e.g., to see if authors tend to cite within rather than across topics). As a partial step in this direction, we also applied the structural topic model (STM) of Roberts et al. (2013) as implemented in the `stm` package for R (Roberts et al., 2014). STM assumes that document-topic proportions are generated around a linear function of document attributes. Applying STM with both citations-per-year and publication date (annual indicators) as inputs yields the topics in the right panel of Figure 1. We used 15 topics, as in our earlier LDA. Again, a topic that focuses on penalized deviance estimation shows the biggest gains in usage over time. Two other topics are also gaining in popularity: one focused on classification (perhaps including much of the machine learning material that was also included in LDA's topic 8) and another that is tough for us to interpret. There is only one big loser over time: topic 15, which is also tough to interpret.

Across both models, STM and LDA, the point of clear agreement is that penalized deviance estimation is a well-defined topic that is gaining in popularity. As we found in our two stage estimation, this topic is associated with higher citations: the fitted STM has a posterior mean effect of an extra 0.5 citations-per-year for an extra 0.1 usage of the penalized deviance topic.

Sensitivity of the citation networks to journal choice

One issue for JJ's analysis is that it includes only a small portion of the statistics publication venues. Excluded are any applied journals (except for the Applications and Case Studies section of JASA), journals from other fields that use statistics and cite statistical research, and the entire machine learning literature. Certain network statistics, such as the betweenness centrality scores, are especially sensitive to any missingness in the network. However, as already pointed out by JJ, even simple degree statistics (i.e., citation counts) are sensitive to the set of journals considered. We illustrate how the set of journals in the dataset affects the hot papers by removing all Biometrika or all JRSS-B papers. There are total of 751 papers published in Biometrika and 411 in JRSS-B during the period considered.

Table 1: "Hot" papers (alphabetically) identified by degree centrality (Column 2; for citation networks of papers), closeness centrality, and betweenness centrality, when papers published in Biometrika are removed. Numbers in column 2-4 are the ranks (shown when the rank is smaller than or equal to 5).

Paper	Citations	Closeness	Betweenness
Bunea et al. (2007)			4
Candes and Tao (2007)	3		
Fan and Li (2004)		3	
Fan and Lv (2008)			1
Fan and Peng (2004)	4	1	
Genovese and Wasserman (2004)	5		
Huang et al. (2008)			2
Johnstone and Silverman (2004)		4	
Koltchinskii (2006)			5
Meinshausen and Bühlmann (2006)	2	5	
Storey et al. (2004)		2	
Zou (2006)	1		
Zou and Li (2008)			3

Table 1 presents "hot" papers when publications from Biometrika are removed from the network. The top 5 hot articles as measured by the number of citations (i.e., by degree) stays almost the same as in Table 3 of JJ. Zou and Hastie (2005) at 5th place is replaced by Genovese and Wasserman (2004) (which is ranked 6th when considering all 4 journals). When using closeness centrality, Meinshausen and Bühlmann (2006) ranks 5th (originally it was ranked 9th), while Hunter and Li (2005) falls to 6th place. Finally, when using betweenness centrality Bunea et al. (2007) and Koltchinskii (2006) are ranked in the top 5, while initially they were not ranked in the top 10 papers.

Table 2 shows the inferred "hot" papers when publications from JRSS-B are removed from the network. Ranking based on citation count gives us the same ordering as when papers

Table 2: "Hot" papers (alphabetically) identified by degree centrality (Column 2; for citation networks of papers), closeness centrality, and betweenness centrality, when papers published in JRSS-B are removed. Numbers in Column 2-4 are the ranks (only shown when the rank is smaller than or equal to 5).

Paper (Area)	Citations	Closeness	Betweenness
Bickel and Levina (2008)			4
Bunea et al. (2007)			3
Candes and Tao (2007)	3		
Drton and Richardson (2004)		4	
Drton and Perlman (2004)		5	
Fan and Peng (2004)	4	1	
Genovese and Wasserman (2004)	5		
Huang et al. (2006)			2
Hunter and Li (2005)		2	
Johnstone and Silverman (2004)		3	
Koltchinskii (2006)			5
Meinshausen and Bühlmann (2006)	2		
Zou (2006)	1		
Zou and Li (2008)			1

published in Biometrika were removed. However, when looking at ranking based on closeness centrality, we observe that Drton and Richardson (2004) and Drton and Perlman (2004) are ranked 4th and 5th, while before they were not in the top 5. Again, using betweenness centrality Bunea et al. (2007) and Koltchinskii (2006) are ranked in the top 5.

We see here that both betweenness and degree rankings, even at the very top ranks, are sensitive to changes in journal set. This type of behavior is well known (e.g., Borgatti et al., 2006), but seeing it in practice is a good reminder: any conclusions you make about network centrality are valid only for the *observed* network. Thus, JJ's results are relevant only for the research universe defined by this small set of largely theoretical journals. We look forward to forthcoming study by the authors, studying a larger set of journals, that might provide more general insights into our profession.

References

Bacallado, S. (2011). Bayesian analysis of variable-order, reversible Markov chains. *The Annals of Statistics 39*, 838–864.

Bickel, P. J. and E. Levina (2008). Covariance regularization by thresholding. *Ann. Statist. 36*(6), 2577–2604.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Borgatti, S. P., K. M. Carley, and D. Krackhardt (2006). On the robustness of centrality measures under conditions of imperfect data. *Social networks 28*(2), 124–136.

Bunea, F., A. B. Tsybakov, and M. H. Wegkamp (2007). Aggregation for Gaussian regression. *Ann. Statist. 35*(4), 1674–1697.

Candes, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist. 35*(6), 2313–2351.

Drton, M. and M. D. Perlman (2004). Model selection for Gaussian concentration graphs. *Biometrika 91*(3), 591–602.

Drton, M. and T. S. Richardson (2004). Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika 91*(2), 383–392.

Fan, J. and R. Li (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc. 99*(467), 710–723.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol. 70*(5), 849–911.

Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist. 32*(3), 928–961.

Farcomeni, A. (2011). Recapture models under equality constraints for the conditional capture probabilities. *Biometrika 98*(1), 237–242.

Genovese, C. and L. Wasserman (2004). A stochastic process approach to false discovery control. *Ann. Statist. 32*(3), 1035–1061.

Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist. 36*(2), 587–613.

Huang, J. Z., N. Liu, M. Pourahmadi, and L. Liu (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika 93*(1), 85–98.

Hunter, D. R. and R. Li (2005). Variable selection using MM algorithms. *Ann. Statist. 33*(4), 1617–1642.

Johnstone, I. M. and B. W. Silverman (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist. 32*(4), 1594–1649.

Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist. 34*(6), 2593–2656.

Lee, Y. K., E. Mammen, B. U. Park, et al. (2010). Backfitting and smooth backfitting for additive quantile models. *The Annals of Statistics 38*(5), 2857–2883.

Massam, H., J. Liu, A. Dobra, et al. (2009). A conjugate prior for discrete hierarchical log-linear models. *The Annals of Statistics 37*, 3431–3467.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist. 34*(3), 1436–1462.

Roberts, M. E., B. M. Stewart, and D. Tingley (2014). stm: R package for structural topic models. *R package vignette*.

Roberts, M. E., B. M. Stewart, D. Tingley, E. M. Airoldi, et al. (2013). The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Schick, A. and W. Wefelmeyer (2004). Estimating invariant laws of linear processes by u-statistics. *Annals of statistics*, 603–632.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol. 66*(1), 187–205.

Taddy, M. (2012). On estimation and selection for topic models. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*.

Taddy, M. (2015). One-step estimator paths for concave regularization. *arXiv:1308.5623*.

Tan, L. S. L., A. H. Chan, and T. Zheng (2015). Topic-adjusted visibility metric for scientific articles. *arXiv 1502.07190*.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc. 101*(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*, 301–320.

Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist. 36*(4), 1509–1533.

Zou, H. and H. H. Zhang (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist. 37*(4), 1733–1751.

Zuo, Y. and H. Cui (2005). Depth weighted scatter estimators. *Annals of statistics*, 381–413.