Discussion of “Co-authorship and citation networks for statisticians” by Pengsheng Ji and Jiashun Jin

Forrest W. Crawford

Department of Biostatistics, Yale School of Public Health

I congratulate the authors on this valuable contribution to the statistics profession, and for their diligent work collecting the co-authorship and citation datasets upon which their analysis is based. The paper provides a valuable perspective on an important aspect of relationships between papers and between individual statisticians in a few of the most prominent journals in the field. The authors’ analyses of these relationships yields new insights, paves the way for future data collection efforts, and provides a valuable dataset for further analytical exploration. In this brief discussion, I will give a few general comments, questions, and give suggestions for future work. I focus on three areas: author, paper, and journal attributes, selection of authors and journals, and the role of time in the process of research collaboration and citation.

1 The role of author, paper, and journal attributes

Collection of additional meta-data, including paper content and author characteristics (current institution, department, PhD institution, time since PhD, dissertation advisor, etc), could potentially yeild additional insight into the complex process of co-athorship and citation. The characteristics of authors themselves may be important. For example, what proportion of authors are students versus professors? Often the order of authorship matters: usually the first author did most of the work, and the last author is in charge. Middle authors have moderate contributions. How does author order arise from co-athorship arrangements? Is the more senior author usually the last author?

One might expect that co-authorship relationships are most common among authors who have been physically proximate in the past or currently. Co-authorship within the same insitution might be most common. Is the same true for co-authorship in the same department? Can we learn about the collaborative character of academic statistics and biostatistics departments by studying the pattern of collaborations within and between them? Online access to scholarly publications, blogs, and researcher websites have made it increasingly easy to identify potential collaborators all around the world. We might expect the prevalence of collaboration across large geographic distances to become more common as information barriers become less pronounced.

2 The role of selection

I am curious about whether the networks derived from the four journals analyzed in this paper can be used to give more general information about the broader network of statistical collaboration and citation. Like the other discussants Regueiro, Sosa, and Rodríguez, I found the communities identified by the authors somewhat puzzling. As the authors say in their disclaimer, these clusters can be hard to interpret. One reason that the communities may not match with our heuristic expectations is the nature of sub-sampling in graphs.

To formalize ideas, let $G = (V, E)$ be the full co-author network of statisticians, however one might choose to define this group, where V is the set of statisticians, and an edge in E indicates a co-authorship relationship. Let J be the set of statistics journals, however we might define it. For each edge $\{i, k\} \in E$, there is an associated set of journals J_{ik} in which i and k have co-authored at least one article. Of course, $J_{ik} = \emptyset$ if i and k have never authored an article together.

Let $H \subseteq J$ be a subset of journals. Suppose now that we find a subgraph $G_H = (V_H, E_H)$ in which V_H consists of all authors that have published at least once in a journal in H, and $\{i, k\} \in E_H$ if for some $j \in J_{ik}$, $j \in H$. Equivalently, G_H is the induced co-authorship subgraph of authors who have published at least once in the journal set H. This setup is essentially a model of selection on vertex attributes. Is the induced subgraph G_H "representative" of G? Does it share any topological properties with G? The answers to these questions speak directly to claims that the field is becoming more collaborative.

It is now well known that sub-samples of networks can be troublesome (Stumpf et al, 2005; Lee et al, 2006; Shalizi and Rinaldo, 2013; Chandrasekhar and Jackson, 2014). Since degree is not preserved by taking the induced subgraph from sub-sampled vertices, centrality measures – especially degree centrality – may not be preserved eitler. Transitivity, clustering, and other network features of G_H may not be generalizable to G. Even if the journals in H were selected at random from J, G_H still might not preserve some topological properties of G. In summary, it is possible that considering a different set $H \subseteq J$ of journals would yield starkly different conclusions about collaboration.

3 The role of time

Co-authorship and collaboration are social processes that evolve over time, but the networks discussed in the paper are static. The authors suggest that the field of statistics has become more competitive, collaborative, and globalized. In addition to their reliance on meta-data not collected here, these assertions seem to indicate a role for time in the evolving pattern of interactions between authors. Associated with each paper in the dataset is a publication date. Of course, publication date is not the same thing as the date of initiation of authorship, the date a manuscript is finished, or even the date is accepted by the journal. But if we assume these journals have similar average review times, we can at least take the time ordering of publication to be a rough measure of the time order of paper authorship.

In addition, edges and other apparent topological patterns in the multi-year co-authorship network correspond to discrete events, ordered in time. For example, a triangle (i, j, k) in the multi-year network may indicate three co-authors (i, j, k) of the same paper, or distinct co-authorship relationships (i, j), (j, k), and (i, k), or some combination, at possibly different times. The time ordering of the papers represented by these edges likely matters in our interpretation of topological features of the network.

We can also understand the co-authorship data as a contact process between authors, and the citation data as a conact process between papers, rather than as networks (e.g. Hawkes and Oakes, 1974; Blundell et al, 2012). While authors exist before and after their publication of any particular paper, articles can only cite those that were previously published (or at least available in citable form). If the publication date of a paper i is t_i, then the citation data may be understood as a directed graph in which an edge from paper i to j can only occur if $t_j < t_i$.

Figure 3 shows co-authorship network "B" by year. Red vertices represent authors of papers appearing that year in one of the journals, and gray vertices indicate authors in the dataset who did not author a paper in those journals in that year. The layout of vertices in the graphs is the same from year to year. The central connected component seems to remain mostly constant in its density from year to year, but in 2012 density of co-authorships is markedly decreased.

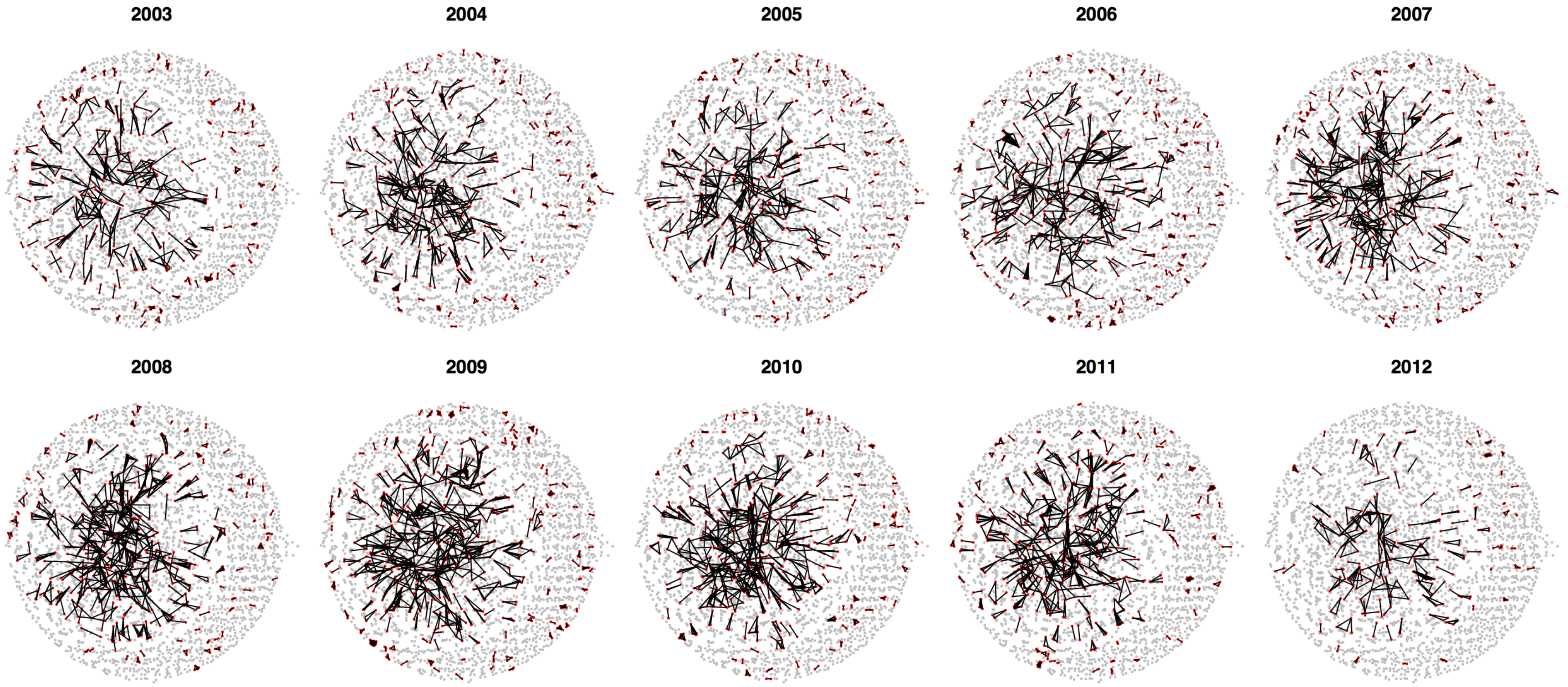

Figure 1: Coauthorship network “B” by year. Vertices represent authors and edges represent co-authorship. Red vertices are authors who published at least one paper in that year. The vertex layout is the same from year to year.

References

Blundell C, Beck J, Heller KA (2012) Modelling reciprocating relationships with hawkes processes. In: Advances in Neural Information Processing Systems, pp 2600–2608

Chandrasekhar AG, Jackson MO (2014) Tractable and consistent random graph models. Tech. rep., National Bureau of Economic Research

Hawkes AG, Oakes D (1974) A cluster process representation of a self-exciting process. Journal of Applied Probability pp 493–503

Lee SH, Kim PJ, Jeong H (2006) Statistical properties of sampled networks. Physical Review E 73(1):016,102

Shalizi CR, Rinaldo A (2013) Consistency under sampling of exponential random graph models. Annals of statistics 41(2):508

Stumpf MP, Wiuf C, May RM (2005) Subnets of scale-free networks are not scale-free: sampling properties of networks. Proceedings of the National Academy of Sciences of the United States of America 102(12):4221–4224