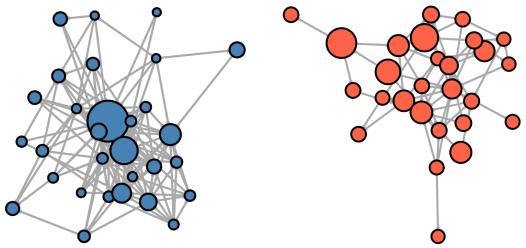
Yicong Jiang and Zheng Tracy Ke's Contribution to the Discussion of "Root and community inference on the latent growth process of a network" by Harry Crane and Min Xu

Yicong Jiang
Ph.D. student of Statistics, Harvard University, Cambridge, USA.
E-mail: yicong_jiang@g.harvard.edu
Zheng Tracy Ke
Associate Professor of Statistics, Harvard University, Cambridge, USA.
E-mail: zke@fas.harvard.edu

We congratulate the authors on an excellent paper! Crane and Xu (2021) proposed novel methods for finding "root nodes" from a single snapshot of a dynamic network process, with several interesting real-data examples. We now consider a new application for finding "root papers" in a citation network. The MADStat dataset (Ji et al., 2022; Ke et al., 2023) consists of the bibtex and citation information of over 83K papers, which we use to construct paper citation networks. Given a keyword (e.g., "Lasso"), let V_0 be the set of papers whose titles contain this keyword, and let V be the set of papers that are either citers or citees of papers in V_0 (we only count the citations recorded in MADStat). We then build a symmetric network on V, with an edge between two papers i and j if either i cites j or j cites i; if the network is disconnected, we restrict it to its giant component. The networks for two keywords, Lasso and Bayesian, are shown in Figure 1. We apply the method in Crane and Xu (2021) to each network to obtain the posterior probability of each node being a root node. The top 6 papers with the highest posterior root probability are in Table 1. In the Lasso network, Tibshirani (1996) is ranked top 1. In the Bayesian network, Gelfand and Smith (1990) is ranked top 1. The results are meaningful and motivate a new application of the proposed method.



	Lasso	Bayesian
#Nodes	2308	18502
#Edges	7137	54794
$d_{\min}$	1	1
$d_{\max}$	1335	718
$\bar{d}$	6.18	5.92

Fig. 1. The Lasso network (blue) and the Bayesian network (red); only the 30 highest-degree nodes are shown. The table on the right provides the summary statistics, where $d_{\max}$, $d_{\min}$, and $\bar{d}$ are the maximum, minimum, and average degrees, respectively.

Table 1. The top 6 papers with the highest posterior root probability in the Lasso network (top) and the Bayesian network (bottom), respectively.

Title	Author(s) & Year	Journal	#Citation	Root Prob.
Regression Shrinkage And Selection Via The Lasso	Tibshirani (1996)	JRSSB	55448	0.50
High-dimensional Graphs And Variable Selection With The Lasso	Meinshausen and Bühlmann (2006)	AoS	4328	0.05
The Adaptive Lasso And Its Oracle Properties	Zou (2006)	JASA	8245	0.03
Simultaneous Analysis Of Lasso And Dantzig Selector	Bickel et al. (2009)	AoS	2800	0.01
The Bayesian Lasso	Park and Casella (2008)	JASA	3453	0.01
Sparsity And Smoothness Via The Fused Lasso	Tibshirani et al. (2005)	JRSSB	3212	0.01
Sampling-based Approaches To Calculating Marginal Densities	Gelfand and Smith (1990)	JASA	9818	0.13
Bayesian Statistics In Medicine: A 25 Year Review	Ashby (2006)	SMed	295	0.11
Bayesian Computation Via The Gibbs Sampler And Related Markov-chain Monte-carlo Methods	Smith and Roberts (1993)	JRSSB	2536	0.08
Bayesian Experimental Design: A Review	Chaloner and Verdinelli (1995)	StSci	2354	0.06
Bayesian Computation And Stochastic-systems	Besag et al. (1995)	StSci	1548	0.05
Bayesian Measures Of Model Complexity And Fit	Spiegelhalter et al. (2002)	JRSSB	14395	0.05

We also suggest some extensions of Crane and Xu (2021). First, the PAPER model is built on the Erdos-Renyi model and does not model degree heterogeneity among nodes. The Erdos-Renyi model can be generalized to accommodate degree heterogeneity (such as a DCBM model with $K = 1$; see Jin et al. (2021)). It will be interesting to see if the PAPER model can be generalized similarly. Second, in the case of multiple roots, we may run community detection first and then apply the algorithm to each community separately. There are fast community detection algorithms (e.g., Jin et al. (2021); Jiang and Ke (2023)) equipped with data-driven choices of the number of communities (Jin et al., 2023). Combining them with the current algorithm will help reduce computational costs and avoid randomness caused by forest partition. We hope these ideas are beneficial. Congratulations to the authors again on their remarkable work!

References

Ashby, D. (2006). Bayesian statistics in medicine: a 25 year review. *Statistics in medicine 25*(21), 3589–3631.

Besag, J., P. Green, D. Higdon, and K. Mengersen (1995). Bayesian computation and stochastic systems. *Statistical science*, 3–41.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector.

Chaloner, K. and I. Verdinelli (1995). Bayesian experimental design: A review. *Statistical science*, 273–304.

Crane, H. and M. Xu (2021). Root and community inference on the latent growth process of a network using noisy attachment models. *arXiv preprint arXiv:2107.00153*.

Gelfand, A. E. and A. F. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association 85*(410), 398–409.

Ji, P., J. Jin, Z. T. Ke, and W. Li (2022). Co-citation and co-authorship networks of statisticians. *Journal of Business & Economic Statistics 40*(2), 469–485.

Jiang, Y. and T. Ke (2023). Semi-supervised community detection via structural similarity metrics. *arXiv preprint arXiv:2306.01089*.

Jin, J., Z. T. Ke, and S. Luo (2021). Improvements on score, especially for weak signals. *Sankhya A*, 1–36.

Jin, J., Z. T. Ke, S. Luo, and M. Wang (2023). Optimal estimation of the number of network communities. *Journal of the American Statistical Association 118*(543), 2101–2116.

Ke, Z. T., P. Ji, J. Jin, , and W. Li (2023). Recent advances in text analysis. *Annual Review in Statistics and Its Applications (in press)*.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso.

Park, T. and G. Casella (2008). The bayesian lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Smith, A. F. and G. O. Roberts (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological) 55*(1), 3–23.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology 64*(4), 583–639.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology 58*(1), 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology 67*(1), 91–108.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.