Summary for JSM talks

1. **Speaker:** Jingfei (Emma) Zhang

   Talk: *Using Maximum Entry-Wise Deviation to Test the Goodness of Fit for Stochastic Block Models*

   Traditional community detection methods always assume the number of clusters k is known. In practice, people use selection criterion such as BIC or hypothesis testing to determine k. However, selection methods usually assume k is a fixed constant and previous hypothesis testing methods only allow the k growing in a slow speed with increasing dimension n. This work tackles the goodness-of-fit testing problem for the Stochastic block model (SBM). Let k be the true number of clusters and k' be the hypothetical number of clusters. This work considers the hypothesis testing:

   $$H_0 : \ k = k' \quad \leftrightarrow \quad H_1 : \ k > k'.$$

   The key idea is that: if we remove the SBM signal correctly, we will have some distributions for the noise matrix to do the test. The proposed test statistics is the maximum deviation across n nodes, $\max_{i\in[n]} \sum_{\hat{z}(j)=a} A_{ij}/|\hat{z}(j) = a|$, where $\hat{z}$ is the estimated assignment. When $\hat{z}$ is strongly consistent, the test statistics asymptotically converges to extreme value distribution under the null and is larger than $\log n$ with high probability under the alternative. Here, the hypothetical k' is allowed to grow as $o(n/\log^2 n)$, and the estimator $\hat{z}$ in Gao et al. (2018) achieves strong consistency under such growing k'. The proposed testing procedures also apply for the degree-corrected SBM.

   Questions & Chat:

   - Do we need to estimate the degree heterogeneity under the degree-corrected SBM?

     Yes. We need to find the estimate of assignment first and then find the plug-in estimate of degree heterogeneity $\boldsymbol{\theta}$. We also need the strong consistency of $\boldsymbol{\theta}$ to obtain a proper statistics.

   - Since the proposed work does not need the singular or eigen-gap assumption in SBM, can we easily extend this work to the higher-order cases?

     Yes. The proposed method also applies for bipartite network or hypergraph since no singular or eigen-gap is required.

   - Why we do not consider the alternative in another direction $k < k'$?

     When $k > k'$, we can fit the SBM with k' by merging a few clusters into one cluster; but when $k < k'$, we can not properly estimate the assignment with more clusters. In

practice, we can propose a sequential testing with k' varying from 1 to a large number.

2. **Speaker:** Zheng (Tracy) Ke

   Talk: *Optimal Estimation of Network Mixed Memberships*

   This talk also tackles the goodness-of-fit problem under the degree-corrected SBM with hypothesis test

   $$H_0 : \ k = k' \quad \leftrightarrow \quad H_1 : \ k > k'.$$

   The test statistics are named as sgnT and sgnQ, which relate to the triangle and quadrilateral composed by the network edges, respectively. The key lemma is that: the clustering algorithm will merge true communities under the under-fitting cases ($k > k'$). The key lemma helps to separate the test statistics under the null and alternative. The singular-gap assumption is required guarantee performance of the test. The theoretical analyses are also different with odd or even k'.

   In addition, a new co-authorship data collected by Tracy and other professors are released to network application.

   Questions & Chat:

   - Is that difficult to extend the sgnT and sgnQ to the hypergraph or multi-layer network?

     The definition of sgnT and sgnQ should be different in hypergraph. We may consider the connection of one node to the other nodes to define similar sgnT and sgnQ in hypergraph. For multi-layer network, we may consider the sgnT in every layer and try to find some ways to combine the results.

3. **Speaker (Poster) :** Min Xu

   Poster & Chat: *Root and Community Inference on Preferential Attachment Model of Networks*

   In epidemiology, people want to detect the "first patient" that is the root of a group infected people. Regular community detection problem does not address such problem. This work considers the network community as a growth process where the edges are connected sequentially. When multiple growth processes exist in the network, the proposed algorithm detect the roots by estimating the probability to be "first patient" and identify the community detection at the same time. Compared with traditional SBM, the proposed method is more similar to degree-corrected SBM, where the degree of the nodes seriously affects the community separation.

4. **Chat:** Guanyu Hu

   (May not be very accurate since I am not familiar with the application background) They bought a new sport data for the soccer games in Champions League. The data records the game information, which is formed as multi-layer networks. However, the networks in each layer may not have the same dimension due to the player changes, and the edges may not have the same meaning since the games happen between different teams/clubs. Tensor methods may be helpful to analysis the player characters.

5. **Speaker:** Rina Barber

   Talk: *Testing the Stability of a Black Box Algorithm*

   We call an algorithm stable if the estimation is unchanged when we resample a small fraction of the data set. For example, KNN and ridge regression are stable due to large sample size and the strong convexity, respectively. However, it is not easy to claim the algorithm stability when we do not know the mechanism of the algorithm; i.e., the black box algorithm. This work propose a hypothesis testing problem for algorithm stability:

   $$H_0: \text{ algorithm is stable} \quad \leftrightarrow \quad H_1: \text{ algorithm is not stable.}$$

   The proposed test statistics is based on the gap between the estimations with full sample size and with the data after dropping the last sample. A binomial test is used with controlled type I error and lower bounds for the power. No assumption on the algorithm and data distribution is made. For future direction, we may consider a weaker definition of stability to get a better power.

6. **Speaker (Poster):** Alan Aw (advised by Yun Song)

   Poster: *Flexible Non-Parametric Tests of Sample Exchangeability and Feature Independence*

   We call the data is exchangable if the distribution of the data is independent with the order of the data. Exchangability is a weaker assumption than the independence. People want to detect exchangability to detect the stratification of the objects (e.g. genes) or to detect the linkage disequilibrium (LD) among the genes. A non-parametric test procedures is proposed under no distribution assumption. (Speaker skips the specific procedures of the test.)

References

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2018). Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185.