Multiplex Network clustering

March 3, 2023

Abstract

Multi-layer networks are very common in real world, for example, friendship networks from various SNS platforms. Clustering aims to separate elements so that an element in one set are more similar to other elements in the same set than elements in another set. In this project, we focus on clustering of nodes and layers in binary multiplex networks, which have binary entries and one-to-one mapping between node sets of a pair of layers. Using both synthetic and real datasets, we compare the performance of the multi-layer clustering methods including Jing et al. (2020), Lei et al. (2020), Fan et al. (2021), Lei and Lin (2022), and so on. From this project, students are expected to learn the notion of multiplex network, and multilayer stochastic block models, and numerical optimization techniques.

We will study about *cluster analysis* for *multiplex networks.* To understand this, we first define two terms, cluster analysis and multiplex networks. Then, we discuss about the project plan.

Cluster analysis aims to group objects (or data points) into subsets, referred to as *clusters*, such that the similarity between two points in the same cluster is greater than that of two points in different clusters. Here, similarity or dissimilarity between two points can be measured by various ways such as distance, correlation, and so on.

To define multiplex network, we first define *multilayer networks.* A multiplex network, $\mathcal{G}$, is defined by L layers, $\{G_1, \ldots, G_L\}$. The ℓth layer is defined by $G_\ell = (V_\ell, E_\ell)$ with a vertex set V_ℓ and an (”intra”-layer) edge set E_ℓ. Also, there would be inter-layer connections between two layers, denoted by E_{ℓ_1,ℓ_2}. Multiplex networks are the special case of multilayer networks with an identical vertex set $V = V_1 = \cdots = V_L$ and no inter-layer edges. For more details about multilayer networks, we refer to [1].

All methods we want to study in this project consider variations of stochastic blockmodels (SBMs) [3] for multiplex networks. SBMs is a probabilistic network models and have been used to understand the node community structures of networks. Once we have communities, we have block probabilities by computing the edge proportions within community/between communities. Multilayer stochastic blockmodels (MSBMs) are extension to multilayer networks. The considered papers use MSBMs or mixture of MSBMs on multiplex networks.

1 Project timeline plan

- Review clustering methods for single-layer networks: hierarchical clustering and spectral clustering (See Chapter 4.3.3 in [6] or Chapter 4.4 in [7]
- For the basic understanding of multilayer networks, read [5, 1].
- Read [8, 2, 4] and write a summary for Background section. (Gradually write)
  - [8] uses MSBMs with tensor decomposition. The goal of this method is to identify “global” node communities.
  - ALMA [2] and TWIST [4] assumes the mixture of MSBMs. They find global or local node communities, layer classes/communities.
- Then, try to implement methods one by one in R. (no later than end of April)
  - Fortunately, TWIST is implemented in R [9].
- If time allows, we may try other methods: [10, 11, 12].
- Using simulated data, compare the multilayer clustering methods (+ single-layer clustering methods on aggregated networks). Potential performance measure would be:
  - Check the ”mis-clustering” rate of each method. Assume we know the true communities $V_1, \ldots, V_K$ and the number of communities K. Assume we estimate communities $\widehat{V}_1, \ldots, \widehat{V}_K$ (after re-ordering to match the index that $V_\ell \cap \widehat{V}_\ell$ is large enough), then compute $\sum_{i \in V_k} \mathbb{I}(i \notin \widehat{V}_k)/|V_k|$.
  - Other performance metrics can be used.
- Using a real dataset, compare the result. (no later than the second week of May)
  - Need to find a dataset you might be interested in (April?). There are a few data repository for network data, but not limited to:
    - `https://networkrepository.com/index.php`
    - UCI Network Data Repository `https://networkdata.ics.uci.edu`
    - Stanford Large Network Data collection `https://snap.stanford.edu/data/`
    - Few other links `https://kateto.net/2016/05/network-datasets/`
    - `https://networks.skewed.de/?search=multilayer`
  - complete simulation and real data analysis
- Write the final report (1–2 times feedback before final submission)

2 Miscellaneous

See Prof. Davison's note for the structure of the final report (Section 2) and several writing tips (Chapter 3)

For examples of github repository: see this or this. The former is a simpler one from a student I advised last semester, which contains python codes and README file. The latter is my repository, which is a repository for an R package and a website for the package. Your repository should have two things: i) source code; ii) a README file.

References

[1] Ginestra Bianconi. *Multilayer Networks*, volume 1. Oxford University Press, 2018.

[2] Xing Fan, Marianna Pensky, Feng Yu, and Teng Zhang. ALMA: Alternating Minimization Algorithm for Clustering Mixture Multilayer Network. *arXiv:2102.10226 [cs, math, stat]*, 2021.

[3] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[4] Bing-Yi Jing, Ting Li, Zhongyuan Lyu, and Dong Xia. Community detection on mixture multilayer networks via regularized tensor decomposition. *The Annals of Statistics*, 49(6), 2021.

[5] M. Kivela, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

[6] Eric D. Kolaczyk. *Statistical Analysis of Network Data.* Springer Series in Statistics. Springer New York, New York, NY, 2009.

[7] Eric D. Kolaczyk and Gábor Csárdi. *Statistical Analysis of Network Data with R.* Use R! Springer, Cham, second edition edition, 2020.

[8] Jing Lei, Kehui Chen, and Brian Lynch. Consistent community detection in multi-layer network data. *Biometrika*, 107(1):61–73, 2020.

[9] Ting Li, Zhongyuan Lyu, Chenyu Ren, and Dong Xia. rMultiNet: An R Package For Multilayer Networks Analysis, 2023. arXiv:2302.04437.

[10] Subhadeep Paul and Yuguo Chen. Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics*, 10(2), 2016.

[11] Subhadeep Paul and Yuguo Chen. Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *The Annals of Statistics*, 48(1), 2020.

[12] Mirko Signorelli and Ernst C. Wit. Model-based clustering for populations of networks. *Statistical Modelling*, 20(1):9–29, 2020.