Community Embeddings with Bayesian Gaussian Mixture Model and Variational Inference

Anton I. N. Begehr
Graduate School of Business
National Research University Higher School of Economics
Moscow, Russia
a.begehr@fu-berlin.de

Prof. Dr. Petr Panfilov
Graduate School of Business
National Research University Higher School of Economics
Moscow, Russia
ppanfilov@hse.ru

***Abstract*—Graphs, such as social networks, emerge naturally from various real-world situations. Recently, graph embedding methods have gained traction in data science research. The graph and community embedding algorithm ComE aims to preserve first-, second- and higher-order proximity. ComE requires prior knowledge of the number of communities K. In this paper, ComE is extended to utilize a Bayesian Gaussian mixture model with variational inference for learning community embeddings (ComE BGMM+VI), similar to ComE+. ComE BGMM+VI takes K as the maximum number of communities and drops components through the trade-off hyperparameter weight concentration prior. The advantage of ComE BGMM+VI over the non-Bayesian ComE for an unknown number of communities K is shown for the small Karate club dataset and explored for the larger DBLP dataset.**

***Index Terms*—graph, embedding, community embedding, ComE, Bayesian, variational inference, Gaussian mixture, expectation maximization**

I. Introduction

Graphs, such as social networks, knowledge graphs, content-rating graphs, and communication networks, emerge naturally from various real-world situations. Analyzing these graphs leads to findings and understanding of the underlying structures, coherences, and dependencies. Recently, methods for embedding graph's nodes into lower-dimensional Euclidean spaces, called graph embeddings, have gained traction in multiple areas of data science research [1].

Community Embeddings, in addition to embedding a graph's nodes through first- and second-order proximity, also preserve higher-order proximity by embedding clusters present in the graph data. The graph and community embedding algorithm ComE aims to preserve first-, second- and higher-order proximity by embedding a graph's nodes and communities[2]. ComE requires prior knowledge of the number of communities K. In this paper, ComE is extended to utilizing a Bayesian Gaussian mixture model with variational inference for learning community embeddings (ComE BGMM+VI), similar to ComE+ published by Cavallari et al. in 2019 [3]. ComE BGMM+VI takes K as the maximum number of communities and drops components through a trade-off hyperparameter.

The reported study was partially supported by RFBR grant №20-07-00958. The paper was prepared within the framework of the HSE University Project Group Competition 2020-2022.

The recent 2017 graph embeddings algorithm ComE is extended similarly to the 2019 ComE+ by taking a Bayesian approach. The open-source code for the Bayesian approach to ComE's community embedding is published on GitHub[1] and serves as a contribution to community embedding research [4].

The original ComE paper *Learning Community Embedding with Community Detection and Node Embedding on Graphs* by Cavallari et al. and the ComE+ paper *Embedding Both Finite and Infinite Communities on Graphs* by Cavallari et al. in combination with the ComE source code have provided the basis and architecture for the community embeddings utilized in this work [2, 3, 5].

Multiple surveys and articles on graph embeddings were consulted to build a full picture of the current state of graph embedding research. Especially the 2018 survey *Graph Embedding Techniques, Applications, and Performance: A Survey* by Goyal and Ferrara and the 2020 paper *On Proximity and Structural Role-based Embeddings in Networks: Misconceptions, Techniques, and Applications* by Rossi et al. have proven to be primary resources for understanding the current landscape of graph embedding research [1, 6].

On part of comparing the Bayesian Gaussian mixture model with variational inference to the Gaussian mixture model with expectation-maximization, the 2006 book *Pattern Recognition and Machine Learning (Information Science and Statistics)* by Bishop includes essential statistics and information science knowledge and explanations [7].

II. Community Embedding

References

[1] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," *Knowledge-Based Systems*, vol. 151, p. 78–94, 7 2018. [Online]. Available: http://dx.doi.org/10.1016/j.knosys.2018.03.022

[2] S. Cavallari, V. W. Zheng, H. Cai, K. C.-C. Chang, and E. Cambria, "Learning community embedding with community detection and node embedding on graphs," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM '17. Association for Computing Machinery, 2017,

[1] at https://github.com/abegehr/ComE_BGMM

p. 377–386. [Online]. Available: https://doi.org/10.1145/3132847.3132925

[3] S. Cavallari, E. Cambria, H. Cai, K. C. Chang, and V. W. Zheng, “Embedding both finite and infinite communities on graphs [application notes],” *IEEE Computational Intelligence Magazine*, vol. 14, no. 3, pp. 39–50, 2019.

[4] A. Begehr. (2020) abegehr/ComE_BGMM. [Online]. Available: https://github.com/abegehr/ComE_BGMM

[5] S. Cavallari. (2017) andompesta/ComE. [Online]. Available: https://github.com/andompesta/ComE

[6] R. A. Rossi, D. Jin, S. Kim, N. K. Ahmed, D. Koutra, and J. B. Lee, “On proximity and structural role-based embeddings in networks: Misconceptions, techniques, and applications,” in *Transactions on Knowledge Discovery from Data (TKDD)*, 2020, p. 36.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.