

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 [3] in [4]. ComE BGMM+VI takes K as the maximum number of communities and drops components through a trade-off hyperparameter.

The recent [5] graph embeddings algorithm ComE is extended similarly to the [6] ComE+ by taking a Bayesian approach. The open-source code for the Bayesian approach to ComE's community embedding is published on GitHub¹ and serves as a contribution to community embedding research.[7]

The original ComE paper [8] by [9] and the ComE+ paper [10] by [11] in combination with the ComE source code have provided the basis and architecture for the community embeddings utilized in this work.[12][13]

Multiple surveys and articles on graph embeddings were consulted to build a full picture of the current state of graph embedding research. Especially the [14] survey Goyal2018by?andthe?paperrossi20tkdd – rolesby?haveproventobepri

On part of comparing the Bayesian Gaussian mixture model with variational inference to the Gaussian mixture model with expectation-maximization, the [15] book Bishop06 by [16] includes essential statistics and information science knowledge and explanations.[17]

II. COMMUNITY EMBEDDING

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.

¹at ComE_{BGMM_GH}