Internship offer: Deep Generative Models for the Joint Analysis of Networks and Continuous data

Supervisors and location

- Teams: INRIA MASSAI (Sophia-Antipolis), LMBP (UCA)
- Supervisors: Marco Corneli, Charles Bouveyron, Pierre Latouche
- Location: Centre INRIA Université Côte d'Azur, 2004 Rte des Lucioles, 06902 Sophia-Antipolis

Contacts

marco.corneli@inria.fr, charles.bouveyron@inria.fr, pierre.latouche@math.cnrs.fr

Context

Since the seminal approach on variational graph auto-encoders of Kipf and Welling [KW16], graph neural networks (GNN, [SGT$^+$08]) and more specifically graph convolution networks (GCN, [CWH$^+$20]), are now widely used in the unsupervised context, for network analysis. The deep latent variable models (DLVM) based strategies are usually used to build node embeddings characterizing the network topology. The supervision team for this internship has been working on networks using computational statistics and machine learning for the last 10 years. Applications of their work in this context to analyze social networks have recieved strong attention during the last French presidential election, with four papers written in LeMonde journal, two in front page, two days in a row.

Project

Existing deep latent variable models (DLVM), relying on graph neural networks (GNN), are specifically designed to analyze network topologies. Covariate variables can also be handled on vertices or edges. However, these extra sources of information are usually seen as given, such that no probabilistic model is used to characterize their distribution. In a predictive perspective, this regression context might be sufficient. For existing techniques, clusters of nodes can then be regarded as tools to explain the residual terms. However, in the unsupervised framework, where interpretability is key, this might strongly affect the quality of the clusters uncovered. Conversely, the extra sources of data should be encompassed in a larger probabilistic model, such that all sources are used for inference and to uncover meaningful clusters. The principal motivation for this task comes from the fact that a great deal of communication involves data such as images. So, the DLVM models should be extended as well as the inference procedure to deal with multiple sources of data on the edges and on the vertices. A principal approach consists in relying on pre-trained deep neural networks (DNN) to obtain continuous features from the images. Strategies based on generative adversarial networks for images [BJV17] can be used for instance. The set of features should then be seen as an extra source of data in high dimension. In order to better capture the relevant information present in the data, a series of model based methods will be considered. The mixture of factor analysers [MP00] as well as the Fisher EM algorithm [JBL21], and the recent infinite mixture of infinite factor [MVG20], are obvious techniques to be investigated. The goal is then to propose a new probabilistic model encapsulating GNN as well as one of the model based methods retained for high dimensional data. Clusters should explain both the construction of the network as well as the nature of the images exchanged. Variational techniques will be used to perform inference using approximated versions of the marginal likelihood.

Objectives

The new methods proposed will strongly improve the existing methods for social network analysis. They will also be key to detect the spread of fake news.

Expected skills

The candidate should be a master 2 student in a statistics / machine learning program, with a strong background in mathematics and computer science.

References

[BJV17] Urs Bergmann, Nikolay Jetchev, and Roland Vollgraf. Learning texture manifolds with the periodic spatial GAN. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 469–477. PMLR, 06–11 Aug 2017.

[CWH$^+$20] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *International Conference on Machine Learning*, pages 1725–1735. PMLR, 2020.

[JBL21] Nicolas Jouvin, Charles Bouveyron, and Pierre Latouche. A bayesian fisher-em algorithm for discriminative gaussian subspace clustering. *Statistics and Computing*, 31(4):1–20, 2021.

[KW16] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[MP00] Geoffrey McLachlan and David Peel. Mixtures of factor analyzers. In *In Proceedings of the Seventeenth International Conference on Machine Learning*. Citeseer, 2000.

[MVG20] Keefe Murphy, Cinzia Viroli, and Isobel Claire Gormley. Infinite mixtures of infinite factor analysers. *Bayesian Analysis*, 15(3):937–963, 2020.

[SGT$^+$08] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.