Bayesian estimation for the generalised random dot product graph

Patrick Rubin-Delanchy

June 23, 2022

Large dynamic networks occur in many areas of cyber-security, and analysts seek to understand network topology and connectivity behaviour for various purposes (anomaly detection, network auditing, etc). One of the more popular applied approaches to network analysis is graph embedding, which refers to the task of representing the nodes as points in space, which can then be used for e.g. exploratory data analysis or as feature vectors in downstream machine-learning tasks.

The most popular graph embedding algorithms, such as node2vec [1] or spectral embedding [6], were not originally based on explicit statistical models. However, statistical models have helped understand and improve them. For example, by considering the stochastic block model, it becomes clear that negative as well as positive eigenvalues should be used for spectral embedding [3], contrary to many earlier recommendations. Similarly, asymptotic analysis of spectral embedding under the random dot product graph (RDPG) indicates that Gaussian clustering, rather K-means, should be used to find network communities [5]. Finally, most relevant to the present proposal, recent work [8] has shown that optimising the RDPG likelihood explicitly, rather than implicitly through spectral embedding, has concrete statistical advantages including reduced asymptotic variance, and the authors later also proposed a Bayesian solution [7].

The RDPG lacks generality and in particular is inadequate for several cyber-security networks because it essentially assumes homophilic connectivity (a friend of a friend is a friend) which is seemingly rare in computer networks. node2vec seems to make similar, homophilic assumptions, although the model it is implicitly fitting is less obvious. The generalised random dot product graph [4] and the graph root distribution [2], its infinite-dimensional counterpart, address this problem, but no Bayesian solution exists. From a few preliminary studies, it seems that the "shape of Bayesian uncertainty" about the latent positions in heterophilic, sparse graphs could be interesting.

The objectives of this work are two-fold:

- Develop a Bayesian approach to fitting the generalised random dot product graph or even the graph root distribution, ideally scalable to large, sparse graphs.
- Theoretically investigate the geometry of Bayesian uncertainty about the latent positions under heterophilic, sparse graphs, and determine the extent to which this is captured by the Bayesian approach above.

Several cyber-security datasets are available for this project, including that used in [4].

References

[1] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[2] Jing Lei. Network representation using graph root distributions. *The Annals of Statistics*, 49(2):745–768, 2021.

[3] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[4] Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E Priebe. A statistical interpretation of spectral embedding: the generalised random dot product graph. *Journal of the Royal Statistical Society: Series B*, 2022. To appear.

[5] Minh Tang and Carey E Priebe. Limit theorems for eigenvectors of the normalized laplacian for random graphs. *The Annals of Statistics*, 46(5):2360–2415, 2018.

[6] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[7] Fangzheng Xie and Yanxun Xu. Optimal Bayesian estimation for random dot product graphs. *Biometrika*, 107(4):875–889, 2020.

[8] Fangzheng Xie and Yanxun Xu. Efficient estimation for random dot product graphs via a one-step procedure. *Journal of the American Statistical Association*, pages 1–14, 2021.