ISSUES WITH SPARSE SPATIAL RANDOM GRAPHS

Francesca Panero [1]

[1] Department of Statistics, London School of Economics and Political Science, (e-mail: f.panero@lse.ac.uk)

ABSTRACT: Spatial networks describe relations among agents that live in a metric space and whose locations affect the probability of connections. Recently, nonparametric Bayesian statistics (BNP) proved itself to be a powerful tool to provide random graph models that mimic real world networks, but no proposals have been made so far to include spatial covariates. I will show how some available models fail in recovering spatial information and conjecture a way to solve the problem.

KEYWORDS: networks, spatial statistics, baysian nonparametrics.

1 Introduction

Relational data can be described by mathematical objects known as graphs, collection of nodes, which represent agents of any nature, connected by edges or links, indicating a relation between those nodes. In applications, a graph is usually called network. Networks describe a plethora of relational phenomena, like transportation, social interactions, email exchanges, protein interactions, internet connections and many more. Network data have been collected extensively in the last decades and have pushed the frontier of research to offer refined models able to fit the complexity of their information.

There are multiple characteristics of real network that researchers try to adhere to when designing a random graph model. The degree of a node is the number of edges departing from it, and the degree distribution of the network is an interesting aspect to study, which in many real examples has been observed to be close to a power-law. Additionally, real networks often display a strictly positive clustering coefficient - defined as number of triangles over triplets - which indicates the presence of transitivity in connections (a friend of a friend will most likely become a friend). Also, the density of edges in many of the gigantic networks we deal with nowadays seems to be very low, meaning that the number of edges does not grow as fast as the number of nodes squared.

The graphon framework received a lot of attention in the last two decades for its possibility of describing node exchangeable graphs, i.e. networks where a reshuffling of the labels of the nodes does not affect the probability of connections. Being exchangeability a convenient property underpinning many Bayesian models, it comes as no surprise that the graphon model became connected with many Bayesian proposals. The graphon also contains as special cases popular models like the stochastic block model and the latent factor model. Nevertheless, this framework is misspecified for sparse networks, being able to fit only dense or empty ones (see Orbanz & Roy, 2015 for a review). In section 2 I will review the model originally proposed in Caron & Fox, 2017 which uses BNP to overcome the sparsity limitation of graphons and fit some of the properties we observe in real data. This model stimulated an interesting line of research under the name of graphex process. The new proposals, though, fail to describe networks whose edges need a spatial component to be described. In section 3 I will show how Caron & Fox, 2017 fails to represent data that feature a strong spatial component. I will include in the comparison the multidimensional scaling algorithm and show how this spatial algorithm fails to describe such data as well. I will finally conjecture how we can move forward with a spatial network model under the graphex framework.

2 A Bayesian nonparametric model for sparse graphs

The model by Caron & Fox, 2017 defines a network as a Poisson point process on the positive real plane, $Z = \sum_{i,j\geq 1} z_{ij}\delta_{(\theta_i,\theta_j)}$,where z_{ij} is equal to 1 if there is an edge between nodes i, j and 0 otherwise, and $\theta_i \in \mathbb{R}_+$ is the label of the node. The model is heterogeneous, since the probability of connection depends on the node sociability weight $w_i \in \mathbb{R}_+$ (as opposed to homogeneous models with equal probability across all pairs of nodes):

$$\mathbb{P}(z_{ij} = 1|(w_k, \theta_k)_{k\geq 1}) = 1 - e^{-2w_iw_j}. \tag{1}$$

To tune the distribution of w, the authors propose $(\theta_k, w_k)_k$ to be sampled from a Poisson process with intensity $\lambda(d\theta)\rho(dw)$, with λ Lebesgue measure and ρ a Lévy measure. Equivalently, we can describe $W = \sum_{i\geq 1} w_i\delta_{\theta_i}$ as distributed according to a homogeneous completely random measure (CRM). CRMs are a BNP building block, being used as flexible prior distributions over functional spaces (Lijoi & Prünster, 2010). Caron & Fox, 2017 assume ρ to be regularly varying at 0 with exponent $\sigma \in [0, 1]$, which intuitively means that ρ behaves similarly to a power function with exponent σ in a neighborhood of 0 (for the formal definition, see Caron *et al.* , 2023). Under this assumption, they prove that the model describes empty, dense or sparse networks (with sparsity level tuned by σ) and that the degree distribution is a power-law with exponent $1+\sigma$ for high degree nodes. Caron *et al.* , 2023 additionally prove that the clustering coefficient of such model is asymptotically strictly positive.

3 Issues of current models with sparse spatial networks

Spatial networks are networks whose nodes live in a metric space, and their positions affect the probability of connections. An example is the network of airports, where nodes are airports and edges represent flight connections between them. An instance of it is available as the network of flight connections in the United States of America in 2010*. We focus on the continental part of the US, excluding Alaska and Hawaii, for a total of 713 airports and 10^4 connections. The network is sparse with power-law degree distribution. We can easily convince ourselves that connections are determined partly by the size of the airport (a "sociability"), and partly by its location.

We fitted eq. 1 to the dataset in order to estimate the sociability of each airport, using a generalised gamma process as prior for the weights (the set up is as described in Caron & Fox, 2017). Once obtained estimates, we sampled 100 networks from the posterior predictive and we compared the clustering coefficient against its true value. Clustering is usually associated with a strong space component, since spatial models favour connections between nodes that are close (therefore inducing transitivity). The clustering coefficient of the real data is 0.50, while the posterior predictive mean is 0.29 (95% credible interval $[0.25, 0.34]$). The BNP model provides a positive value, but still the true value sits far away from the estimated one, suggesting that sociability is not enough to capture the underlying dynamics of the airport data.

Another possibility to fit such data is to use a multidimensional scaling algorithm, which takes a pairwise similarity matrix between nodes (in our case, the binary adjacency matrix) and computes latent locations for the nodes which minimise a loss function known as strain (Mead, 1992). Applying the algorithm to the dataset and fixing a 2-dimensional latent space, we obtain figure 1. On the left side, longitude is plotted against the two projections, showing that none of them is able to recover the true locations (the results for latitude are similar). The orange dots represent the nodes with highest degree (hubs). On the right, where nodes are shown in the 2-dimensional latent space, we can clearly see that the positions are determined by the degree of the nodes, since the hubs are all projected in a tight central position.

*https://toreopsahl.com/datasets/

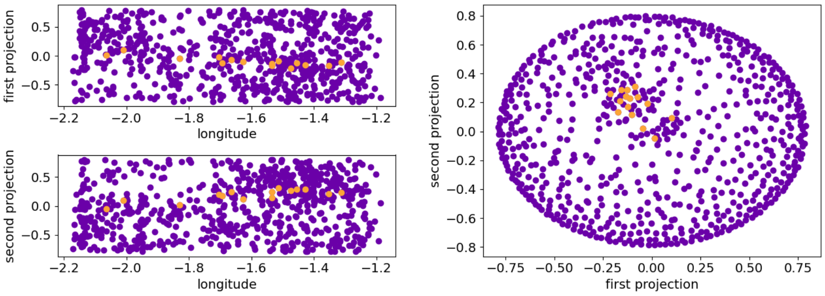

Figure 1. *2-dimensional MDS. On the left, longitude against the two projections, on the right nodes in the latent space. Orange represents nodes with high degree.*

The experiments suggest that a model to describe sparse spatial networks is needed. I conjecture that this could be a modification of eq. 1 with an additional spatial component. The model would inherit the interesting properties of sparsity, power-law degrees and interpretability. This would be beneficial not only for networks with a concrete notion of space, but also for those whose connections can be described by similarity of nodes measured in an abstract latent space (e.g. for qualitative covariates with no notion of distance).

References

Caron, F., & Fox, E. 2017. Sparse Graphs using Exchangeable Random Measures. *Journal of the Royal Statistical Society B*, **79**, 1–44. Part 5.

Caron, F., Panero, F., & Rousseau, J. 2023. On sparsity and power-law and clustering properties of graphex processes. *Advances in Applied Probability, to appear*.

Lijoi, A., & Prünster, I. 2010. Models beyond the Dirichlet process. *In:* Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (eds), *Bayesian Nonparametrics*. Cambridge University Press.

Mead, AL. 1992. Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society: Series D*.

Orbanz, P., & Roy, D. M. 2015. Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**(2), 437–461.