A Stochastic Block Hypergraph model

Marc → Alexis et Jean-Daniel

A proposal for a model of a hypergraph generalization of the stochastic block model, using the clustering connection probability P_{ij} and displaying explicitely the hyperedge formation process.

QUICK INTRODUCTION

The stochastic block model (SBM) has been used widely as a canonical model for community detection (see the review [1]). It is a very simple model of a graph with communities/clusters. It is a generative model for data and benefits from a ground truth for the communities.

The standard community detection problem is the statistical recovery problem of the clusters: from the graph, can we reconstruct the sets $\{C_i\}$? Depending on the parameter values there are different regimes (recovery impossible, easy, etc. [2–4]).

There are some papers proposing a model for a SBM for hypergraphs. See for example [5]. Most of the models proposed recently [6–9] consider hypergraphs with fixed degree and/or edge sizes (which in general do not correspond to empirical observations).

In [5], the authors propose a generative approach to hypergraph clustering based on a degree-corrected hypergraph stochastic blockmodel. This model generates clustered hypergraphs with heterogeneous degree distributions and hyperedge sizes. It is based on a probability distribution over the space of possible hypergraphs. Very general, but a bit difficult to use and also difficult to connect with real hypergraphs and processes at play in the formation and evolution of these structures...

In [7], the authors study the problem of community detection in a random hypergraph model which they call the stochastic block model for k-uniform hypergraphs. (model introduce in [6]). Each hyperedge appears in the hypergraph independently with the probability depending on the community labels of the vertices involved in the hyperedge. More precisely the standard simple model for $k-$uniform hypergraph with two clusters is the following one (introduced in [6], see also [7]).

We denote by V ($|V| = N$) the set of vertices of the hypergraph H and k an integer $k \geq 2$. p and q are numbers between 0 and 1 (possibly depending on N). The collection of size k subsets of V is given by $\binom{V}{k}$. The so-called $k-$stochastic block model for hypergraphs $k-$HSBM with parameters k, N, p, q, denoted by HSBM(N, p, q, k) is a model which samples a $k-$uniform hypergraph on the vertex set V according to the following rules:

- σ is a vector in $\{\pm 1\}^{|V|}$ chosen uniformly among those with equal numbers of $+1$ and -1. These $+/-1$ label the two different communities considered here.
- Each hyperedge in $\binom{V}{k}$ appears independently with probability

$$\text{Prob}(e) = \begin{cases} p & \text{if } \sigma_1 = \sigma_2 = \ldots = \sigma_k \\ q & \text{otherwise} \end{cases} \tag{1}$$

Recent work also construct random hypergraphs with communities [10, 11] and generalize the Poisson stochastic block model. I am not convinced of the usability and generality of this type of models...

Other remarks. It is unclear how to generalize the previous model to the case with more than 2 clusters. Most models that I found in the litterature are a bit obscur and difficult to use. Moreover, I didn't find a model that explicitely construct a stochastic block hypergraph starting from the quantity P_{ij} alone.

SIMPLE DEFINITIONS

The number $N = |V|$ of vertices is called the order of the hypergraph, and the number of hyperedges $M = |E|$ is usually called the size of the hypergraph. The size of an hyperedge $|e_i|$ is the number of its vertices. The degree of a vertex is then simply given by the number of hyperedges to which it is connected. A simpler hypergraph considered in many studies is obtained when all hyperedges have the same cardinality k and is then called a k-uniform hypergraph (the rank of a hypergraph is $r = \max_E |e|$ and the anti-rank $\overline{r} = \min_E |e|$, and when both quantities are equal the hypergraph is uniform). A 2-uniform hypergraph is then a standard graph.

THE STOCHASTIC BLOCK MODEL (SBM)

We have N nodes and a partition of these nodes into m disjoints communities (subsets) $C_1, C_2, \ldots, C_m$. The (symmetric) probability matrix P of connections is assumed to be known.

For constructing the SBM, two vertices $u \in C_i$ and $v \in C_j$ are connected by an edge with probability P_{ij}.

A special is given by the matrix

$$P_0(p, q) = \begin{bmatrix} p & q & \ldots \\ \vdots & \ddots & \\ q & & p \end{bmatrix}$$

which allows us to tune the ratio intra- versus interconnections. In particular, the case $p \gg q$ corresponds to the presence of well-defined clusters while other cases such as $p \sim q$ are more fuzzy. This type of model serves as a benchmark for community detection algorithm for example.

A (SIMPLE) HYPERGRAPH GENERALIZATION

We would like to find a model for a stochastic block hypergraph that satisfies a number of requirements:

- The minimal input is the number of nodes (N), the set of communities ($C_1, C_2 \ldots, C_K$), and the connection probability matrix P_{ij}.
- The matrix P specifies the number of hyperedges and their size (if you make a loop on all pair of nodes -see below). We can however fix the number of hyperedges and their sizes if needed.
- For $2-$uniform hypergraph, the model should recover the standard SBM
- The code should be'simple' enough - in particular, it should explicitly display the formation process of hyperedges.

Here is such a proposal. We start from N vertices and we construct a hyperedge e in the following general way:

1. We first choose a node u_1 at random. We assume that $u_1 \in C_i$. At this point the hyperedge is $e = \{u_1\}$.
2. We then choose a new node u_2 at random. We assume that $u_2 \in C_j$. We add u_2 to the hyperedge e with probability P_{ij}.
3. After a number n of steps, the hyperedge is of the form $e_n = \{u_1, u_2, \ldots, u_n\}$ where the community of u_l is denoted by $u_l \in C_{i(l)}$ (ie. $i(l)$ is the community which u_l belongs to). We now choose another node u_{n+1} at random and we have to choose the probability $P(u_{n+1} \to e_n)$ that u_{n+1} will belong to the hyperedge e_n. The main point here is that we can specify the formation process of the hyperedge. In other words, what drives the composition of a hyperedge ? There are several choices for this probability and a simple one is the following

$$P(u_{n+1} \to e_n) = \frac{1}{n} \sum_{j=1}^{n} P_{i(j)i(n+1)} \tag{2}$$

This choice corresponds to the average of all probabilities, but other choices are possible depending on the specific formation process. We could have chosen

$$P(u_{n+1} \to e_n) = \max_j P_{i(j)i(n+1)} \tag{3}$$

or the min. Both cases (the average or the max) make sense I believe.

4. We stop the construction of the hyperedge: (A) when all other nodes are tested (in which case the probability matrix P fixes the number and size of hyperedges), (B) when the size of the hyperedge reaches a value k which is fixed (we then obtain what is called a $k-$uniform hypergraph that mathematicians like), or (C) when the size of the hyperedge reaches a random value k distributed according to a distribution $P(k)$ (given empirically for example). The standard graph SBM is recovered for (B) with $k = 2$.

If we want to include additional information such as the degree distribution we have to assign to each node a degree k_i which will limit the number of hyperedges it belongs to (seems doable).

POSSIBLE MEASURES

Many measures are available for hypergraphs. Walks, paths and centrality measures can be defined and other measures such as the clustering coefficient can be extended to hypergraphs [12–16]. A recent study investigated the occurrence of higher-order motifs [17] and community detection was also considered [5, 18]. We can also measure the statistics of the intersection between two hyperedges as being the number of nodes they have in common [5].

In our case, we could limit ourselves to the main characterization of the hypergraph: we could measure the distribution of degree $P(k)$, the distribution of hyperedge sizes $P(m)$, the distribution of the overlap between two hyperedges $P(O)$. Maybe something else related to clustering (?) We could measure this for the simple case where $P = P_0(p, q)$ and varying p and q. In this case we could probably compute analytically some averages.

[1] Abbe E. Community detection and stochastic block models: recent developments. The Journal of Machine Learning Research. 2017 Jan 1;18(1):6446-531.

[2] Mossel, Elchanan; Neeman, Joe; Sly, Allan (February 2012). "Stochastic Block Models and Reconstruction". arXiv:1202.1499

[3] Decelle, Aurelien; Krzakala, Florent; Moore, Cristopher; Zdeborová, Lenka (September 2011). "Asymptotic analysis of the stochastic block model for modular networks

and its algorithmic applications". Physical Review E. 84 (6): 066106.

[4] Abbe, Emmanuel; Bandeira, Afonso S.; Hall, Georgina (May 2014). "Exact Recovery in the Stochastic Block Model". arXiv:1405.3267

[5] Chodrow, P.S., Veldt, N., Benson, A.R. Generative hypergraph clustering: From blockmodels to modularity. Science Advances 7.28 (2021): eabh1303.

[6] D. Ghoshdastidar, A. Dukkipati, Consistency of spectral partitioning of uniform hypergraphs under planted partition model. Adv. Neural Inf. Process. Syst. 27, 397–405 (2014).

[7] C. Kim, A. S. Bandeira, M. X. Goemans, Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. arXiv:1807.02884 [math.PR] (8 July 2018).

[8] M. C. Angelini, F. Caltagirone, F. Krzakala, L. Zdeborová, Spectral detection on sparse hypergraphs, in 2015 53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015 (IEEE, 2016), pp. 66–73.

[9] Z. T. Ke, F. Shi, D. Xia, Community detection for hypergraph networks via regularized tensor power iteration. arXiv:1909.06503 [stat.ME] (14 September 2019).

[10] Ruggeri N, Contisciani M, Battiston F, De Bacco C. Community detection in large hypergraphs. Science Advances. 2023 Jul 12;9(28):eadg9159.

[11] Ruggeri N, Battiston F, De Bacco C. A framework to generate hypergraphs with community structure. arXiv preprint arXiv:2212.08593. 2023 Jun;22.

[12] Zlatic, V., Ghoshal, G., Caldarelli, G. (2009). Hypergraph topological quantities for tagged social networks. Physical Review E, 80(3), 036118.

[13] Chodrow, P. S. (2020). Configuration models of random hypergraphs. Journal of Complex Networks, 8(3), cnaa018.

[14] Battiston, F., et al. Networks beyond pairwise interactions: structure and dynamics. Physics Reports 874 (2020): 1-92.

[15] Joslyn, Cliff A., et al. Hypernetwork science: from multidimensional networks to computational topology. International Conference on Complex Systems. Springer, Cham, 2020.

[16] Aksoy, S. G., Joslyn, C., Marrero, C. O., Praggastis, B., Purvine, E. (2020). Hypernetwork science via high-order hypergraph walks. EPJ Data Science, 9(1), 16.

[17] Lotito, Q. F., Musciotto, F., Battiston, F., Montresor, A. (2022). Exact and sampling methods for mining higher-order motifs in large hypergraphs. arXiv preprint arXiv:2209.10241.

[18] Turnbull, K., Lunagómez, S., Nemeth, C., Airoldi, E. (2019). Latent Space Modelling of Hypergraph Data. arXiv preprint arXiv:1909.00472.