EasyChair Preprint

№ 5438

Applications of Renormalizable Binary Fitness-Based Network Models

Aurélien Hazan

April 30, 2021

"Applications of renormalizable binary fitness-based network models" Aurélien Hazan, Université Paris-Est Créteil .

A recent binary fitness-based renormalizable network models is fit to empirical networks across several levels. First, the case when a "natural" hierarchical partition is available is examined, and examplified with trade data at a finer scale than in the original paper. The goodness of fit at a fixed level is compared to the output of established fitness-based probabilistic models.

Then the goodness of fit of the renormalizable model is assessed when the hierarchical partition is generated by state-of-the art model-based hierarchical clustering methods, which adds a null model to the original algorithm.

1 Introduction

Reducing the complexity of networks to increase interpretability is becoming all the more necessary as the size of empirical graphs made available for analysis increases.

Community-finding [1], coarse-graining, backboning [2], are different paths to this aim, guided by the necessity to simplify the network while preserving some properties of interest. Methods differ by the type of property they preserve (e.g. topological, spectral, dynamical).

In the present article, applications of the recently proposed fitness-based renormalizable network model by Garuccio et al. [3] are examined. The model tackles the issue of reducing networks complexity with a multi-scale and probabilistic standpoint, given the following hypotheses: a hierarchical partition is known a priori, for example in the form of a dendrogram; nodes are associated with known fitnesses, that sum across aggregation; edges bear a binary value.

This model is analytically tractable, estimates the edge probability a level l p_{ij}^{l}, and depends on just one parameter δ that fits the original network density, i.e. at level $l = 0$. Optionnally it can take dyadic information (such as distances between nodes, or group membership) as an input. Recent works started tackling the weighted case but the topic is still open [4].

This model lies at the intersection of several research fields, the first of which is the physics-oriented complex networks literature, inspired by renormalization in statistical physics.While geographical coarsening [5] considers graphs embedded in a 2d space used to perfom box-covering, fractal-oriented studies [6] use a shortest-path-distance on a graph to perform renormalization without needing an embedding space. In specific cases some quantitites such as the degree distribution can shown to be preserved. Spectral methods [7] truncate the set of eigenvalues of the graph Laplacian and average eigenvectors in order to preserve random-walk characteristics.

Secondly, the model by Garuccio et al. is related to model-based hierarchical clustering. For example latent space methods first project network nodes in an embedding space built from data, in the spirit of the Latent Space Model [8] where homophily increases if nodes get closer. Traditional clustering can then be performed as in the Latent Position Cluster Model (LPCM) [9], and

applied recursively in latent space, as was proposed in [10,11] upon projecting in a hyperbolic space. In the same family of model-based hierarchical clustering, the Stochastic Block Model (SBM) was extended to the hierarchical case in [12,13], where the former builds on the degree-corrected SBM [14]. Both algorithms rely on Monte-Carlo methods. As evoked in [14,15], the SBM and related models (e.g. block configuration model) can be formulated as an exponential random graph (ERG), that constitutes another important group of random network models of widespread use, including for clustering purpose.

In the present article the following questions will be examined: if a given empirical graph comes with a "natural" or expert-defined partition and a set of node fitnesses, is it well fit by the renormalizable model of Garuccio et al., and why ? This will be addressed in sec. 3 with trade data aggregated by product nomenclature level. Are the hierarchical clustering method depicted in the introduction well fitted by the renormalizable model, and why ?

2 The renormalizable model by Garuccio et al.

In this section, a simplified version of the model by Garuccio et al. [3] is presented, keeping their notations: $\mathbf{A}^{(l)}$ is the adjacency matrix of a binary undirected graph, that results from the iterated coarse-graining of an original matrix $\mathbf{A}^{(l)}$, with dimension $N_0 \times N_0$. Non-overlapping partition functions $\mathbf{\Omega}_l$ define partitions of the node set at level l. The partitions defined at all levels are considered as known *a priori*, for example in the form of a dendrogram. The chosen coarse-graining rule is simple: if there is at least an edge between two nodes i_l, j_l belonging respectively to super-nodes i_{l+1} and j_{l+1}, then the super-nodes are connected as well. This can be written: $a_{i_{l+1},j_{l+1}} = 1 - \prod_{i_l \in i_{l+1}} \prod_{j_l \in j_{l+1}} (1 - a_{i_l,j_l})$, where $i_l \in i_{l+1}$ means that node i_l at level l belongs to super-node i_{l+1} at level $l+1$. Random adjacency matrix are then considered, and associated with the probability $P_l(\mathbf{A}^{(l)})$. Hypothesizing independent links leads to the expression $\prod_{i_l=1}^{N_l} \prod_{j_l=1}^{i_l} \left[p_{i_l,j_l}^{(l)}\right]^{a_{i_l,j_l}^{(l)}} \left[1-p_{i_l,j_l}^{(l)}\right]^{1-a_{i_l,j_l}^{(l)}}$. Renormalizability of the model is expressed by the scale-invariance property, for any $l \geq m \geq 0$:

$$P_l(\mathbf{A}^{(l)}|\mathbf{\Theta}_l) = \sum_{\{\mathbf{A}^{(m)}\} \to \mathbf{A}^{(l)}} P_m(\mathbf{A}^{(m)}|\mathbf{\Theta}_m) \tag{1}$$

where the sum is on the set $\{\mathbf{A}^{(m)}\}$ of matrices that lead to $\mathbf{A}^{(l)}$ under successive application of the renormalization rule, and $\mathbf{\Theta}_l$ is a parameter depending only on $\mathbf{\Theta}_m$ and the partitions.

It is shown by the authors that there is a unique solution to the problem under scale-invariance: $p_{i_l,j_l(\delta)} = 1 - e^{-\delta x_{i_l} x_{j_l}}$ with the fitnesses obeying : $x_{i_{l+1}} = \sum_{i_l \in i_{l+1}} x_{i_l}$ where x_{i_l} are node-specific predefined parameters, the *fitnesses*. The interested reader is referred to Garuccio et al. [3] for more detail and the treatment of the dyadic case, that is omitted here.

This model is probabilistic, generative in the sense that one can sample from the set of adjacency matrices $\{A^{(l)}\}$ at all levels, and has an anlytical expression (in contrast to MCMC-based methods) which functional form is preserved across renormalization. Only one graph sample is required to fit the model. At any level $p_{i_l,j_l(\delta)}$ can be compared to other fitness-based models such as the FiCM model [16].

3 Empirical networks with a natural partition and summable fitnesses

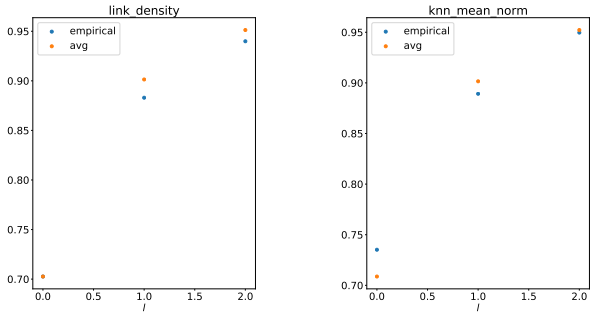

Fig. 1 Macro-scale properties of MRIO network across hierarchical levels l, empirical vs average under the null model. (a) link density; (b) $\bar{k}_{nn}$.

The usefulness of the above model will be examplified using data at a finer scale than in the original paper, namely Multi-Regional Input-Output data from Exiobase 3 that are broken down at the product level. They cover 44 countries (plus 5 Rest of World regions), 163 industries and 200 products, and represent monetary flow between industrial sectors (industry-by-industry tables), or between products (product-by-product tables). The products and industries can be grouped by family according to various correspondance tables. Here we use the Nace[1] classification at levels 1 and 2. Keeping the number of regions to 49, we generate aggregate undirected binary graphs with the following dimensions: At level 0, 163 industries are kept, which results in a vertex set of size 7987. At level 1, 88 industry groups remain, the resulting dimension is 4312. At level 2, 21 industry groups remain, the resulting dimension is 1029.

To assess goodness of fit to the model, macro-scale properties were computed for the three product levels as shown in Fig. 1. Fig. 1(a) shows that the empirical link density at level 0 is the same as the average link density under the null model, which is expected given the model definition. Further, it can

[1] NACE is the statistical classification of economic activities in the European Communities. Version 2 is used here.

be observed that at levels 1 and 2 the empirical and average quantities rise -which is expected since the network is getting denser upon renormalization- and remain close to each other. Fig. 1(b) shows $\bar{k}_{nn}$, the average nearest neighbor degree (ANND) averaged over the vertex to yield a scalar value. The two values show a significant discrepancy at level 0, and converge to a common value as the level increases.

In the rest of the paper the issue of the renormalizability of hierarchical probabilistic models produced by state-of-the-art hierarchical clustering algorithms will be examined and a typology of situations that favour "renormalizability" will be established

References

1. M.E.J. Newman, M. Girvan, Physical Review E **69**(2), 026113 (2004). DOI 10.1103/PhysRevE.69.026113. URL https://link.aps.org/doi/10.1103/PhysRevE.69.026113
2. V. Gemmetto, A. Cardillo, D. Garlaschelli, Irreducible network backbones: unbiased graph filtering via maximum entropy. Tech. rep. (2017). _eprint: 1706.00230
3. E. Garuccio, M. Lalli, D. Garlaschelli, Multiscale network renormalization: scale-invariance without geometry. Tech. rep. (2020). _eprint: 2009.11024
4. T.D. Lardy, A Renormalizable Random Graph Model. Tech. rep. (2018)
5. B.J. Kim, Physical Review Letters **93**(16), 168701 (2004). DOI 10.1103/PhysRevLett.93.168701. URL https://link.aps.org/doi/10.1103/PhysRevLett.93.168701
6. L.K. Gallos, C. Song, H.A. Makse, Physica A: Statistical Mechanics and its Applications **386**(2), 686 (2007). DOI 10.1016/j.physa.2007.07.069. URL https://linkinghub.elsevier.com/retrieve/pii/S0378437107007741
7. D. Gfeller, P. De Los Rios, Physical Review Letters **99**(3), 038701 (2007). DOI 10.1103/PhysRevLett.99.038701. URL https://link.aps.org/doi/10.1103/PhysRevLett.99.038701
8. P.D. Hoff, A.E. Raftery, M.S. Handcock, Journal of the American Statistical Association **97**(460), 1090 (2002). DOI 10.1198/016214502388618906. URL http://www.tandfonline.com/doi/abs/10.1198/016214502388618906
9. M.S. Handcock, A.E. Raftery, J.M. Tantrum, Journal of the Royal Statistical Society: Series A (Statistics in Society) **170**(2), 301 (2007). DOI 10.1111/j.1467-985X.2007.00471.x. URL http://doi.wiley.com/10.1111/j.1467-985X.2007.00471.x
10. M. Zheng, G. García-Pérez, M. Boguñá, M. Serrano, Scaling up real networks by geometric branching growth (2020). _eprint: 1912.00704
11. M. Boguñá, I. Bonamassa, M. De Domenico, S. Havlin, D. Krioukov, M. Serrano, Nature Reviews Physics **3**(2), 114 (2021). DOI 10.1038/s42254-020-00264-4. URL http://www.nature.com/articles/s42254-020-00264-4
12. T.P. Peixoto, Physical Review X **4**(1), 011047 (2014). DOI 10.1103/PhysRevX.4.011047. URL https://link.aps.org/doi/10.1103/PhysRevX.4.011047
13. V. Lyzinski, M. Tang, A. Athreya, Y. Park, C.E. Priebe, IEEE Transactions on Network Science and Engineering **4**(1), 13 (2017). DOI 10.1109/TNSE.2016.2634322. URL http://ieeexplore.ieee.org/document/7769223/
14. B. Karrer, M.E.J. Newman, Physical Review E **83**(1), 016107 (2011). DOI 10.1103/PhysRevE.83.016107. URL https://link.aps.org/doi/10.1103/PhysRevE.83.016107
15. J. van Lidth de Jeude, R. Di Clemente, G. Caldarelli, F. Saracco, T. Squartini, Complexity **2019**, 1 (2019). DOI 10.1155/2019/5120581. URL https://www.hindawi.com/journals/complexity/2019/5120581/
16. D. Garlaschelli, M.I. Loffredo, Physical Review Letters **93**(18) (2004). DOI 10.1103/PhysRevLett.93.188701. URL https://link.aps.org/doi/10.1103/PhysRevLett.93.188701