A random model that relies on maximal bicliques to preserve the overlaps in bipartite networks

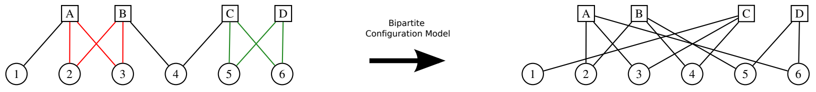
Fabien Tarissan[1]* and Lionel Tabourier[2]

[1] Université Paris-Saclay, CNRS, ENS Paris-Saclay, UMR 7220, ISP, France
[2] Sorbonne Universités, CNRS, LIP6, F-75005 Paris, France

Context. Many real-networks, also refered to as *complex networks*, lend themselves to the use of graphs in order to analyse their structure and model their properties. Since the seminal papers of Barabási and Watts, one usually considers that, whatever the context in which they emerge, all networks share non trivial properties such as a low density, a low average distance, an heterogeneous degree distribution, a high local density, etc.

Such properties distinguish those networks from classic random graph models such as the ones generated by the Erdős-Rényi model which only reproduce the density of the networks. As a consequence, significant effort is dedicated to the elaboration of random models able to capture more intricate properties. Among them, one can cite the Barabási-Albert model which succeeds in producing a heterogeneous (scale-free) degree distribution but fail in generating graphs with a high local density, the Watts and Strogatz model which generates networks with the opposite features or the Configuration Model [3] which generate random graphs with a prescribed degree sequence but with a low local density. All in all, and despite the different attempts, generating a graph exhibiting all expected properties is still an open issue.

The purpose of this study is to present a new step toward that goal by exploiting the bipartite version of the configuration model. Indeed, although useful, the representation of networks as unipartite graphs does not account for the inherent complexity induced by the hierachical structure observed in most real networks. This observation led the scientific community to turn to *bipartite graphs* to describe such complex structure when possible. This formalism allows to define explicitly two disjoint sets of nodes and the links only relate a node of one set to a node of the other set. The natural extension of the configuration model to bipartite graphs allows to preserve the degree of every nodes while shuffling the links, as depicted below:

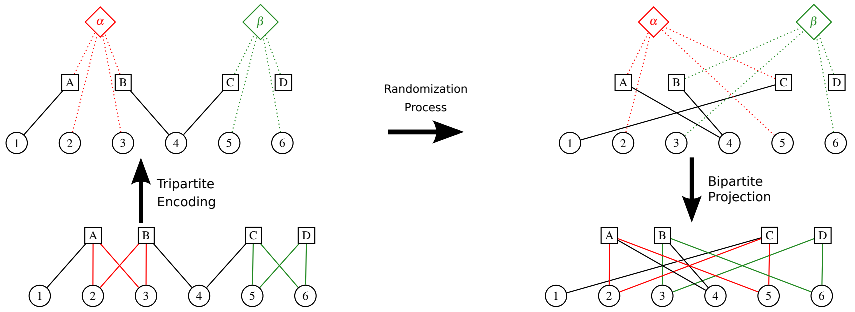


However, as illustrated in the picture, such a model can easily disturb key patterns of the structure. Although the degree distribution is preserved, the two bicliques (in red and green) completely vanish after the randomization due to a slight modification of

*This work is funded in part by CNRS under grant n 245 709 (PICS project Récital).

the links. To that regard, recent studies showed that overlaps (top nodes connected to common bottom nodes) are ubiquitous and important patterns in bipartite networks [4].

In order to overcome this issue, we propose in this paper a generative model able to preserve both the degree sequence and the overlaps of real networks. It relies on the encoding of those patterns in a third level, defining thus a *tripartite* graph, on which we perform the randomization. More precisely, we first perform the enumeration of all maximal bicliques in the bipartite graph, then encode the bicliques in a third level before performing a randomization preserving the encoding. Finally, we project the obtained tripartite graph into its corresponding bipartite structure:



One key operation in this method relies on the tripartite encodings of the bipartite structure. We tested several natural heuristics which select the bicliques in a given order to create the tripartite encoding: a random selection, a selection that maximizes the number of links encoded and one that maximizes the number of nodes captured.

Results show that all heuristics lead to generating bipartite graphs in which the overlaps are preserved. We show in addition that several other properties emerge naturally with much more accuracy than with a standard bipartite configuration model.

Results. In order to validate the approach, we tested the models on 9 datasets that have an underlying bipartite structure. Due to space limitation, we only show the results on three representative datasets: `HepB` is a network featuring scientists and the articles that they coauthored, collected from Medline repository using the keyword *Hepatitis B*, `BPSE` is a network built from the proteins of bacteria *Burkholderia pseudomallei* and the biochemical reactions they take part in, and `Youtube` contains the membership of Youtube users as collected in 2007 [2].

For each network, we computed several properties both on the original bipartite graphs and on the ones generated by the models. More precisely, let $G = (\top, \bot, E)$ be a bipartite graph, where $\top$ is the set of *top* nodes, $\bot$ the set of *bottom* nodes, and $E \subseteq \top \times \bot$ the set of links between $\top$ and $\bot$. We denote by $N(u)$ the set of neighbors of u in the bipartite graph and by $N_2(u)$ its neighbors at distance 2. We computed several nodes characteristics related to the overlaps: the *bipartite coefficient* [1] based on the Jaccard index defined as $\mathtt{bip}(u) = \frac{\sum_{v \in N(u)} \mathtt{cc}(u,v)}{|N(u)|}$ where $\mathtt{cc}(u,v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$, the *dispersion*

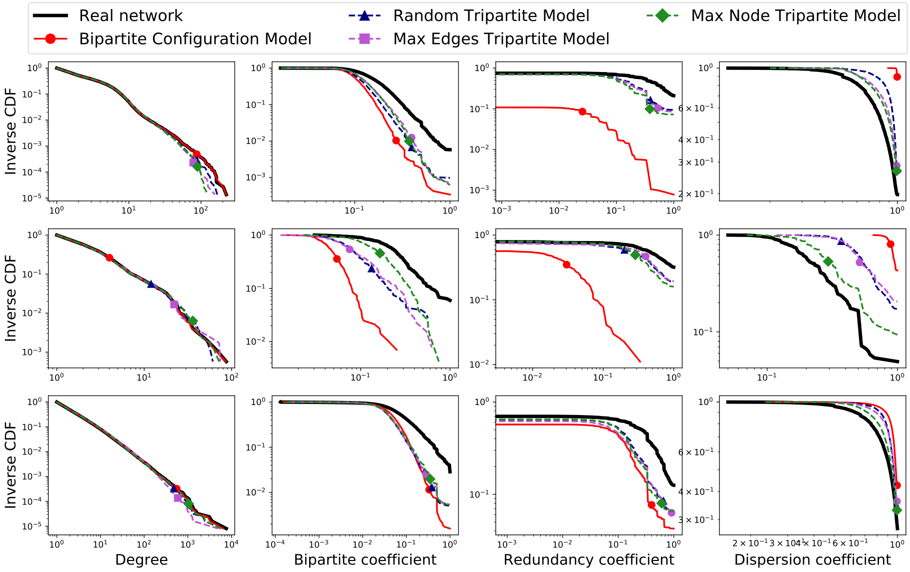


Fig. 1. Inverse cumulative distribution of the degree distribution (first column), the bipartite clustering coefficient (second column), the redundancy coefficient (third column) and the dispersion coefficient (fourth column) for `HepB` (top), `BPSE` (middle) and `Youtube` (bottom).

coefficient [4] defined as $\texttt{disp}(u) = \frac{|N_2(u)|}{\sum_{v \in N(u)} (|N(v)|-1)}$ and the *redundancy coefficient* [1] defined as $\texttt{rd}(u) = \frac{|\{(v,w) \in N(u) \times N(u) \text{ s.t. } \exists u' \neq u, (u',v) \in E \text{ and } (u',w) \in E\}|}{\frac{|N(u)|(|N(u)|-1)}{2}}$.

Figure 1 presents the results for the 3 datasets and the distribution of all characteristics considered. For all features examined here the tripartite models succeed in preserving the properties better than the configuration model applied on the bipartite structure. This is particularly true for the redundancy and dispersion coefficients.

References

1. M. Latapy, C. Magnien, and N. Del Vecchio. Basic notions for the analysis of large two-mode networks. Social Networks, 30(1):31–48, January 2008.
2. A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07), San Diego, CA, October 2007.
3. M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graphs with arbitrary degree distribution and their applications. Phys. Rev. E., 64, July 2001.
4. R. Tackx, F. Tarissan, and J.-L. Guillaume. Revealing intricate properties of communities in the bipartite structure of online social networks. In Proceedings of the 2015 IEEE 9th International Conference on Research Challenges in Information Science, RCIS'15, pages 321–326. IEEE, 2015.