On the importance of correlation in graph reconstruction

Simon Lizotte,[1,2] Jean-Gabriel Young,[2,3,4] and Antoine Allard[1,2,4]

[1] *Département de physique, de génie physique et d'optique, Université Laval, Québec, QC, Canada G1V 0A6*
[2] *Centre interdisciplinaire en modélisation mathématique, Université Laval, Québec, QC, Canada G1V 0A6*
[3] *Department of Mathematics and Statistics, University of Vermont, Burlington VT, USA*
[4] *Vermont Complex Systems Center, University of Vermont, Burlington VT, USA*

The structure of empirical networks is often unknown: we generally observe measurements of pairwise interactions and not the network itself. Some form of post-processing is needed to convert these data to networks. Recent work [1–3] shows that a Bayesian framework can be used to generate a distribution of graphs compatible with all the available information instead. Crucially, to keep solutions tractable, these work assume that the edges are conditionally independent, an assumption that can be violated in practice. For example, triadic closure and clustering are two well-known phenomena inducing correlations among edges in empirical networks [4, 5].

Here, we introduce a minimal Bayesian network reconstruction framework that can account for correlations. In the model, we account for correlation by using a hypergraph comprised of 2-edges and 3-edges. The 2-edges and 3-edges are supposed to exist independently *a priori* with probabilities q and p respectively. To obtain a pairwise description of the interactions, these hypergraphs are projected onto a graph with two edge labels: 2-edges become "regular edges" and 3-edges become a triangle of "correlated edges". The likelihood of the observations (e.g., number of interactions between two individuals) is then a Poisson mixture model where the projection labels determine the type of measurement made (no interaction, regular edge and correlated edge).

As an uncorrelated baseline, we also study a network model in which weak and strong ties are *a priori* independently distributed: each weak edge exists independently with probability q_1 and each strong edge exists independently with probability q_2 among the remaining unconnected pairs. Again, the strength of the interactions determines the type of measurements made.

We develop sampling algorithms for these two models and show how to fit them to empirical data. As an example, we use Zachary's karate club hypergraph obtained with Young et al. [6]. In a regime where the Poisson distributions are well separated, both models identify all effective edge types correctly. However, we see in Fig. 1 that when distributions start overlapping, the model with correlation can better reconstruct the structure using the information contained in the hypergraph neighbourhood of the pairwise interactions.

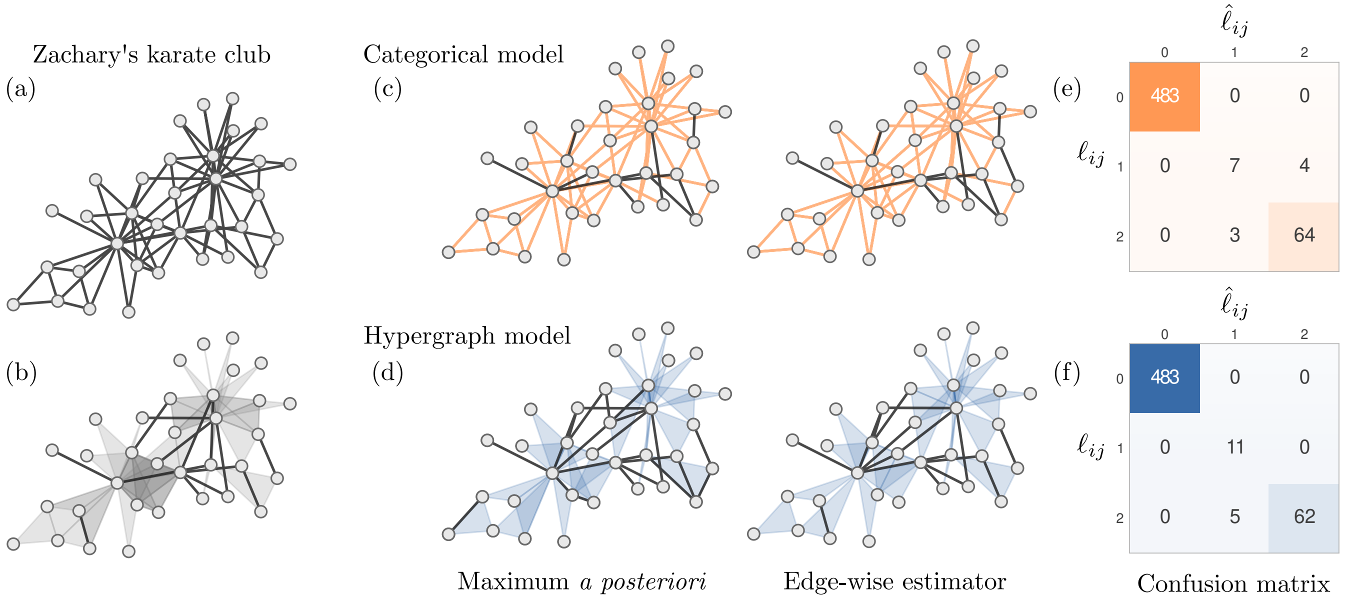

Figure 1: **Average structure *a posteriori*, maximum *a posteriori* (MAP) estimator and confusion matrix inferred by the hypergraph and multiplex graph models on Zachary's karate club hypergraph.** The synthetic pairwise observations were generated using 0.01, 20 and 30 as the averages of the Poisson distributions for unconnected pairs, 2-edges and pairs in 3-edges respectively. (a) Zachary's karate club (b) Hypergraph version of Zachary's karate club using Young and al. [6] algorithm (c) Results for the graph without correlation (d) Results for the model with correlation. The average structure is determined by the interactions which occur in more than half of the posterior sample and confusion matrices are based on the effective edge type of the average structure.

[1] Nat. Phys. **14**, 542–545 (2018).
[2] Phys. Rev. X **8**, 041011 (2018).
[3] J. Complex Netw. **8**, cnaa046 (2021).
[4] Phys. Rev. E **84**, 066117 (2011).
[5] Phys. Rev. E **90**, 042806 (2014).
[6] Commun. Phys. **4**, 1–11 (2021).