From Relational Data to Graphs

Inferring Significant Links using Generalized Hypergeometric Ensembles

Giona Casiraghi, Vahan Nanumyan, Ingo Scholtes, Frank Schweitzer
Chair of Systems Design
ETH Zürich
www.sg.ethz.ch

gcasiraghi@ethz.ch @gi_ona

1 Research Question

The inference of network topologies from relational data is an important problem in data analysis. Exemplary applications include the reconstruction of social ties from data on human interactions, the inference of gene co-expression networks from DNA microarray data, or the learning of semantic relationships based on co-occurrences of words in documents. Solving these problems requires **statistically principled techniques to infer significant links in noisy relational data.**

2 Generalized Hypergeometric Ensembles

We propose a new statistical modeling framework to address this challenge. The framework builds on **generalized hypergeometric ensembles**, a class of generative stochastic models that give rise to analytically tractable probability spaces of directed, multi-edge graphs [1-3].

Our construction of a statistical ensemble follows the general idea of the Molloy-Reed model, which is to randomly shuffle the topology of a given network while preserving node degrees. Different from this model, we assume a sampling of multi-edges such that the sequence of *expected degrees* of nodes is preserved. For this, for each n^2 pair of nodes i and j, we first define the maximum number Ξ_{ij} of multi-edges that can possibly exist between i and j. We then consider the construction of a random graph realization as an urn problem, where exactly *m* multi-edges are randomly sampled (without replacement) from an urn with n^2 balls with different colors. Each color represents the possible edges between a particular node pair.

For scenarios where we have additional information on factors that influence the formation of edges, we can further generalize the procedure as follows: We introduce a **propensity matrix Ω** whose entries Ω_{ij} capture relative tendency of a node i to form an edge specifically to node j. The probability distribution resulting from such a biased sampling process is given by the multivariate Wallenius' non-central hypergeometric distribution [4]:

$$\Pr(\mathbf{A}) = \left[\prod_{i,j} \binom{\Xi_{ij}}{A_{ij}}\right] \int_0^1 \prod_{i,j} \left(1 - z^{\frac{\Omega_{ij}}{S_\Omega}}\right)^{A_{ij}} dz$$

The probability to observe a particular number A_{ij} of edges between a pair of nodes i and j can be calculated from the marginal distribution as

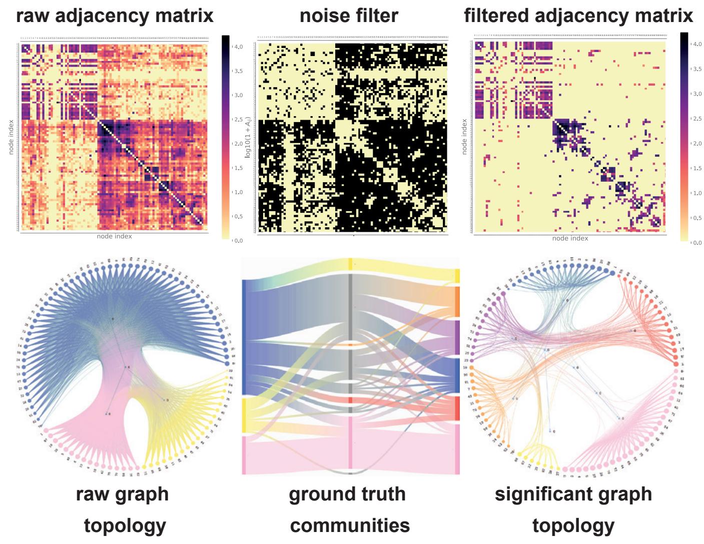
$$\Pr(A_{ij} = \hat{A}_{ij}) = \binom{\Xi_{ij}}{\hat{A}_{ij}} \binom{M - \Xi_{ij}}{m - \hat{A}_{ij}} \cdot \int_0^1 \left[\left(1 - z^{\frac{\Omega_{ij}}{S_\Omega}}\right)^{\hat{A}_{ij}} \left(1 - z^{\frac{\bar{\Omega}_{\setminus(i,j)}}{S_\Omega}}\right)^{m - \hat{A}_{ij}} \right] dz$$

We obtain a general framework of statistical ensembles which (i) allows to encode arbitrary a priori tendencies of nodes to interact, and (ii) provides an analytical expression for the probability to observe a given number of interactions between any pair of nodes.

For a given observed adjacency matrix Aij and a significance threshold α, this **allows us to identify significant links** by filtering the weighted adjacency matrix A by a threshold $\Pr(A_{ij} \leq \hat{A}_{ij}) > 1 - \alpha$ based on the equation above. This can be seen as assigning *p*-values to dyads *(i,j)*, obtaining a *high-pass noise filter* for entries in the adjacency matrix.

3 Inferring Significant Links from Noisy Data

We demonstrate how our framework can be used to infer significant links in two relational data sets. (RM) captures time-stamped proximities between students and faculty at MIT recorded via smart devices. (ZKC) covers frequencies of encounters between members of a university Karate club.



Illustrating our method, the figure above shows the entries of the (observed) adjacency matrix **Â** for (RM). Using a threshold of α=0.01 we use our methodology to obtain a high-pass noise filter (top center), where black entries correspond to pairs of nodes with non-significant links. Applying this filter to the original matrix yields a noise-filtered matrix (top right). While in the raw network (bottom left) there are 721,889 multi-edges amounting to 2,952 distinct links, after filtering there are 626 (21.2 %) significant links left (bottom right). A comparison of community structures detected by a stochastic block model confirms that **communities inferred in the filtered graph better correspond to ground truth communities** (bottom center).

4 Conclusion and Outlook

Our work makes three important contributions:

(1) We provide an **analytically tractable statistical model** of directed and undirected multi-edge graphs that can be used for inference and learning tasks.

(2) Our work highlights a - to the best of our knowledge - previously unknown **relation between random graph theory and Wallenius' non-central hypergeometric distribution**.

(3) Different from existing ensembles such as, e.g., the configuration model, our f**ramework can be used to encode prior knowledge** on factors that influence the formation of relations, thus tuning the random baseline.

Our method opens perspectives for a statistically principled network inference that accounts for effects that are not purely random. Our work advances the theoretical foundation for statistical relational learning. It also highlights that **model selection and hypothesis testing are crucial prerequisites that should precede the application of network analysis.**

5 References

1. G Casiraghi, V Nanumyan, I Scholtes, F Schweitzer: **Generalized Hypergeometric Ensembles: Statistical Hypothesis Testing in Complex Networks**, arXiv 1607.02441
2. G Casiraghi: **Multiplex Network Regression: How do relations drive interactions?**, arXiv 1702.02048
3. G Casiraghi, V Nanumyan, I Scholtes, F Schweitzer: **From Relational Data to Graphs: Inferring Significant Links using Generalized Hypergeometric Ensembles**, In Proc. of the 9th Intern. Conference on Social Informatics, 2017
4. KT Wallenius: **Biased Sampling: the Noncentral Hypergeometric Probability Distribution**, PhD thesis, 1963