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

ischoltes@ethz.ch @ingo_s
www.ingoscholtes.net

1 Research Question

The inference of graphs from relational data is an important problem in data mining and machine learning. 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 definition of a statistical ensemble follows the general idea of the configuration model, which is to randomly shuffle the topology of a given graph while preserving node degrees. Different from this model, our approach assumes a **(biased) sampling of multi-edges** such that the sequence of *expected node degrees* is preserved. Our approach works as follows:

- For each n^2 pair of nodes i and j, we 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 **introduce a propensity matrix Ω** whose entries Ω_{ij} capture the relative tendency of a node i to form an edge specifically to node j.

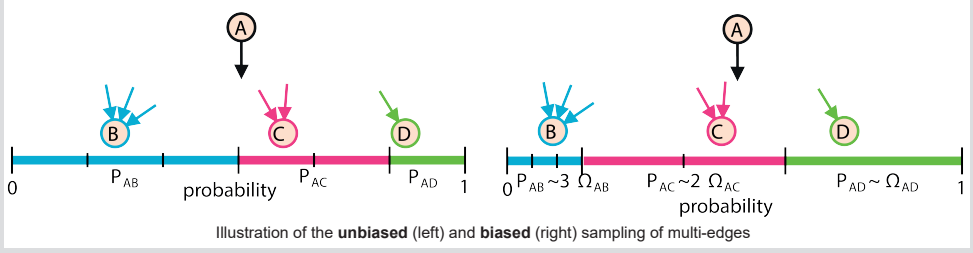


Illustration of the **unbiased** (left) and **biased** (right) sampling of multi-edges

The probability distribution resulting from such a biased sampling 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 $\hat{A}_{ij}$ of edges between a pair of nodes i and j can be calculated from the marginal distribution:

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

We obtain a framework of statistical ensembles which

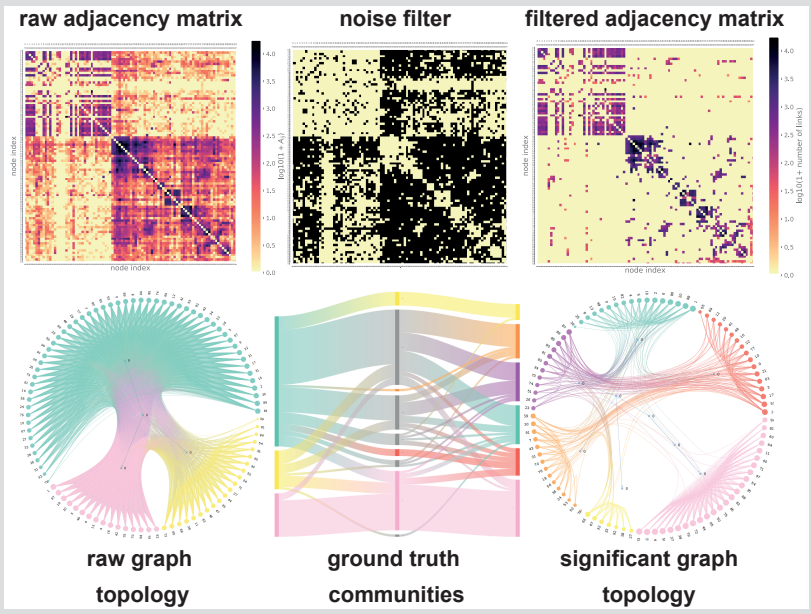
i) allows to encode arbitrary a priori tendencies of nodes to interact, and

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 Â and a significance threshold α, **this ensemble allows to identify significant links** by filtering the weighted adjacency matrix Â 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 the inference of significant links from noisy data in a **data set that captures time-stamped proximity relations between students and faculty** recorded via smart devices in the RealityMining project.



The figure above shows the entries of the (observed) adjacency matrix Â (top left). Using α=0.01 we use our method to calculate a high-pass noise filter (top center), where black entries correspond to non-significant links. Applying this filter to Â yields a noise-filtered adjacency matrix (top right). The raw graph (bottom left) has 721,889 multi-edges amounting to 2,952 distinct links. The significant graph (bottom right) has 626 significant links (21.2 % of the original graph). A visual comparison of clusters detected by a stochastic block model confirms that **communities in the filtered graph better correspond to the ground truth** (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 previously unknown **relation between random graphs and Wallenius' non-central hypergeometric distribution**.

3) Different from existing ensembles such as, e.g., the configuration model, our **framework 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 inference of graphs that accounts for effects that are not purely random. With this **we advance the theoretical foundation of statistical relational learning**. Our work 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