ETH zürich

Multiplex Network Regression

a Statistical Framework for Multidimensional Data Analysis

Giona Casiraghi

Chair of Systems Design
ETH Zürich
gcasiraghi@ethz.ch @gi_ona

1 Introduction

Complex networks are often generated from datasets of observed, repeated interactions. The question is whether these interactions are *random events*, or whether they are *driven by existing relations* between the nodes. We propose a new nonlinear parametric model to perform **statistical regression on networks**.

2 Multiplex Perspective

Suppose that we have a dataset consisting of *m* recorded interactions between *n* elements and *r* different types of relations between them.
We can encode the interactions in a graph with *n*=|*V*| vertices and *m* edges.

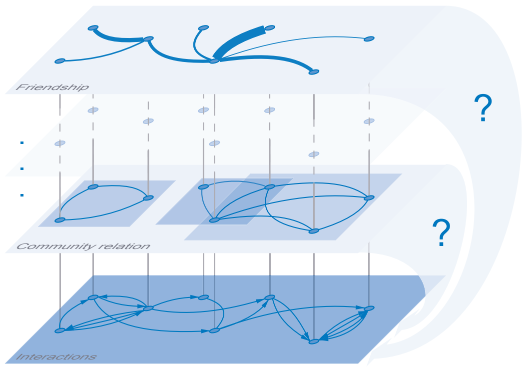


Fig. 1. *Multiplex network* representation of a system, where different types of ***relation layers*** governs the interactions between system elements. But how do relations drive interactions?

Since two individuals may interact more than once, multiple edges may exist between the same couple of vertices, giving rise to a **multi-edge** graph.
In the following we will refer to this graph as the interaction layer *I*.

For each type of the *r* relations, we can generate a graph that encodes dyadic relations between the elements of the system as **weighted edges** between vertices.
The weight of each edge encodes the strength of the relation.
We will refer to these *r* graphs as the relational layers. We can now define the multiplex network **M** generated by the *r+1* layers and *n*=|*V*| vertices.

3 Regression with Generalised Hypergeometric Ensembles

The method is based on an extension to multiplex networks of the **generalised hypergeometric ensembles**. These define a class of analytically tractable ensembles for weighted directed networks that we have recently developed. They contain random graphs generated by **merging combinatorial effects** and **arbitrary relations** between vertices.
I is then distributed according to the Wallenius non-central hypergeometric distribution:

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

with

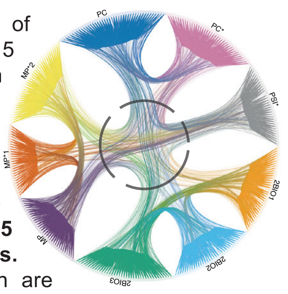
$$S_\Omega = \sum_{i,j}\Omega_{ij}(\Xi_{ij} - A_{ij})$$

The distribution above is defined by the two quantities Ξ and Ω. Ω encodes the **propensity** of pairs of vertices to connect, and Ξ the probability that pairs of vertices are connected because of **combinatorial effects**. We assume the entries of the matrix of Ξ are built according to the configuration model, the most general way to encode combinatorial effect generated by the different activity level of vertices. It means that **vertices that are more active**, i.e., have **higher degree**, are **more likely to interact**. Hence, Ξ is completely defined by *I*. On the other hand, Ω depends on the relational layers as follows, where the parameters β are **fitted to the data** by means of MLE.

$$\Omega := \prod_{l=1}^{r}\mathbf{R}_l^{\beta_l}$$

4 Application to SocioPatterns

Recorded contacts of **327 students** over 5 days, represented in the graph of interactions on the right. The dataset contains 5 additional types of relations between the students encoded as **5** more **relational layers.**



Results of regression are reported in the table below with pseudo-r-squared.

Coefficients:

	Estimate	*Std.Err*	*odds*	
gender	0.0771632	0.0021205	1.1944368	***
class	1.3919905	0.0047051	24.6598537	***
topic	0.9995999	0.0089452	9.9907912	***
friendship	0.6954883	0.0024251	4.9600760	***
facebook	-0.0888839	0.0023650	0.8149221	***

R2:

McFadden R2	*Cox Snell R2*
0.5472486	0.9800223

5 Conclusion

The method allows to *quantify the influence* that each layer, i.e. the independent variables, has *on the interaction counts*, i.e. the dependent variable.
Additionally, we can quantify the **significance** of the estimated model, its **goodness-of-fit** and the **effect size** of each parameter fit.

The model separates **random** from **deterministic influences** on interactions. We hence identify how much known relations drive the interactions.
In conclusion, the method proposed is a major advance for the analysis of relational datasets and complex networks.

6 References

1. Casiraghi, G. Multiplex Network Regression: How do relations drive interactions? arXiv Prepr. arXiv1702.02048 (2017)
2. Casiraghi, G., Nanumyan, V., Scholtes, I. & Schweitzer, F. Generalized Hypergeometric Ensembles: Statistical Hypothesis Testing in Complex Networks. arXiv Prepr. arXiv1607.02441 (2016)
3. Mastrandrea, R., Fournet, J. & Barrat, A. Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. PLoS One 10, 1–26 (2015).