# Multiplex Network Regression: How do relations drive interactions?

**Giona Casiraghi***

February 8, 2017

**Abstract**

We introduce a statistical method to investigate the impact of dyadic *relations* on complex networks generated from *repeated interactions*. It is based on generalised hypergeometric ensembles, a class of statistical network ensembles developed recently. We represent different types of known relations between system elements by weighted graphs, separated in the different layers of a *multiplex network*. With our method we can regress the influence of each relational layer, the independent variables, on the interaction counts, the dependent variables. Moreover, we can test the *statistical significance* of the relations as explanatory variables for the observed interactions. To demonstrate the power of our approach and its broad applicability, we will present examples based on synthetic and empirical data.



## 1 Introduction

We often deal with datasets of *observed repeated interactions* between elements of a system. These datasets are used to generate networks where the elements are represented by vertices and interactions by edges. We ask whether these interactions are random events or whether they are driven by existing relations between the elements. To answer this question, we propose a statistical model to regress *relations*, which we identify as *independent variables*, on a network created from interactions, which we will refer to as *dependent variables*.

In general, a regression model explains dependent variables as a function of the independent ones, accounting for random effects. Here, we assume that the observed interactions are driven by different relations which are masked by *combinatorial effects*. With combinatorial effects, we mean that elements that interact more in general are also more likely to interact with each other, even if they have no relations. This problem is well known in network theory, where it is referred to as *degree-correction* (see e.g. [10, 16, 18]). For example, the fact that two individuals have contact very often can be explained by multiple reasons. They may interact because they are friends, because they work together, or simply because they are very active, and hence have high chances to meet. Therefore, to have a full understanding of the system, we have to disentangle relations from combinatorial effects.

*E-mail: gcasiraghi@ethz.ch. Chair of Systems Design, ETH Zürich, Weinbergstrasse 56/58, 8092 Zürich, Switzerland.

Datasets of interactions are ubiquitous across disciplines. Examples of these are recorded contacts between individuals (e.g. SocioPatterns [12, 23], Reality Mining [5]), mutualistic interactions between species in ecology [4, 14], economical transactions between countries and firms [8, 17], and collaborations between firms [24]. In these cases, researchers are interested in learning whether the observed interactions are driven by relations between the elements of the system. They ask whether friendship plays a role in the contacts between students [12], whether *homophily* drives interactions within social networks, i.e. whether individuals sharing similar characteristics are more likely to interact [13], or whether collaborations between companies are driven by geographical distance or industrial sector similarity [24].

There exist few statistical models addressing the problem of *quantifying* the interdependence between observed edges and dyadic relations. This problem is exacerbated by the fact that the dyadic relations represented in complex networks are not independent from one another.

Because of the non-independency of dyadic relations, ordinary least squares regression models are inappropriate to analyse network data [11]. To partially overcome their limits, Krackhardt [11] introduced a regression method based on the quadratic assignment procedure developed by Hubert and Schultz [9]. Other statistical methods commonly used in the analysis of social networks are based on exponential random graph models (ERGMs) or on their extensions (see e.g. [19–22]). Although being effective under specific conditions, all these methods have been developed for *unweighted* graphs. This means that they are not suited for datasets which contain repeated interactions, that need to be represented as *integer-weighted* graphs. The solution to this issue is to *threshold* the interactions to obtain an unweighted graph (e.g. [3]). Clearly, this approach does not exploit all the information available in the data, and therefore may produce sub-optimal results [1]. A step forward in the analysis of weighted networks has recently been proposed in [6]. The authors introduce a Bayesian approach to edge formation, which allows to encode a broad class of hypothesis for the formation of weighted edges in complex networks. However, this approach does not take into account the combinatorial effects found in interaction data.

We can hence summarise the limitations of existing methods into two main issues. First, many of them are not appropriate for weighted graphs. Second, they do not take into account the combinatorial effects typical of interaction data. To solve these problems, in this article we propose a new model to perform statistical regression on networks. Our method is based on an extension of generalised hypergeometric ensembles (gHypE), a class of statistical network ensembles we have recently introduced (see [2]). gHypEs contain random graphs generated by merging arbitrary relations between vertices and combinatorial effects. Moreover, thanks to their analytical formulation, we can statistically test the *significance* of the regression model against the observed interactions.

We demonstrate the power of our approach and its broad applicability with examples based on synthetic and empirical data. As case study, we will use the SocioPattern dataset provided in [12]. The data available consist of an interaction network, built from recorded contact counts between high-school students, and of further information such as student's gender, class membership and topic, self-reported friendship relations, and Facebook connections.

# 2 Methodology

## 2.1 Network Representation

Relational datasets as the one provided in [12], consist of interaction counts and a collection of dyadic relations and vertex attributes. Vertex attributes, such as community membership or gender, often yield strong relations between individuals, as individuals in the same community tend to interact more than individuals in different ones. We can study this type of data representing it as a *multiplex network*. Multiplex networks are a special class of interconnected multilayer networks where the vertices of each layer correspond (cf. fig. 1) [7].

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. 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 $\mathcal{I}$. For each type of relation, we can generate a graph that encodes the 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 $\mathcal{R}_l$ with $l \in [1, r]$. Let now $\mathcal{M}$ be the multiplex network generated by the $r + 1$ layers and $n = |V|$ vertices. Figure 1 illustrates the multiplex approach we take.

In the following, we propose a framework to perform statistical regressions with network layers. We assume the multi-edged graph $\mathcal{I}$ to be the dependent variable and the remaining layers $R_l$ to be the independent – explanatory – variables. The model that results has the following form:

$$\mathcal{I} = f(\mathcal{R}_1, \ldots, \mathcal{R}_r; \beta_1, \ldots, \beta_r), \tag{1}$$

for some function $f : \mathbb{R}^{V\times V} \times \cdots \times \mathbb{R}^{V\times V} \times \mathbb{R}^r \to \mathbb{N}^{V\times V}$, where the parameters $\beta_l$, $l \in [1, r]$ are the parameters of the regression model corresponding to each layer $\mathcal{R}_l$.

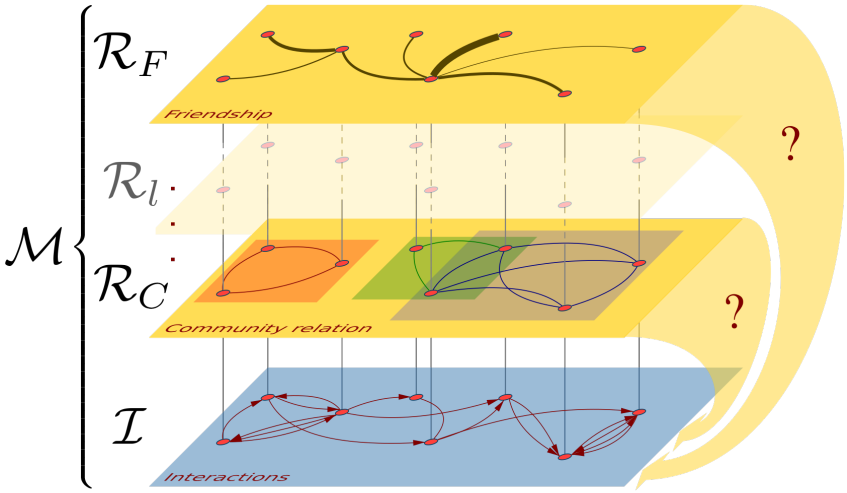


Figure 1: The multiplex network representation of a relational dataset. The bottom layer (blue) captures the interaction counts that are observed. The top layers (yellow) encodes different types of relations, like weighted friendship links, or community membership. The model we propose allows us to understand *how* these relational layers impact interactions.

## 2.2 Statistical Model

We want to model the interaction layer $\mathcal{I}$, which is a multi-edged graph with fixed number of edges $m$. To do so, we treat $\mathcal{I}$ as a realisation from a *generalised hypergeometric ensemble* $\mathbb{E}(n, m)$ [2], with $n$ vertices and $m$ edges. We indicate with $\mathbf{A}$ the adjacency matrix of the interaction layer $\mathcal{I}$ and $A_{ij}$ with $i, j \in V$, its elements. Similarly, let $\mathbf{R}_l$ be the adjacency matrix of the relational layer $\mathcal{R}_l$ and with $\beta \in \mathbb{R}^r$ the $r$-vector of regression coefficients. $\mathcal{I}$ is then distributed according to the Wallenius non-central hypergeometric distribution [2, 25]

$$\Pr(\mathcal{I}|\mathcal{R}) = \left[\prod_{i,j}\binom{\Xi_{ij}}{A_{ij}}\right]\int_0^1 \prod_{i,j}\left(1 - z^{\frac{\Omega_{ij}}{S_\Omega}}\right)^{A_{ij}} dz \tag{2}$$

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

The distribution in eq. (2) is defined by the two quantities $\Xi$ and $\Omega$. $\Omega$ encodes the propensity of pairs of vertices to connect, and $\Xi$ the probability that pairs of vertices are connected because of combinatorial effects, as described in [2]. We assume the entries of the matrix of possible edges $\Xi$ are built according to the configuration model. This is the most general way to encode combinatorial effect generated by the different activity, i.e. degree, of vertices. It means that vertices that are more active, i.e. have higher degree, are more likely to interact. Hence, $\Xi$ is

completely defined by $\mathcal{I}$. On the other hand, $\Omega$ depends on the relational layers $\{\mathcal{R}_l\}_{l\in[1,r]}$ as follows:

$$\Omega := \prod_{l=1}^{r} \mathbf{R}_l^{\beta_l}. \tag{3}$$

We can now specify the statistical model in eq. (1). We take $f$ as the expectation of the hypergeometric network ensemble $\hat{\mathbb{E}}(n,m)$ that maximises the probability of observing $\mathcal{I}$, given the relational layers $\{\mathcal{R}_l\}_{l\in[1,r]}$:

$$\mathcal{I} = \mu[\hat{\mathbb{E}}(n,m)|\mathcal{R}_1,\dots,\mathcal{R}_r]. \tag{4}$$

Estimating the model in eq. (4) is therefore equivalent to find maximum likelihood estimators (MLE) for the parameter vector $\beta$ in eq. (2).

Equation (2) shows that the likelihood of $\beta$ given the observed graph $\mathcal{I}$ is defined by

$$L(\beta|\mathcal{I}) = \left[\prod_{i,j}\binom{\Xi_{ij}}{A_{ij}}\right]\int_0^1 \prod_{i,j}\left(1 - z^{\frac{\prod_{l=1}^{r} R_{l,ij}^{\beta_l}}{S_\beta}}\right)^{A_{ij}} dz \tag{5}$$

with $S_\beta = \sum_{i,j}\prod_{l=1}^{r} R_{l,ij}^{\beta_l}(\Xi_{ij} - A_{ij})$.

Although the numerical maximisation of eq. (5) is difficult, for $m \ll \sum_{ij}\Xi_{ij}$ we can approximate the multivariate hypergeometric distribution with a multinomial distribution. Therefore eq. (5) as a function of $\beta$ can be approximated up to constants by

$$L(\beta|\mathcal{I}) \sim \prod_{i,j\in V}\left(\frac{\Xi_{ij}\prod_{l=1}^{r} R_{l,ij}^{\beta_l}}{\sum_{i,j\in V}\Xi_{ij}\prod_{l=1}^{r} R_{l,ij}^{\beta_l}}\right)^{I_{ij}}. \tag{6}$$

We obtain the MLE $\hat{\beta} = \text{argmax}_\beta(L(\beta|\mathcal{I}))$ of eq. (6) by solving numerically the system given by $\nabla L(\beta) = 0$. Each component of the gradient of the log-likelihood $\nabla\log(L(\beta))$ is then given by

$$\frac{\partial\log(L(\beta|\mathcal{I}))}{\partial\beta_l} = -m\frac{\sum_{ij}\log(R_{l,ij})\Xi_{ij}\prod_{l=1}^{r} R_{l,ij}^{\beta_l}}{\sum_{ij}\Xi_{ij}\prod_{l=1}^{r} R_{l,ij}^{\beta_l}} + \sum_{ij} I_{ij}\log(R_{l,ij}) \tag{7}$$

### 2.3 General Regression Model

The model described in the previous section can be generalised to account for multiple observations of the multiplex $\mathcal{M}$. For example, suppose we have data about contacts between students in a school, and we have collected the same type of data for different schools. Let's assume now we want to learn whether gender plays the same role in the interactions across all the schools. This implies that while the relations between the individuals change for different observations, e.g.

gender distribution in different schools, the effect that the relations have on the interactions remains constant. In other words, the relational layers change for each observation, i.e. $\mathcal{R}^{(i)} \neq \mathcal{R}^{(j)}$, where $i$ and $j$ are different observations. On the other hand, the parameter $\beta$ quantifying the effect of the relations on the interactions is assumed to be constant, i.e. $\beta^{(i)} = \beta^{(j)} = \beta$.

Suppose to have $N$ independent observations of the multiplex $\mathcal{M}$, each denoted $\mathcal{M}^{(i)}$. We assume that the influence of the independent layers $R_l^{(i)}$ on the dependent layer $\mathcal{I}^{(i)}$ is fixed, i.e. for each observation $i$, $\beta^{(i)} = \beta \, \forall i \in N$.

Since each observation $\mathcal{I}^{(i)}$ is independent and follows the distribution of the gHypE (eq. (2)), the joint probability distribution is just the product of each probability. Therefore the likelihood of the parameter vector $\beta$ is given by

$$L(\beta|\mathcal{I}^{(0)}, \mathcal{I}^{(1)}, \ldots, \mathcal{I}^{(N)}) := \prod_{i=1}^{N} L(\beta|\mathcal{I}^{(i)}), \tag{8}$$

where $L(\beta|\mathcal{I}^{(i)})$ is defined as in eq. (5). It is worth noting that the interaction layers $\mathcal{I}^{(i)}$ come from the same class of distribution but are not identically distributed. This is true unless the number of edges $M^{(i)} = M$ and the matrix $\Xi^{(i)} = \Xi$ are constant for each observation $(i)$.

Given the likelihood in eq. (8), we can derive the MLE $\hat{\beta}$ of the parameter $\beta$. Denoting with $l(\beta|\mathcal{I}^{(i)})$ the log-likelihood of $\beta$ and by

$$\hat{l}(\beta) = \frac{1}{N} \sum_{i=1}^{N} l(\beta|\mathcal{I}^{(i)}) \tag{9}$$

the average log-likelihood, $\hat{\beta}$ is defined as follows:

$$\hat{\beta} = \text{argmax}_{\beta} \left( \hat{l}(\beta) \right). \tag{10}$$

## 2.4 Model Selection and Significance

Recall we have a multiplex $\mathcal{M}$ with $r + 1$ layers. Suppose we have estimated the statistical regression model defined in eq. (4). We know the MLEs $\{\hat{\beta}_l\}_{l\in[1,r]}$ corresponding the $r$ relational layers $\{\mathcal{R}_l\}_{l\in[1,r]}$, and each of their values quantifies the *strength* of the effect each layer has on the interaction layer $\mathcal{I}$. Are all these parameters significant? In other words, we want to quantify the *statistical significance* of the model with all parameters $\{\hat{\beta}_l\}_{l\in[1,r]}$, and compare it to a model with less parameters. This allows to select the statistical model with the highest significance, and to disregard the parameters and the layers with non-significant effect on the interactions.

We want to compare which of two statistical models defined by the sub-multiplexes $\{\mathcal{R}_l\}_{l\in[1,q]}$ and $\{\mathcal{R}_l\}_{l\in[1,q+s]}$ as in eq. (1), one with $q$ and the other one with $q + s$ relational layers, better describes the observed interaction layer $\mathcal{I}$.

Both models are described by eq. (5) with the appropriate layers chosen as predictors. The two models are nested, as one is a special case of the other. In fact, the model defined by $\{\mathcal{R}_l\}_{l\in[1,q]}$ can be obtained by setting to 0 the $s$ coefficients $\{\beta_l\}_{l\in[q,q+s]}$ corresponding to the $\{\mathcal{R}_l\}_{l\in[q,q+s]}$ layers in the second model (cf. eq. (3)).

We can perform *model selection* by means of the *likelihood ratio test*, which is the most powerful test at significance level $\alpha$, according to the Neyman-Pearson lemma. In particular, we can identify the null hypothesis $H_0$ by the model defined by $\{\mathcal{R}_l\}_{l\in[1,q]}$ with $\tilde{q}$ parameters, and the alternative hypothesis $H_1$ by the model defined by $\{\mathcal{R}_l\}_{l\in[1,q+s]}$, with $\tilde{s}$ more parameters. This allows to test whether the explaining power of the more complex model with $\tilde{q}+\tilde{s}$ parameters is high enough to justify the increase in complexity.

The likelihood ratio statistic $\Lambda(\mathcal{I})$ is defined as

$$\Lambda(\mathcal{I}) = \frac{L(\beta_0|\mathcal{I})}{L(\beta_1|\mathcal{I})} = \frac{L(\beta_q|\mathcal{I})}{L(\beta_{q+s}|\mathcal{I})}. \tag{11}$$

According to eq. (5), $\Lambda(\mathcal{I})$ is given by

$$\Lambda(\mathcal{I}) = \frac{\int_0^1 \prod_{i,j} \left(1 - z^{\frac{\prod_{l=1}^{q} R_{l,ij}^{\beta_l}}{S_{\beta_0}}}\right)^{A_{ij}} dz}{\int_0^1 \prod_{i,j} \left(1 - z^{\frac{\prod_{l=1}^{q+s} R_{l,ij}^{\beta_l}}{S_{\beta_1}}}\right)^{A_{ij}} dz}. \tag{12}$$

Given the two nested models described above, and chosen a significance level $\alpha$, we will select the more complex model only if we can reject the null hypothesis by computing the p-value corresponding to $\Lambda(\mathcal{I})$.

In the case of the regression problem described in section 2.3, with multiple observation of the multiplex $\mathcal{M}$, the same procedure applies. In fact the models are still nested and the likelihood ratio test can be performed by choosing the joint likelihood functions as in eq. (8) to compute the likelihood ratio statistics.

The likelihood ratio distribution converges to a $\chi^2$ distribution with $d$ degrees of freedom, in accordance with Wilks' theorem [26], as the number of observations $N$ and the number of observed dyads $A_{ij} \to \infty$. In fact, if the network is large enough, i.e. if the number of non-zero $\Xi_{ij}$ is large, Wilks' theorem holds even for a single observation ($N = 1$) and the likelihood ratio distribution can be approximated by the $\chi^2$ distribution. The number of degrees of freedom $d$ is equal to the difference in the number of parameters $(q + s) - s = s$ between the two models, plus the number of degrees of freedom of the $\{\mathcal{R}_l\}_{l\in[q,q+s]}$ additional layers. By performing a stepwise selection, we can find the best model, where only the significant layers are used. In each step of

the selection process, a layer is added to or removed from the model according to the result of the likelihood-ratio test.

Once the best model has been selected, we can quantify the goodness of the model itself by means of its Mahalanobis distance, as described in [2].

# 3 Applications

We showcase our method with two case studies. In the first one we illustrate how we can detect *homophily*, i.e. the tendency of individuals to interact with similar others, in a network of repeated interactions. In the second case study, we apply our technique to the SocioPattern dataset [12], to measure the strength and the significance of the effect of each layer of information provided on the observed number of interactions.

## 3.1 Homophily detection

In this first example, we analyse a synthetic dataset. We generate random interactions between individuals divided into two groups, encoding a preference for individuals of the same group to interact more between each other than with the other group. Our method allows to estimate the relative intensity of the homophily encoded in the data, i.e. how much individuals of the same group are more likely to interact within the group than with individuals from the other group. Furthermore, we can identify a *threshold* for its detectability in graphs from repeated interactions.

We generate 500 replicates of a random graph with $n = 200$ vertices divided into 2 equally sized groups. We place $m = 5000$ edges at random, such that the probability to create an edge between two vertices of the same group is $p_1$ and the probability to create an edge between two vertices of different groups is $p_2$. There is homophily if the odds-ratio $\omega = p_1/p_2 > 1$, i.e. if the probability to observe an edge between two vertices of the same group is $\omega > 1$ times higher than that to observe an edge between vertices of different groups. An example of adjacency matrix with strong homophily ($\omega = 18$) is provided in fig. 2.

In this setup, the interaction layer corresponds to the synthetic graph, while there is only one relational layer $\mathcal{R}_1$ that encodes the presence of the homophily relation. We can encode homophily in the relational layer $\mathcal{R}_1$ by assigning $R_{1,ij} = 1$ if $i, j$ are in the same group, and $R_{1,ij} = \epsilon < 1$ if $i, j$ are in different groups.

To test whether the effect of homophily on the interaction is significant, we proceed as described in section 2. We build a statistical model with one independent variable $\mathcal{R}_1$ and one parameter

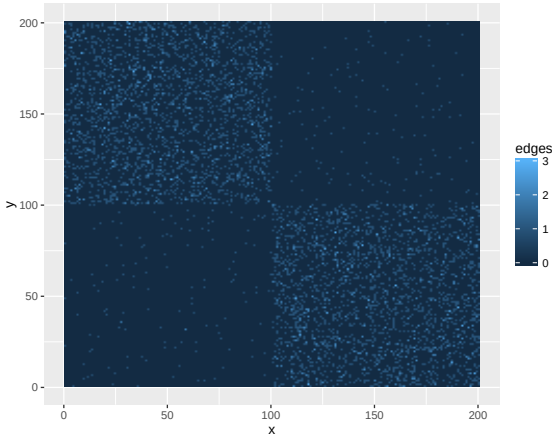


Figure 2: The adjacency matrix of a random graph with 200 nodes and 5000 edges divided into 2 identical groups. It shows homophily characterised by an odds-ratio $\omega = 18$ between in-group and out-group relations.

$\beta_1$. We can then obtain the MLE $\hat{\beta}_1$ as described in section 2.2. Finally, we can perform a likelihood-ratio test to test for the presence of homophily.

We build the test as follows. The null-hypothesis is the absence of homophily. Hence we assume homogeneous relations between vertices, i.e. $R_{0,ij} = 1$ for all vertices $i, j$. The alternative hypothesis is the presence of homophily, encoded as described above. The closer to 0 the maximum likelihood estimator $\hat{\beta}_1$ is, the weaker the observed homophily. The greater $\hat{\beta}_1$ is, the stronger the observed homophily.

In the case of $\omega = 18$ (cf. fig. 2), we can reject the null-hypothesis with p-value of 0 (quantile $> 4000$ of a $\chi^2$ with 2 degrees of freedom).

In this example, we can identify a threshold of detectability for homophily of $\omega = 1.11$, i.e. the null-hypothesis is rejected at 0.05 level 454 out of 500 tests ($> 90\%$), performed with random replicates generated using the same parameters. This threshold can be further lowered by increasing the number of observations available to perform the regression ($N > 1$). Compared to methods based on *modularity* (e.g. [15]), this approach allows to detect homophily with a very low threshold. In fact, the average value for modularity is 0.0266 on 500 replicates, which, in the absence of other knowledge, does not allow to say anything about the structure of a single observed graph.

To quantify the intensity of the homophily relation we can look at the value of $\epsilon^{\hat{\beta}_1}$. The smaller it is, the stronger the relation. Moreover, in this simple example we can recover the odds-ratio

$\omega = p_1/p_2$. By taking the ratio $\hat{\omega} = 1/\epsilon^{\hat{\beta}_1}$, we can estimate $\omega$. In particular, in the case of $\omega = 18$, we obtain $\hat{\omega} \approx 18.06$ averaging over 500 random replicates.

### 3.2 High school contacts analysis

In the second case study, we analyse the dataset provided in [12]. The data consists of recorded contacts between 327 students over 5 days that we represent in the graph of interactions. The dataset contains 5 additional types of relations between the students that we encode as 5 relational layers.

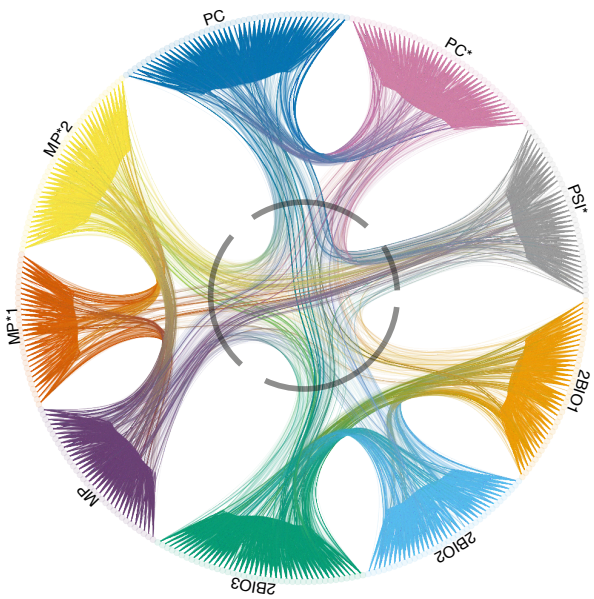


Figure 3: The graph obtained from the contacts between students. Each student is coloured according to its class membership and the internal ring groups classes on similar topic. From this figure it is clear that most of contacts happen between students of the same class, and there is a preference for contacts between students attending classes on the same topic.

A first relational layer $\mathcal{R}_C$ reflects the separation of students into the 9 classes. We want to limit the encounters between students attending different classes, as it can be observed in fig. 3. To build $\mathcal{R}_C$ we can set $R_{C,ij} = 1$ if $i, j$ are in the same class, and $R_{C,ij} = \epsilon < 1$ if $i, j$ are in different classes.

A second relational layer $\mathcal{R}_T$ is built according to the topic of the different classes students take. The 9 classes can be grouped into 4 topical areas, of 3,3,2,1 classes respectively. There are 3 classes of type "MP" (MP, MP01, MP02), two of type "PC" (PC and PC0), one of type

"PSI" (PSI0) and 3 of type "BIO" (2BIO1, 2BIO2, 2BIO3). This separation is highlighted by the internal ring in fig. 3. The layer $\mathcal{R}_T$ is defined similarly to $\mathcal{R}_C$, setting $R_{T,ij} = 1$ if $i, j$ attend classes in the same topical area, and $R_{T,ij} = \epsilon < 1$ if not.

A third relational layer $\mathcal{R}_G$ is built using the gender of the students. We hypothesise homophily would partially play a role in student interactions. We build the graph as described for the synthetic example, in section 3.1.

The dataset provides also information about actual friendship relations between the students. We build the fourth layer $\mathcal{R}_f$ using the social networks obtained from *self-reported friendship relations*, and a fifth layer $\mathcal{R}_F$ using the provided *Facebook connections*. We assume the absence of an edge in either $\mathcal{R}_f$ or $\mathcal{R}_F$ is not enough to disallow an interaction to happen. Hence, we set the weight of the relations between students which are not "friends" in either of the two layers to $\epsilon < 1$.

We speculate that $R_C$ will have a very strong influence on the interactions, since the division into classes act as sort of physical boundaries. In general, we would assume the information provided by the two friendship layers will be comparable, as the reported friendship relations should be part of the Facebook connections. However, in the dataset studied the two layers have a low correlation (0.115), due to the fact that nearly half of the friendship relations (326 out of 668) are not accounted by the 4515 Facebook connections. Hence, we cannot predict whether one of the two layers does or does not have a significant effect on the interactions, without fitting the regression model.

We build a regression model with 5 predictors as described in section 2. The MLEs of the parameters corresponding to the 5 layers are given in table 1. As we expected, from the result of

| | MLE | Significance | $\omega$ |
|---|---|---|---|
| $\beta_C$ | 1.37171072 | $\star\star\star$ | 23.5348113 |
| $\beta_T$ | 0.92233844 | $\star\star\star$ | 8.3625444 |
| $\beta_G$ | 0.08556497 | $\star\star\star$ | 1.2177692 |
| $\beta_f$ | 0.74534302 | $\star\star\star$ | 5.5634350 |
| $\beta_F$ | −0.08775806 | − | 0.8170374 |

Table 1: Fitted parameters for the 5-layer model and the significance of the respective sub-models, obtained by adding one layer after the other. 3 stars correspond to a p-value $p < \alpha = 0.05$, the dash to p-value $p > \alpha = 0.1$. For the regression we used $\epsilon = 0.1$. The third column report the odds-ratio for each parameter.

the regression we can see a strong effect of the separation between classes. The value of $\beta_C \gg 0$ implies an odds-ratio of 23.5348113 for the probability of an encounter between classmates against

an encounter of students of different classes, given everything else equal. This means that there are 23 more chances that two classmates meet, compared to encounters between students of different classes. Class topics is also a driving force for the interactions. A contact between students attending classes on the same topic is 8 times more likely to be observed than a contact between students attending classes on different topics. The value of $\beta_G$ supports the presence of weak homophily in the encounters between students, with an odds-ratio of 1.2177692. The effect of self-reported friendship is also positive as expected, however the negative value of $\beta_F$, close to 0, hints that Facebook connections have little effect on the encounters.

We can proceed to study the *significance* of the effect provided by each relational layer. To do so we follow a stepwise selection method as described in section 2.4. We introduce one effect after the other, and we test whether the model with the added effect is better than the one without. The test is the likelihood-ratio test described above, where the degrees of freedom are given by the difference of the number of parameters between the new model that we want to test and the previous one. The first effect is tested against a null-model where the interaction graph is assumed to come from a gHypE with $\Omega = 1$. The test for the gender layer has 2 degrees of freedom, one to identify the gender groups and one to quantify the intensity of homophily. The test for the classes layer has 9 degrees of freedom, 8 to identify the classes and 1 to quantify the intensity of the separation (we assume that the intensity of the relation driving the interactions is homogeneous across classes). The tests for the friendship layer and the Facebook layer have both as many degrees of freedom as relations plus 1.

We find that the effect of the Facebook layer $\mathcal{R}_F$ is the only non-significant, cf. table 1. We can suppose that the information provided by the self-reported friendship relations $\mathcal{R}_f$ and $\mathcal{R}_F$ is similar. In fact, a model obtained by adding only the layer $\mathcal{R}_F$ instead of $\mathcal{R}_f$ *is* significant. However, when both relational layers are used as independent variables, only the layer $\mathcal{R}_f$ of self-reported friendship relations is significant, thus showing that the information provided by $\mathcal{R}_F$ can be discarded.

From this example we can conclude that in the dataset studied, the observed interactions are strongly influenced by the subdivision into classes and topic, as can also be visualised in fig. 3. Gender homophily is also relevant, and there is a significant positive effect of friendship on the contacts recorded.

# 4 Conclusion

In this article, we have proposed a new statistical model to quantify how observed interactions depend on relations, in the framework of multiplex networks. The model is based on the assumption that interactions between elements of a system are driven by two factors. The first factor is

the existence of relations between elements, such as friendship or homophily. The second is the combinatorial randomness caused by the activity of the elements. Elements that are more active are more likely to interact with each other, even if they are unrelated.

Different from common approaches used in network analysis, our methodology has been specifically designed to deal with multi-edge graphs. It therefore allows to use the whole data available, without the need of thresholding it to obtain unweighted, i.e. binary, graphs. In fact, repeated interactions between elements of a system generate multi-edge graphs, where the vertices correspond to the elements of the system. Similarly, relations can have varying intensity and should be encoded in weighted graphs. This is why thresholding the data into binary networks can be a waste of useful information.

Our model separates random and deterministic influences on interactions, accounting for the randomness as combinatorial effects. We hence identify *how much* known relations drive the interactions. To achieve this, we base our regression model on generalised hypergeometric ensembles, a class of statistical network ensembles we have recently introduced. The formulation of our model allows to estimate the strength of the dependence between relations and interactions, together with its statistical significance.

Studying how different relations drive observed interactions is not only necessary to increase the understanding of a system, it is also needed to control the dynamics of a system. In fact, to do so we have to appropriately modify the relations that are the driving forces underlying its behaviour. Similarly, if we want to increase the resilience of a system, we want to affect the relations that are responsible for its weaknesses. Having a clear understanding on how and which relations impact the behaviour of the elements of a system is a necessary condition to properly control it.

In conclusion, the method we propose is a major advance for the analysis of relational datasets and complex networks. By allowing the study of multi-edge and weighted graphs, it increases the breadth of applicability of network theory. In future work, it will allow to identify missing interactions, according to null-models based on known relations. Thanks to this, it will be possible to uncover *unknown relations* between elements of a system.

## Acknowledgment

The author thanks S. Schweighofer, G. Vaccario and F. Schweitzer for useful discussion, and V. Nanumyan for designing figs. 1 and 3.

# References

[1] Ahnert, S. E.; Garlaschelli, D.; Fink, T. M. A.; Caldarelli, G. (2007). Ensemble approach to the analysis of weighted networks. *Physical Review E* **76(1)**, 016101. ISSN 1539-3755.

[2] Casiraghi, G.; Nanumyan, V.; Scholtes, I.; Schweitzer, F. (2016). Generalized Hypergeometric Ensembles: Statistical Hypothesis Testing in Complex Networks. *arXiv preprint arXiv:1607.02441* .

[3] Cranmer, S. J.; Desmarais, B. A. (2011). Inferential network analysis with exponential random graph models. *Political Analysis* **19(1)**, 66–86.

[4] Dicks, L. V.; Corbet, S. a.; Pywell, R. F. (2002). Compartmentalization in plant-insect flower visitor webs. *Journal of Animal Ecology* **71(1)**, 32–43. ISSN 0021-8790.

[5] Eagle, N.; Pentland, A. (2006). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing* .

[6] Espín-Noboa, L.; Lemmerich, F.; Strohmaier, M.; Singer, P. (2017). A Hypotheses-driven Bayesian Approach for Understanding Edge Formation in Attributed Multigraphs. vol. 693. ISBN 978-3-319-50900-6, pp. 3–16.

[7] Garas, A. (ed.) (2016). *Interconnected Networks*. Understanding Complex Systems, Cham: Springer International Publishing. ISBN 978-3-319-23945-3.

[8] Garas, A.; Argyrakis, P.; Rozenblat, C.; Tomassini, M.; Havlin, S. (2010). Worldwide spreading of economic crisis. *New Journal of Physics* **12(11)**, 113043. ISSN 1367-2630.

[9] Hubert, L.; Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology* **29(2)**, 190–241. ISSN 00071102.

[10] Karrer, B.; Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83(1)**, 16107.

[11] Krackhardt, D. (1988). Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social Networks* **10(4)**, 359–381. ISSN 03788733.

[12] Mastrandrea, R.; Fournet, J.; Barrat, A. (2015). Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS ONE* **10(9)**, 1–26. ISSN 19326203.

[13] McPherson, M.; Smith-Lovin, L.; Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* **27(1)**, 415–444. ISSN 0360-0572.

[14] Memmott, J. (1999). The structure of a plant-pollination food web. *Ecology Letters* **2**, 276–280.

[15] Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103(23)**, 8577–8582.

[16] Newman, M. E. J. (2015). Generalized Communities in Networks. *Physical Review Letters* **115(8)**, 088701. ISSN 0031-9007.

[17] Onnela, J.-P.; Chakraborti, A.; Kaski, K.; Kertész, J.; Kanto, A. (2003). Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E* **68(5)**, 056110. ISSN 1063-651X.

[18] Peixoto, T. P. (2014). Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. *Physical Review X* **4(1)**, 011047. ISSN 2160-3308.

[19] Snijders, T.; Spreen, M.; Zwaagstra, R. (1995). The Use of Multilevel Modeling for Analysis of Personal Networks: Networks of Cocaine Users in an Urban Area. *Journal of Qualitative Anthropology* **5(2)**, 85–105.

[20] Snijders, T. A. (2011). Statistical Models for Social Networks. *Annual Review of Sociology* **37(1)**, 131–153. ISSN 0360-0572.

[21] Snijders, T. A. B. (1996). Stochastic actor-oriented models for network change. *The Journal of Mathematical Sociology* **21(1-2)**, 149–172.

[22] Snijders, T. A. B.; van de Bunt, G. G.; Steglich, C. E. G. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks* **32(1)**, 44–60. ISSN 03788733.

[23] Stehlé, J.; Voirin, N.; Barrat, A.; Cattuto, C.; Isella, L.; Pinton, J. F.; Quaggiotto, M.; van den Broeck, W.; Régis, C.; Lina, B.; Vanhems, P. (2011). High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE* **6(8)**. ISSN 19326203.

[24] Tomasello, M. V.; Napoletano, M.; Garas, A.; Schweitzer, F. (2016). The rise and fall of R&D networks. *Industrial and Corporate Change* , dtw041ISSN 0960-6491.

[25] Wallenius, K. T. (1963). *Biased Sampling: the Noncentral Hypergeometric Probability Distribution*. Ph.d. thesis, Stanford University.

[26] Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics* **9(1)**, 60–62. ISSN 0003-4851.