The likelihood-ratio test for multi-edge network models

Journal of Physics: Complexity

OPEN ACCESS

PAPER

The likelihood-ratio test for multi-edge network models

Giona Casiraghi*

ETH Zürich, Chair of Systems Design, Weinbergstrasse 56/58, Zürich, Switzerland

* Author to whom any correspondence should be addressed.

E-mail: gcasiraghi@ethz.ch

RECEIVED
22 February 2021

REVISED
7 May 2021

ACCEPTED FOR PUBLICATION
24 May 2021

PUBLISHED
24 June 2021

Keywords: likelihood-ratio test, multi-edge network, complex system, hypothesis testing, model selection

Abstract

The complexity underlying real-world systems implies that standard statistical hypothesis testing methods may not be adequate for these peculiar applications. Specifically, we show that the likelihood-ratio (LR) test's null-distribution needs to be modified to accommodate the complexity found in multi-edge network data. When working with independent observations, the p-values of LR tests are approximated using a χ^2 distribution. However, such an approximation should not be used when dealing with multi-edge network data. This type of data is characterized by multiple correlations and competitions that make the standard approximation unsuitable. We provide a solution to the problem by providing a better approximation of the LR test null-distribution through a beta distribution. Finally, we empirically show that even for a small multi-edge network, the standard χ^2 approximation provides erroneous results, while the proposed beta approximation yields the correct p-value estimation.

1. Overview

Complex systems are notoriously challenging to analyze due to the large number of interdependencies, competitions, and correlations underlying their dynamics. To deal with these issues, data-driven studies of complex systems are based—either directly or indirectly—on the careful formulation of models representing different hypotheses about the system. The validation of these hypotheses is performed by comparing how well different models fit some observed data $\mathcal{G}$. Principled *model selection* is most probably the central problem of data analysis.

Model selection and statistical hypothesis testing have been intensively investigated in the general cases of, e.g., linear and generalized statistical regression models [6, 22]. However, less attention has been devoted to developing hypothesis testing methods specific to network models and network data, commonly used to study complex systems. In this article, we investigate how one standard statistical test, the *likelihood-ratio* (*LR*) *test*, needs to be modified when dealing with *multi-edge network data*.

Model selection is addressed by operationalizing the principle of parsimony, one of the fundamental concepts in statistical modeling. Statisticians usually view the principle of parsimony as a *bias* versus *variance tradeoff*: bias decreases, and variance increases as the complexity of a model increases. The fit of any model can be improved by increasing the number of parameters. However, a tradeoff with the increasing variance must be considered when selecting a model to validate a statistical hypothesis. Parsimonious models should achieve the proper tradeoff between bias and variance. Box *et al* [3] suggested that the principle of parsimony should lead to a model with 'the smallest possible number of parameters for adequate representation of the data'. Data-driven selection of a parsimonious model is thus at the core of scientific research.

We can roughly summarise model selection methods into two groups that address the principle of parsimony differently. Model selection based on statistical tests and model selection based on information-theoretic methods [6]. Prominent examples of information-theoretic methods are the AIC [1, 2], the BIC [28], or description length minimisation [24, 27]. Studying how such methods fare when faced with network data complexity is beyond this article's scope.

The LR test is instead one of the most common examples of statistical tests used for hypothesis testing. Statistical tests allow performing hypothesis testing in the following way. They evaluate how far a **test statistic**

λ falls from an appropriately constructed null-model. In the LR test, the test statistic λ is the ratio between the likelihood L_0 of the model X_0 representing the null-hypothesis and the likelihood L_a of a more complex model X_a representing the alternative hypothesis to be tested. The better the alternative hypothesis is compared to the null, the smaller the test statistic's absolute value λ.

How small need λ be to reject the null hypothesis in favor of the alternative? This depends on the null-distribution of λ. In other words, it depends on the null-hypothesis and its corresponding null-model X_0. Assuming that the null-hypothesis is correct, we could generate realizations $\tilde{\mathcal{G}}$ of model X_0 that represent all the possible forms the null-hypothesis could have taken in the data. This provides a *null-distribution* for the test statistic, i.e., a baseline distribution of the test statistic assuming that the null hypothesis was true. If the alternative hypothesis does not fit the observed data well, we can expect the probability of observing λ from the null-distribution to be relatively large. The reason for this is that X_a does not fit the data better than X_0. In other words, there is no (statistical) evidence that we need the more complex model X_a to explain the data, and the null-model X_0 is sufficient. If the alternative hypothesis is considerably better than the null, the probability of observing λ under the null-hypothesis will be small. This would give statistical evidence to reject the null hypothesis in favor of the alternative. The p-value of the LR test is precisely the probability of observing a value from the null distribution as small or smaller than λ.

Standard implementations of the LR test have been developed assuming that the data consist of many independent observations of the same process (i.i.d. observations) [22]. Under these circumstances, Wilk's theorem provides a widely used approximation of the null-distribution of λ to a χ^2 distribution [25]. Analyzing complex systems, we are often faced with *multi-edge network data*. These data consist of m repeated—and possibly time-stamped—edges (i, j) representing interactions between n different agents i, j, the vertices of the network. Examples of such datasets arise in multiple fields, e.g., in the form of human or animal interactions. Usually, such repeated edges are explicitly dependent on each other. More specifically, repeated edges are the events defining the network that are observed and recorded. The crucial assumption required by Wilk's theorem is that such events are independent from each other (and identically distributed, i.i.d.), i.e., that the presence or absence of some edges does not affect the presence or absence of other edges. The underlying assumption in network science, instead, is that the opposite of i.i.d. is true: that the presence or absence of edges between some vertex pairs do affect edges between other vertex pairs. This is critical in the presence of phenomena typically found in complex social systems, such as triadic closure [26], structural balance [18], degree–degree correlations [23], and other network effects.

So what is the implication of such interdependencies? The dependence between different observations of a complex system means that some of the statistical tests' properties will not hold when analyzing network data. In particular, we show that the null-distribution of the test statistic λ of the LR test needs to be modified to accommodate such dependence. When this is not done, the results obtained applying LR tests for hypothesis testing cannot be relied upon.

To illustrate this, we employ the generalized hypergeometric ensemble of random graphs (gHypEGs) [9, 12] to model multi-edge network data. The gHypEG allows the encoding of different types of hypotheses in a model, from simple ones like block structures [8] to more complex ones, akin to statistical regression models [5, 7]. Such models can then be used to evaluate different hypotheses about the data [4].

2. Statistical hypotheses and gHypEG

2.1. Multi-edge network data and hypothesis formulation

In this article, we deal with multi-edge network data. Each observation $e_{i\to j}$ consists of an interaction from a system's agent i to another agent j. All observations $E = \{e_{i\to j}\}$ can be collected as directed edges in a multi-edge network $\mathcal{G}(V, E)$, where V is the set of all interacting agents, the vertices of the network. The matrix $\mathbf{A}$ denotes the adjacency matrix of the network $\mathcal{G}$. Each of its entries A_{ij} reports the number of edges from vertex i to vertex j, or in other words, the number of observed interactions between agent i and agent j.

Within this framework, statistical hypotheses are formulated in terms of *random graph models*. Simple examples of such hypotheses are:

(a) Each agent has the same potential to interact as any other.

(b) Different agents have different potentials to interact.

(c) Agents are separated into two distinct groups, and agents of one group are more likely to interact with each other than with agents of the other group.

(d) Agents are separated into n distinct groups, and agents of one group are more likely to interact with each other than with agents of the other groups.

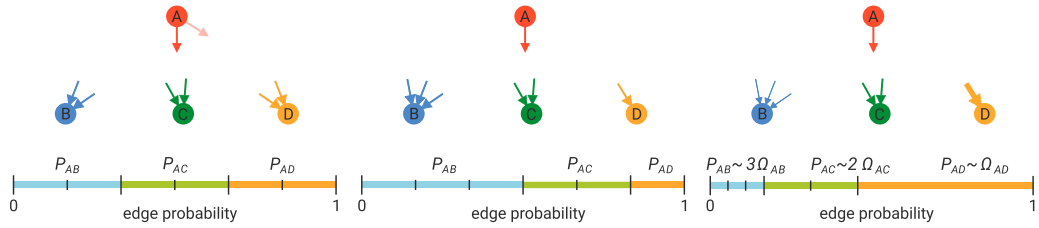

Figure 1. Probabilities of connecting different agents according to gHypEG, from the perspective of an agent *A*. Graphical illustration of (left) uniform edge probabilities, (center) the probability of connecting two vertices is a function of degrees, and (right) the probability of connecting two vertices is function of degrees and propensities Ω (represented as the edge-weight).

More complex social and dynamical hypotheses can obviously be formulated depending on the system studied. Testing these hypotheses requires encoding them in statistical models and then comparing the fit of these models against each other or against some null-models [22]. We employ the gHypEG as the statistical model encoding such hypotheses to deal with multi-edge network data. While this is not the only option, we choose the gHypEG because of its versatility and suitability to model multi-edges. In the next sections, we show (i) how to formulate a LR test between two of such hypotheses and (ii) how the null-distribution of the LR test needs to be modified to fit complex network data.

2.2. The generalised hypergeometric ensembles of random graphs (gHypEG)

Before discussing the LR test details, we briefly introduce the gHypEGs. A more formal presentation is provided in [9].

The general idea underlying the gHypEG is to sample edges at random from a predefined set of possible edges. Hypotheses about the system from which the data is observed can be encoded by either (a) changing the number of possible edges in such a set and (b) changing the odds of sampling an edge between two vertices instead of others, i.e., by specifying different edge sampling weights, or biases. Figure 1 provides a graphical illustration of such a process from the perspective of an agent *A*.

On the left-hand side of figure 1, the number of possible ways *A* can interact with the other agents is the same: there are two *edge-stubs* for each vertex. Moreover, the odds of sampling one edge-stub instead of another is 1. Each edge has thus the same sampling weight, which is denoted as Ω_{ij} in the following, where i, j are vertices in V. According to the model just described, the probability of observing a multi-edge network $\mathcal{G}$ with m edges depends only on the number Ξ of possible edges between each pair of vertices. This scenario gives rise to a uniform random graph model similar in spirit to the $G(n, p)$ model of Erdös and Rènyi [16]. The process of sampling m edges from a collection of $n^2\Xi$ possible edges, i.e. Ξ possible edges for each pair of vertices in a directed network with selfloops, is described by the hypergeometric distribution [9]:

$$\Pr(\mathcal{G}|\Xi) = \binom{n^2\Xi}{m}^{-1} \prod_{i,j\in V} \binom{\Xi}{A_{ij}}. \tag{1}$$

The reason for is that the process described above corresponds to a standard urn problem. The probability of sampling m objects without replacement from an urn containing $n^2\Xi$ total objects of n^2 different types is given by the multivariate hypergeometric distribution. In equation (1), the object sampled are the edges, and their 'types' correspond to the pair of vertices i, j to which they are incident. The probability of sampling any single object is the same for all of them, and the urn contains the same number Ξ of objects of each type.

By setting $\Xi = (m/n)^2$, we ensure that the *average degree* of the observed network $\mathcal{G}$ is preserved by the model. This first scenario corresponds to the hypothesis (a) listed above: each agent has the same potential to interact. Furthermore, in a directed network with self-loops, $\Xi = m^2/n^2$ corresponds to the maximum likelihood estimation (MLE) of the only model parameter Ξ. We refer to the resulting hypergeometric network model as *regular model*.

The central illustration of figure 1 highlights a different case. The odds between the different interactions are still identical. Therefore, there is no preference for *A* to interact with any of the other agents. However, the actual possibilities of interactions vary between the different agents: each agent has a different number of edge-stubs for *A* to connect to. This scenario encodes a different potential of interaction for the different agents, usually reflected in a heterogeneous degree distribution found in the network $\mathcal{G}$. This model encodes hypothesis (b) above. In practice, this hypothesis requires setting different values Ξ_{ij} for the number of possible edges between each pair vertices i, j. The probability of observing a network $\mathcal{G}$ according to this model changes

as follows:

$$\Pr\left(\mathcal{G}|\boldsymbol{\Xi}\right) = \binom{\sum_{ij}\Xi_{ij}}{m}^{-1} \prod_{i,j\in V}\binom{\Xi_{ij}}{A_{ij}}, \tag{2}$$

where the matrix $\boldsymbol{\Xi}$ contains all different entries Ξ_{ij}. In the urn parallel, Ξ_{ij} denotes the number of objects of each type present in the urn. While the probability of sampling any single object is still the same for all objects, the probability of sampling an object of a given type increases with its abundance in the urn Ξ_{ij}.

The value of Ξ_{ij} can be freely chosen to encode different properties of the system studied (see, e.g., [4]). For example, if we were studying a citation network consisting of citations between scientists, we could set Ξ_{ij} to $p_i \cdot p_j$, where p_x is the number of articles published by scientist x. $p_i \cdot p_j$ would then encode all the possible ways scientist i *could have* cited scientist j, through all their respective publications. In most cases, though, Ξ_{ij} is taken to be $k_i^{\text{out}} \cdot k_j^{\text{in}}$, where k_x^{in} is the observed in-degree of agent x, and k_x^{out} its observed out degree. This hypergeometric network model corresponds to a soft *configuration model* [17], and defines a network model that preserves the observed degree sequences in expectation [9].

The two models described so far are both characterized by the absence of sampling biases, i.e., interaction preferences between specific vertex pairs that go beyond what is prescribed by the number of edge-stubs and degrees. GHypEG further expands this formulation modifying the hypergeometric configuration model with additional information available about the system. Specifically, the probability of connecting two vertices depends not only on the observed degrees (i.e., number of stubs) but also on an independent *propensity* of two vertices to be connected. Such propensities introduce non-degree related effects into the model. This result is achieved by changing the *odds* of connecting a pair of vertices instead of another. The right side of figure 1 illustrates this case, where A is most likely to connect with vertex D, even though D has only one available stub.

We collect these edge propensities in a matrix $\boldsymbol{\Omega}$. The ratio between any two elements Ω_{ij} and Ω_{kl} of the propensity matrix gives the odds-ratio of observing an edge between vertices i and j instead of k and l, independently of the degrees of the vertices. The probability of a graph $\mathcal{G}$ depends on the stubs' configuration specified by $\boldsymbol{\Xi}$, and on the odds defined by $\boldsymbol{\Omega}$. Such a probability distribution is described by the multivariate Wallenius' non-central hypergeometric distribution [14, 31]:

$$\Pr(\mathcal{G}|\boldsymbol{\Xi},\boldsymbol{\Omega}) = \left[\prod_{i,j}\binom{\Xi_{ij}}{A_{ij}}\right]\int_0^1 \prod_{i,j}\left(1 - z^{\frac{\Omega_{ij}}{S_\Omega}}\right)^{A_{ij}} \mathrm{d}z \tag{3}$$

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

This model can again be seen as an urn problem. Differently from the cases above, though, the probability of sampling a single object from the urn is not anymore the same for all object types. Specifically, some objects are more likely to be sampled even when correcting for their abundance Ξ_{ij}. Such a *non-central* multivariate probability distribution is the Wallenius distribution provided in equation (3).

2.3. Undirected models

Equations (1) to (3) define models for directed multi-edge graphs. However, they can be easily adapted to the undirected case. The undirected version of the model is the degenerate case of its directed counterpart, where the direction of the edges is ignored [9]. In particular, this implies that the random vector corresponding to the undirected model has half the dimensions of the directed one, because any undirected edge between two vertices i and j can either be generated as a directed edge (i, j) or as a directed edge (j, i). A naive approach would simply restrict the directed model to the $n(n+1)/2$ components corresponding to the upper-triangle and diagonal of the adjacency matrix. However, this approach ignores (a) the symmetry in the process of constructing possible edges, since (i, j) and (j, i) denote the same edge, and (b) the fact that eventual selfloops needs to be accounted differently than normal edges as they lack this symmetry.

To address these issues, the corresponding undirected version of the general undirected model with parameter $\boldsymbol{\Xi}$ and $\boldsymbol{\Omega}$ of equation (3) reads as

$$\Pr(\mathcal{G}|\boldsymbol{\Xi},\boldsymbol{\Omega}) = \left[\prod_{i<j}\binom{2\Xi_{ij}}{A_{ij}}\prod_{l\in V}\binom{\Xi_{ll}}{A_{ll}/2}\right]\int_0^1 \prod_{i<j}\left(1 - z^{\frac{\Omega_{ij}}{S_\Omega}}\right)^{A_{ij}}\prod_{l\in V}\left(1 - z^{\frac{\Omega_{ll}}{S_\Omega}}\right)^{A_{ll}/2} \mathrm{d}z. \tag{4}$$

The corresponding undirected versions of equations (1) and (2) are obtained following the same reasoning. A formal proof for this correspondence between directed and undirected models is provided in [9].

2.4. Encoding hypotheses

By constraining the number of free parameters in $\mathbf{\Omega}$, we can specify hypotheses about the data changing the sampling odds for different vertex pairs. For example, we can cluster vertices into multiple groups and verify whether the odds of observing interactions *within a group* and *between a group* are different [8]. The resulting model is similar to a degree-corrected stochastic block model [19]. Alternatively, we can specify $\mathbf{\Omega}$ to encode endogenous network properties. E.g., $\mathbf{\Omega}$ can be utilized to encode *triadic closure* [5], to verify whether pairs whose interactions will close triads in the network are more likely than others. Finally, different effects contributing to the odds of observing some interactions instead of others can be composed together to formulate more complex hypotheses [7].

The advantage of the approach just described is the ability to encode a wide range of statistical hypotheses within the same modelling framework. This has the practical benefit of allowing the comparison of very different models, as they can all be formulated by means of the same probability distribution of equation (3). Different hypotheses are thus encoded by appropriately choosing the free parameters in $\mathbf{\Xi}$ and $\mathbf{\Omega}$.

For clarity, we will focus on simple hypotheses such as those described in the previous section. However, the results shown do hold for any combinations of $\mathbf{\Xi}$ and $\mathbf{\Omega}$. In the particular case of encoding group structures, $\mathbf{\Omega}$ takes the following form:

$$\Omega_{ij} := \omega_{g_i,g_j}, \tag{5}$$

where g_i is the group of agent i, g_j is the group of agent j, and ω_{g_i,g_j} is the propensity of sampling an edge between group g_i and group g_j. In the presence of two different groups of vertices (A, B), there are three possible values that Ω_{ij} can take: $\omega_{AA}, \omega_{BB}, \omega_{AB}$ (assuming that $\omega_{AB} = \omega_{BA}$). The ratio ω_{AA}/ω_{AB} gives the odds between sampling an edge within group A and an edge between group A and B given a value of Ξ.

3. The likelihood-ratio (LR) test

We now illustrate how the LR test is used to test a null-hypothesis against an alternative hypothesis about the observed system. The data are used to define a graph $\mathcal{G}$ with adjacency matrix $\boldsymbol{A}$. Let H_r be some statistical hypothesis. Here, we always assume that each hypothesis is defined by a gHypEG model X_r that can be encoded by a propensity matrix $\mathbf{\Omega}_r$ and a combinatorial matrix $\mathbf{\Xi}_r$. Each model is characterized by several free parameters that we want to fit to the data $\mathcal{G}$, such that the probability of observing $\mathcal{G}$ is maximized. This requirement corresponds to performing a MLE of the free parameters.

Likelihood-ratio statistic. Assume now we have two hypotheses we want to test against each other. Let H_0 denote the null-hypothesis and let H_a denote the alternative. The corresponding models are defined in terms of $\mathbf{\Omega}_0$, $\mathbf{\Xi}_0$ and $\mathbf{\Omega}_\mathrm{a}$, $\mathbf{\Xi}_\mathrm{a}$. To test the alternative hypothesis against the null, we use the LR statistic $\lambda(0, a)$, defined as follows:

Definition 3.1. (LR statistic). Let $\mathcal{G}$ be a graph, X_0 be the model corresponding to the null-hypothesis, and X_a the model corresponding to the alternative hypothesis. The likelihood ratio statistic $\lambda(0, a)$ is given by

$$\lambda(0, a) := \frac{L(\mathbf{\Xi}_0, \mathbf{\Omega}_0|\boldsymbol{A})}{\sup(L(\mathbf{\Xi}_0, \mathbf{\Omega}_0|\boldsymbol{A}), L(\mathbf{\Xi}_\mathrm{a}, \mathbf{\Omega}_\mathrm{a}|\boldsymbol{A}))}, \tag{6}$$

where $L(\mathbf{\Xi}_r, \mathbf{\Omega}_r|\boldsymbol{A}) = \Pr(\mathcal{G}|\mathbf{\Xi}_r, \mathbf{\Omega}_r)$ denotes the likelihood of model X_r given the network $\mathcal{G}$.

Through the LR statistic, we can perform two types of tests. First, we can perform a standard model selection test to compare a simpler model against a more complex model. This test corresponds to verify whether there is enough evidence in the data that justifies the more complex model or whether the simpler model fits the data well enough. In this scenario, the simpler model corresponds to the null-hypothesis, while the more complex model to the alternative.

Second, the LR test can be used to perform a goodness-of-fit test. This test allows verifying the quality of the fit of a model X_r. By defining the alternative hypothesis with a model X_full that perfectly reproduces the observed data (in expectation), we can test whether the fit of the model X_r is as good as such an overfitting model [22]. In the framework of gHypEGs, the alternative hypothesis is obtained by specifying the parameter matrix $\mathbf{\Omega}_\mathrm{full}$ such that the expectation of X_full corresponds to the observation $\mathcal{G}$. This model is the maximally complex model that can be specified with a gHypEG and has as many free parameters as entries in the adjacency matrix [9].

Let us now assume that the two models corresponding to the alternative and null hypotheses are *nested*. This means that $\mathbf{\Xi}_0$ can be written as a special case of $\mathbf{\Xi}_\mathrm{a}$, and $\mathbf{\Omega}_0$ as a special case of $\mathbf{\Omega}_\mathrm{a}$. Thus, the null-model (with fewer parameters) can be formulated by constraining some of the alternative model parameters. Thanks to Wilks' theorem [25], if the two models are nested, the number of observations m is large, and the

observations are independent, the distribution of λ under the null-hypothesis can be written in terms of

$$D(0,a) := -2\log(\lambda(0,a)), \tag{7}$$

and can be approximated by the χ^2 distribution with as many degrees of freedom as the difference of degrees of freedoms between the two models. Letting ν be the difference of degrees of freedom between the null and the alternative models, the p-value of the LR test between the two hypotheses is computed as follows:

$$p-\text{value} := \Pr(\chi^2(\nu) \geqslant D(0,a)). \tag{8}$$

We reject the null hypothesis in favour of the alternative if the p-value is smaller than some threshold α.

Distribution of λ under the null-hypothesis. The question that remains to be answered is whether the conditions provided by multi-edge network data allow Wilks' theorem's application. Unfortunately, in most real-world scenarios, the answer to this question is negative. This is a known issue in statistics, where it arises in the context of multinomial goodness-of-fit tests [13, 20, 21, 30]. Because Wallenius' multivariate non-central hypergeometric distribution converges to the multinomial distribution (a formal proof can be found in [33]), we use the results obtained for multinomial tests to find a better approximation for the null distribution of $D(0,a)$ than the χ^2 approximation of Wilks' theorem. Specifically, following the work of Smith et al [30], we propose approximating the distribution of $D(0,a)$ with a beta distribution.

Theorem 1. (Approximation of LR statistics distribution). *The distribution under the null-hypothesis of $D(0,a)$, defined as in equation (7), for m large can be approximated by a Beta(α,β) distribution with parameters*

$$\alpha = \frac{\mu\left[D(0,a)\right]}{M\cdot\sigma^2\left[D(0,a)\right]}\cdot\left(\mu\left[D(0,a)\right]\cdot\left(M-\mu\left[D(0,a)\right]\right)-\sigma^2\left[D(0,a)\right]\right), \tag{9}$$

and

$$\beta = \left(M-\mu\left[D(0,a)\right]\right)\cdot\frac{\alpha}{\mu\left[D(0,a)\right]}, \tag{10}$$

where M denotes the upper limit of the image of $D(0,a)$, $\mu\left[D(0,a)\right]$ its expectation and $\sigma^2\left[D(0,a)\right]$ its variance.

We now sketch a proof for theorem 1. If a random variable X is distributed on $[0,M]$, then $\Pr(X/M \leqslant x)$ can be approximated by a beta distribution with some parameters α and β [30]. Using the methods of moments, we express the two distribution parameters α and β in terms of the first two moments of the distribution $\Pr(X/M \leqslant x)$ that we want to approximate. Writing α and β as a function of the mean $\mu[X]$ and variance $\sigma^2[X]$ of X we obtain equations (9) and (10), where X has been substituted by $D(0,a)$ [15]. Finally, for a given m and a fixed number of vertices, there is only a finite number of possible graphs that exists. Thus, the image of $D(0,a)$ under the null-hypothesis is discrete, as $\lambda(0,a)$ can take only a finite number of values. The larger the value of m, the more graphs exists, and the better is the beta approximation of the discrete distribution of $D(0,a)$ under the null-hypothesis.

In some special cases there exist analytical solutions for $\mu\left[D(0,a)\right]$ and $\sigma^2\left[D(0,a)\right]$. For example, Smith *et al* [30] provided analytical expressions for the mean and variance of the null distribution of $D(0,a)$ in the case of a multinomial goodness-of-fit test, where all events in the multinomial are equiprobable. These expressions could be used here as an approximation of LR-tests where the null model is the regular model given by equation (1). Because the multivariate hypergeometric distribution with a single parameter Ξ can be approximated by an equiprobable multinomial distribution [33], choosing a regular model as null-hypothesis matches the equiprobable null model studied in Smith *et al* [30]. In most situations, though, an analytical expression for the first two moments of $D(0,a)$ is not available, and we resort to a numerical estimation of them. While a general solution would be optimal, thanks to the ability to generate samples provided by gHypEG models, the parameters' numerical estimation can be nevertheless performed with ease. Specifically, we can estimate $\mu\left[D(0,a)\right]$ and $\sigma^2\left[D(0,a)\right]$ through the sample mean and sample variances of the empirical null distribution. To generate the complete null distribution of the LR statistic numerically, we would need a considerable number of realisations. In the case of large networks, this is infeasible. Exploiting theorem 1 instead, we only need to estimate the first two moments of the distribution under the null hypothesis, which can be done reliably with a small number of realisations generated from the model corresponding to the null-hypothesis [29]. Furthermore, for the general model of equation (3), following Smith *et al* [30], we can estimate the upper limit M of the range of $D(0,a)$ as

$$M = -2m\log\left(\min_{i,j}\left[\Omega_{ij}\Xi_{ij}\right]/\sum_{i,j}\Omega_{ij}\Xi_{ij}\right). \tag{11}$$

In the equation above, $\min_{i,j}\left[\Omega_{ij}\Xi_{ij}\right]/\sum_{i,j}\Omega_{ij}\Xi_{ij}$ denotes the smallest edge probability $p_{\min}$ defined by the null model. Thus, $-2\log\left(p_{\min}^{m}/1^{m}\right)$ is the largest possible value that $D(0,a)$ can take assuming a multinomial approximation of the (generalised) hypergeometric distribution, which is acceptable for m large [33].

The `R` package `ghypernet`[1] [10] provides an implementation of the LR test for gHypEG models. The package is open source and can be obtained from the CRAN `R` packages repository.

4. Simulation studies

In a first simulation study, we highlight how theorem 1 allows the estimate of the null-distribution of the LR test efficiently. Specifically, we show that the asymptotic results of theorem 1 are acceptable even in the case of a small network with a limited number of edges, and that the moments of the beta distribution can be estimated from a small number of synthetic realisations s.

In the absence of analytical expressions for the null-distribution, we are required to estimate it numerically to be able to compute p-values. The direct approach to do so consists in generating all graphs possible under the null-hypothesis. In practice, we generate a large number s_{null} of realisations from the model defining the null-hypothesis. For each of these realisations, we compute the LR statistic λ and $D(0,a)$ (cf equation (7)). The resulting empirical distribution of $D(0,a)$ is a numerical approximation of the null-distribution of the test [22]. The larger the number of realisations s_{null} used to generate such an empirical null-distribution, the more accurate it will reproduce the 'true' unknown distribution.

Thanks to theorem 1, instead, we do not need to generate the null-distribution. It is sufficient to estimate its first two moments to compute its parameters α and β. Computationally, this is much more efficient [29], as the number of synthetic realisations under the null-hypothesis that are needed to obtain a good estimate of the first two moments is smaller. In the following, we investigate how well the approximating null-distribution defined by theorem 1 and estimated with the methods of moments matches the empirical null-distribution.

In figure 2(a), we show how the empirical and the approximating distribution become more and more similar when the number of edges m increases (for a fixed number of vertices). Specifically, we compute the Kolmogorov–Smirnov distance D_{KS} between the empirical distribution and the approximating beta distribution for a regular model with 40 vertices and an increasing number of indirected edges m. The lower is D_{KS}, the closer are the two distribution. According to the Kolmogorov–Smirnov test, of which D_{KS} is the test statistic, for small values of D_{KS} the two distributions cannot be statistically distinguished. We see that the value of D_{KS} sharply decreases for increasing values of m (note the log-scale for the y axis in the plot). In figure 2(b), instead, we show the results of applying Kolmogorov–Smirnov's test comparing the LR's empirical distribution against the beta distribution fitted to increasing sample sizes s_{Beta}. The example is constructed from a 40 vertices random graph and fixing $m = 500$. In both cases, the null model is chosen to be a regular model.

To build the empirical distribution of the LR statistic, we take $s_{\text{null}} = 500\,000$ realisations under the null hypothesis. We then proceed to compute the parameters of the approximating beta distribution using sample mean and variance computed from an increasing number of independent samples s_{Beta} (figure 2(b)), or using all realisations to ensure the best fit (figure 2(a)). The results show (a) that increasing m the fit of the approximating beta distribution improves, and (b) with a limited sample size $s_{\text{Beta}} \sim 1000 \ll s_{\text{null}}$, most of the observations give a p-value for the Kolmogorov–Smirnov test larger than 0.05. Hence, with a limited sample size s_{Beta}, the empirical null-distribution obtained from s_{null} is not significantly different from the beta distribution whose parameters have been estimated from the s_{Beta} realizations.

In the second simulation study, we generate a random undirected graph with $n = 100$ vertices and $m = 400$ undirected edges uniformly distributed between each vertex pairs. Utilizing the LR test, we can test the null-hypothesis (a) that each vertex has the same potential of interactions against the alternative hypothesis (b) that different vertices/agents have different interaction potentials. As explained in the previous sections, (a) is encoded by a *regular model* with one parameter, and (b) by a *configuration model*. This test corresponds to testing that the degree distribution deviates from that of the regular model. We expect that the test returns high p-values because we choose the null-hypothesis to match the random graph's generating process. The results, obtained from 1000 repetitions of the experiment, confirm this hypothesis with a p-value of the median λ of 0.44. Similarly, we perform the same experiment generating a random undirected graph from the standard configuration model with a heterogeneous degree distribution. To ease the comparison with the example above, we define it by a degree sequence sampled from a geometric distribution with mean chosen such that

[1] https://ghyper.net

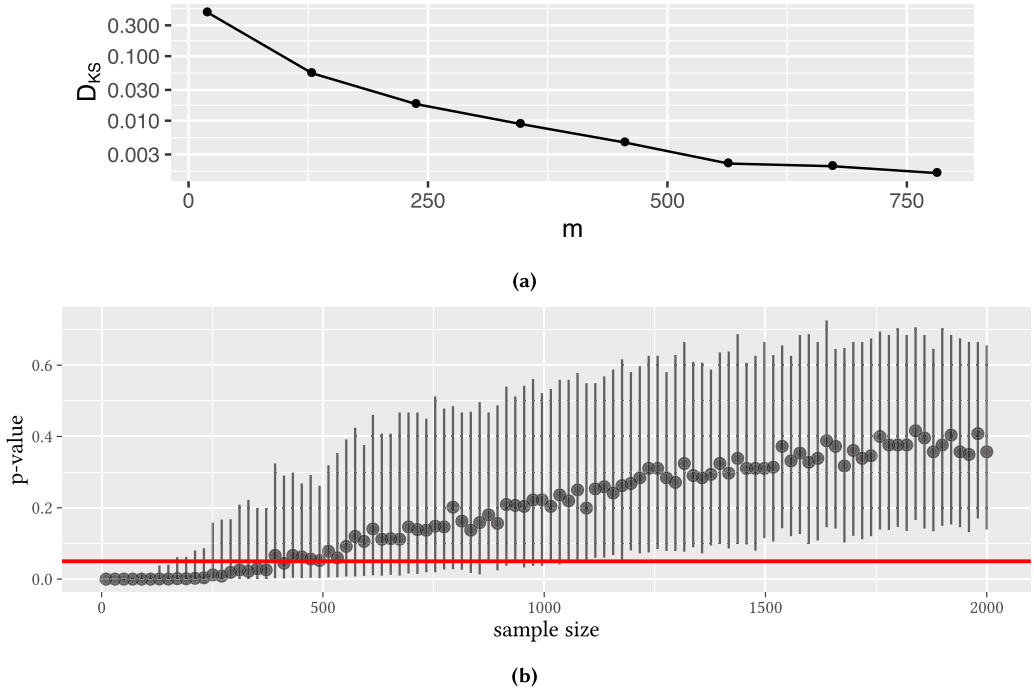

Figure 2. Quality of the beta approximation of the empirical LR null distribution. (a) Kolmogorov–Smirnov statistic D_{KS} for the comparison of the empirical and approximating distribution computed increasing the number m of edges in a regular model. A smaller value D_{KS} corresponds to a better fit of the approximating distribution to the empirical distribution. (b) P-values of a Kolmogorov–Smirnov test for the LR null-distribution. Each point is the p-value (y-axis) of the median statistic resulting from 500 tests of the empirical distribution against the approximating beta distribution, computed from an increasing number of realisations s (x-axis). The length of the lines denotes 1.5 inter-quantile ranges. The horizontal red line shows the 0.05 significance threshold for the Kolmogorov–Smirnov test. It is evident a clear trend where the p-value increases with increasing sample size. That highlights an increase of the goodness-of-fit of the limiting beta distribution for the empirical distribution.

the expected number of edges in the graph is $m = 400$. This effectively corresponds to generating data according to the hypothesis (b). In this case, we expect small p-values from the same test done before because the generating model of the data corresponds to the alternative hypothesis. Repeating the experiment 1000 times, for the largest recorded λ we obtain a p-value $< 1e-20$.

5. Case study

Finally, we provide a case study involving an empirical graph. We use Zachary's Karate club (ZKC) [32] as a test case. ZKC consists of 34 vertices and 231 undirected multi-edges. As with most empirical graphs, its degree sequence is skewed (empirical skewness is 1.456). Hence, we expect that the test we performed before—comparing the null-hypothesis of a regular model against the hypergeometric configuration model—should reject the null-hypothesis. Performing such a LR test gives a p-value $< 1e-20$, which confirms our expectations.

We can further exploit this example to compare the empirical distribution of $D(0, a)$ under the null hypothesis with the χ^2 distribution and the beta distribution. The result is shown in figure 3(a), where the shaded area corresponds to the empirical complementary cumulative distribution (CCDF) under the null-hypothesis, computed from $s_{\text{null}} = 500\,000$ synthetic realisations, and the two curves show the CCDFs of the χ^2 and beta distributions respectively. Note that the value of $D(0, a)$ for ZKC is 300.338, which is out of scale on the right side of the x-axis of figure 3(a). In this case, it appears that the beta distribution and the χ^2 distribution provide similar fits to the empirical null-distribution. However, when we perform a two-sided Kolmogorov–Smirnov test, we get a p-value of $1.45e-05$ for the χ^2 distribution, which means that we can reject the null hypothesis that the empirical distribution follows a χ^2. Performing the same test for the beta distribution, we get a p-value of 0.4211. That means that we cannot reject the null-hypothesis that the empirical distribution follows a beta. While this is not enough to claim that the distribution of $D(0, a)$ is a beta, it gives confidence on using the asymptotic results of theorem 1 even for small graphs like ZKC.

We perform a second experiment to highlight how much, in extreme cases, the χ^2 distribution can deviate from the empirical distribution of $D(0, a)$ under the null-hypothesis, thus heavily biasing p-value estimates. For the ZKC, we now perform a goodness-of-fit test of the hypergeometric configuration model. The alternative

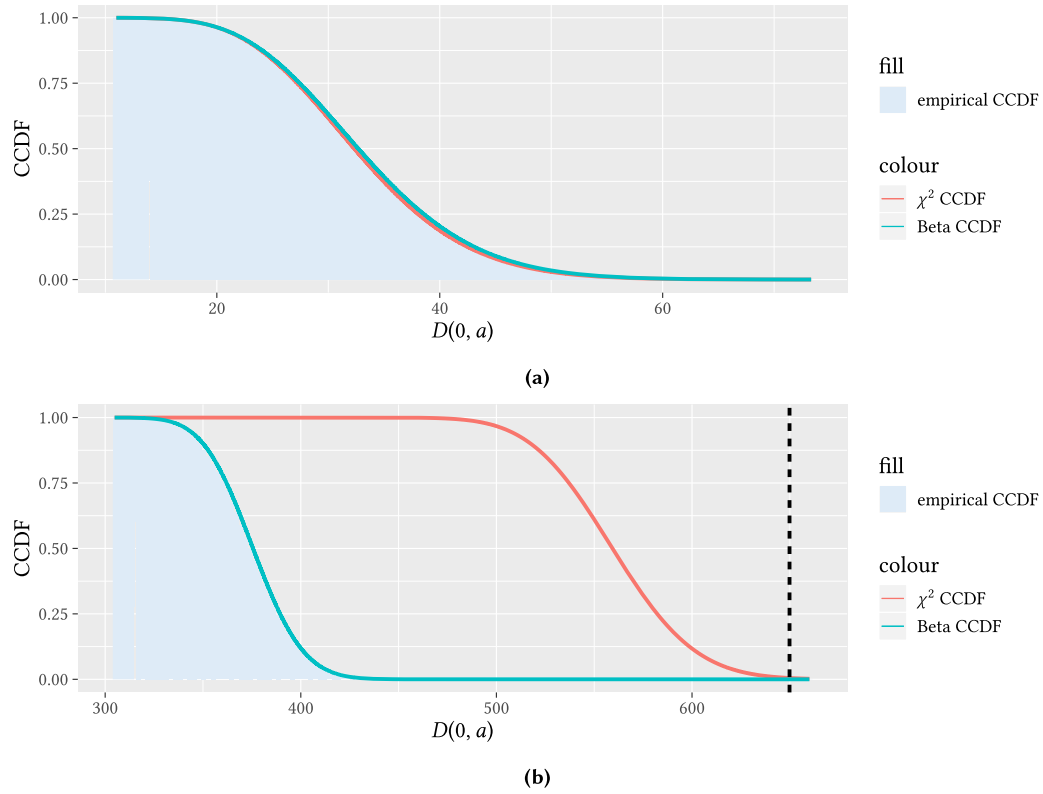

Figure 3. Empirical distribution of $D(0, 2)$ for two LR tests perform on ZKC. In the top figure, we perform a LR test for the null-hypothesis that ZKC comes from a regular model against the alternative hypothesis that ZKC comes from a configuration model. The results show that there is strong evidence to reject the null hypothesis. In the bottom figure, we perform a goodness-of-fit test for the hypergeometric configuration model on ZKC. Also, in this case, the results show a bad fit of the null-model. While in the top-figure, the χ^2 and the beta approximations of the LR statistic's empirical distribution give a relatively good fit, in the bottom case, it is clear that the χ^2 does not approximate well the empirical distribution. The shaded area denotes the empirical CCDF of $D(0, 2)$ under the null-hypothesis, the orange line its χ^2 approximation, and the green line its beta approximation. The vertical line denotes the value of the LR statistic for ZKC. In the top figure, such a line is out of the boundaries on the right side and hence not plotted.

hypothesis is encoded by the maximally complex model fitted by gHypEGs and results in a model that fixes the expected graph as the observed one, as explained above. The fit of its parameters is performed according to what described in [9]. The test can be interpreted as how well the null-model fits the data, which is entirely encoded in the full model. This test results in a p-value of $1.69e-30$. That means that the configuration model is not a good model for the ZKC. This result is hardly surprising, given the well-known community structure present in the empirical graph. In figure 3(b), we show the empirical distribution of $D(0, a)$. There, we notice that while the beta distribution provides a visually good fit, the χ^2 distribution is heavily shifted to the right. The two-sample Kolmogorov–Smirnov test confirms this result, providing a p-value of 0.169 for the beta distribution, and $< 2.2e-16$ for the χ^2 distribution.

In this last example, it is essential to note that using the χ^2 distribution would provide a misleading result. Comparing the value of $D(0, a)$ for the empirical graph, we see that it is on the right tail of the χ^2 distribution. Computing a p-value from this distribution would result in a p-value of ≈ 0.005. That means that in this case, we would only weakly reject the null-hypothesis, giving the wrong impression that the ZKC could come from an extreme realization of a simple configuration model. However, this is ruled out by looking at the LR statistics' empirical distribution or simply comparing an empirical graph with a realization from the hypergeometric configuration model [11]. The beta approximation provided by theorem 1 solves this issue, providing a better estimate of the null-distribution of the LR test and thus a reliable p-value.

6. Discussion

The study of complex systems is intertwined with network science and advanced multivariate statistics. Hypothesis testing and model selection methods, in particular, need to account for the complexity underlying observations from such systems. Because interactions between system agents tend not to be independent, many standard statistical methods should be employed with care when dealing with network data.

This article has investigated how the LR test needs to be adapted to deal with network models. LR tests provide a practical methodology for selecting different network models and testing statistical hypotheses. However, the characteristics of multi-edge networks require us to adapt the test null-distribution to account for the underlying complexity of network data. When this is not done, we incur the risk of over- (or under-) estimating the p-values of the statistical test, generating contradictory results, as shown in the case study above. With theorem 1, we provide the means to correctly estimate the p-values for LR tests by means of a beta distribution. Finally, we provide an implementation of the methods described through the open source `R` package `ghypernet`. Even though our analysis is focused on the LR test, similar issues may arise with other statistical tools applied to complex networks.

The main limitation of the results presented in this article is the need to numerically estimate the first two empirical moments of the statistic's null distribution. Although this can be performed easily using our implementation, we will investigate analytical asymptotic estimates for the parameters needed in future research.

Acknowledgments

The author thanks Frank Schweitzer for his support and Georges Andres for his detailed comments.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Giona Casiraghi https://orcid.org/0000-0003-0233-5747

References

[1] Akaike H 1973 Information theory and an extension of the maximum likelihood principle *Int. Symp. on Information Theory* ed B Petrov and F Csaki pp 267–81
[2] Akaike H 1974 A new look at the statistical model identification *IEEE Trans. Autom. Control* **19** 716–23
[3] Box G E P, Jenkins G M and Reinsel G C 1994 *Time Series Analysis: Forecasting and Control* (*Wiley Series in Probability and Statistics*) (New York: Wiley)
[4] Brandenberger L, Casiraghi G, Andres G, Schweighofer S and Schweitzer F 2021 Why online does not equal offline: comparing online and real-world political support among politicians https://doi.org/10.31235/osf.io/j4fp6
[5] Brandenberger L, Casiraghi G, Nanumyan V and Schweitzer F 2019 Quantifying triadic closure in multi-edge social networks *Proc. of the 2019 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining* (New York: ACM) pp 307–10
[6] Burnham K P and Anderson D R (ed) *Model Selection and Multimodel Inference* 2004 (New York: Springer)
[7] Casiraghi G 2017 Multiplex network regression: how do relations drive interactions? (arXiv:1702.02048)
[8] Casiraghi G 2019 The block-constrained configuration model *Appl. Netw. Sci.* **4** 123
[9] Casiraghi G and Nanumyan V 2018 Generalised hypergeometric ensembles of random graphs: the configuration model as an urn problem (arXiv:1810.06495)
[10] Casiraghi G and Nanumyan V 2020 GHYPERNET v1.0.1: fit and simulate generalised hypergeometric ensembles of graphs
[11] Casiraghi G, Nanumyan V, Scholtes I and Schweitzer F 2016 Generalized hypergeometric ensembles: statistical hypothesis testing in complex networks (arXiv:1607.02441)
[12] Casiraghi G, Nanumyan V, Scholtes I and Schweitzer F 2017 From relational data to graphs: inferring significant links using generalized hypergeometric ensembles *Social Informatics. SocInfo 2017* (*Lecture Notes in Computer Science*) ed G L Ciampaglia, A Mashhadi and T Yasseri (Berlin: Springer) pp 111–20
[13] Chapman J-A W 1976 A comparison of the Chi squared, −2 log R, and multinomial probability criteria for significance tests when expected frequencies are small *J. Am. Stat. Assoc.* **71** 854–63
[14] Chesson J 1978 Measuring preference in selective predation *Ecology* **59** 211–5
[15] Elderton W P 1906 *Frequency-Curves and Correlation* (London: Institute of Actuaries)
[16] Erdős P and Rényi A 1959 On random graphs I *Publicationes Mathematicae Debrecen* **6** 290–7
[17] Fosdick B K, Larremore D B, Nishimura J and Ugander J 2018 Configuring random graph models with fixed degree sequences *SIAM Rev.* **60** 315–55
[18] Heider F 1946 Attitudes and cognitive organization *J. Psychol.* **21** 107–12
[19] Karrer B and Newman M E J 2011 Stochastic blockmodels and community structure in networks *Phys. Rev.* E **83** 16107
[20] Koehler K J and Larntz K 1980 An empirical investigation of goodness-of-fit statistics for sparse multinomials *J. Am. Stat. Assoc.* **75** 336–44
[21] Larntz K 1978 Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics *J. Am. Stat. Assoc.* **73** 253–63
[22] Lehmann E L and Romano J P (ed) 2005 *Testing Statistical Hypotheses* (*Springer Texts in Statistics*) (New York: Springer)
[23] Mondragón R J 2020 Estimating degree–degree correlation and network cores from the connectivity of high-degree nodes in complex networks *Sci. Rep.* **10** 5668
[24] Peixoto T P 2013 Parsimonious module inference in large networks *Phys. Rev. Lett.* **110** 148701

[25] Rao C R 1973 *Linear Statistical Inference and its Applications* vol 2 (New York: Wiley)
[26] Rivera M T, Soderstrom S B and Uzzi B 2010 Dynamics of dyads in social networks: assortative, relational, and proximity mechanisms *Annu. Rev. Sociol.* **36** 91–115
[27] Rosvall M and Bergstrom C T 2008 Maps of random walks on complex networks reveal community structure *Proc. Natl Acad. Sci.* **105** 1118–23
[28] Schwarz G 1978 Estimating the dimension of a model *Ann. Stat.* **6** 461–4
[29] Shore H 1995 Fitting a distribution by the first two moments (partial and complete) *Comput. Stat. Data Anal.* **19** 563–77
[30] Smith P J, Rae D S, Manderscheid R W and Silbergeld S 1981 Approximating the moments and distribution of the likelihood ratio statistic for multinomial goodness of fit *J. Am. Stat. Assoc.* **76** 737–40
[31] Wallenius K T 1963 Biased sampling: the noncentral hypergeometric probability distribution *PhD Thesis* Stanford University
[32] Zachary W W 1977 An information flow model for conflict and fission in small groups *J. Anthropol. Res.* **33** 452–73
[33] Zingg C, Casiraghi G, Vaccario G and Schweitzer F 2019 What is the entropy of a social organization? *Entropy* **21** 901