Model Selection for Degree-corrected Block Models

Xiaoran Yan
Computer Science
University of New Mexico
everyan@cs.unm.edu

Jacob E. Jensen
Computer Science
Columbia University
2timesjay@gmail.com

Florent Krzakala
ESPCI ParisTech
and CNRS
fk@espci.fr

Cristopher Moore
Santa Fe Institute
and University of New Mexico
moore@santafe.edu

Cosma Rohilla Shalizi
Statistics Department
Carnegie Mellon University
cshalizi@cmu.edu

Lenka Zdeborová
Institut de Physique Théorique, CEA Saclay
and CNRS
lenka.zdeborova@cea.fr

Pan Zhang
ESPCI ParisTech
and CNRS
pan.zhang@espci.fr

Yaojia Zhu
Computer Science
University of New Mexico
yaojia.zhu@gmail.com

Abstract

A central problem in analyzing networks is partitioning them into *modules* or *communities*, clusters with a statistically homogeneous pattern of links to each other or to the rest of the network. One of the best tools for this is the *stochastic block model*, which in its basic form imposes a Poisson degree distribution on all nodes within a community or block. In contrast, *degree-corrected* block models allow for heterogeneity of degree within blocks. Since these two model classes often lead to very different partitions of nodes into communities, we need an automatic way of deciding which model is more appropriate to a given graph. We present a principled and scalable algorithm for this model selection problem, and apply it to both synthetic and real-world networks. Specifically, we use belief propagation to efficiently approximate the log-likelihood of each class of models, summed over all community partitions, in the form of the Bethe free energy. We then derive asymptotic results on the mean and variance of the log-likelihood ratio we would observe if the null hypothesis were true, i.e. if the network were generated according to the non-degree-corrected block model. We find that for sparse networks, significant corrections to the classic asymptotic likelihood-ratio theory (underlying χ^2 hypothesis testing or the AIC) must be taken into account. We test our procedure against two real-world networks and find excellent agreement with our theory.

1 Introduction

In many real-world networks, nodes divide naturally into *communities*, clusters with dense internal ties which are only weakly connected to the rest of the graph. More generally, they can divide into *modules* or *functional communities*, where nodes in the same group connect to the rest of the network in similar ways. Discovering such communities is an important part of modeling networks [24], as community structure offers clues to the processes which generated the graph, on scales ranging from face-to-face social interaction [31] through social-media communications [1] to the organization of food webs [3, 18]. Since communities often reflect functional groupings in the underlying system, community membership is also useful for predicting the attributes of nodes, for predicting links between nodes, and for statistically controlling for unobserved node attributes.

The *stochastic block model* (SBM) [11, 15, 27, 2] has, deservedly, become one of the most popular generative models for community detection. It splits nodes into communities or *blocks*, within which all nodes are *stochastically equivalent* [28]. That is, the probability of an edge between any two nodes depends only on which blocks they belong to, and all edges are independent given the nodes' block memberships. Block models are highly flexible, representing assortative, disassortative and satellite community structures, as well as combinations thereof, in a single generative framework [21, 22]. Their asymptotic properties, including phase transitions in the detectability of communities, can be determined exactly using tools from statistical physics [10, 9].

Despite this flexibility, SBMs impose real restrictions on networks; notably, the degree distribution within each block is asymptotically Poisson. This makes the SBM implausible for many real-world networks, where the degrees within each community are highly inhomogeneous. Fitting the SBM to such networks tends to split the high- and low- degree nodes in the same community into distinct blocks; for instance, dividing both liberal and conservative political blogs into high-degree "leaders" and low-degree "followers" [1, 16]. To avoid this effect, and allow degree inhomogeneity within blocks, there is a long history of generative models where the probability of an edge depends on node attributes θ_u as well as their group memberships (e.g. [19, 25]). Here we use the variant due to [16], called the *degree-corrected* (DC) block model, where the expected number of edges between u and v is proportional to $\theta_u\theta_v$.

We often lack the domain knowledge to choose between the ordinary and the degree-corrected block model, and so are faced with a classic problem of statistical model selection. The classic frequentist approaches to model selection are largely based on likelihood ratios [6], and we follow that approach here. Since both SBM and DC models have many hidden variables, calculating likelihood ratios is itself non-trivial; the likelihood must be summed over all partitions of nodes into blocks, so (in statistical physics terms) the log-likelihood is a free energy. We approximate this free energy using belief propagation, giving a highly scalable algorithm that can deal with large sparse networks in nearly linear time. However, even with the likelihoods in hand, it turns out that the usual χ^2 theory for likelihood ratios relies on approximations which are invalid in our setting, because of the dependency and sparsity of network data. We derive the correct asymptotics under certain assumptions, recovering the classic asymptotics in the limit of dense graphs, but finding that significant corrections are needed in the sparse case. Numerical experiments confirm the validity of our expressions, and we apply our method to a range of real and synthetic networks.

2 Poisson Stochastic Block Models

We have an observed graph G with n nodes and m edges. We assume G is undirected, though the directed case is only notationally more cumbersome. We want to split the nodes into k communities, taking k to be given *a priori*. (We will address estimating k elsewhere.) To do this, we need to decide whether to use the ordinary or the degree-corrected block model.

Traditionally, stochastic block models are applied to simple graphs, where each entry A_{uv} of the adjacency matrix follows a Bernoulli distribution. Following e.g. [16], we use a multigraph version of the block model, where the A_{uv} are independent and Poisson-distributed. (For simplicity, we ignore self-loops.) In the sparse network regime we are most interested in, this Poisson mode differs only negligibly from the original Bernoulli model [23], but the former is easier to analyze.

2.1 The Ordinary Stochastic Block Model

In both models, each node u has a latent variable $g_u \in \{1, \ldots, k\}$ indicating which of the k blocks it belongs to. The block assignment is then $g = \{g_u\}$. The g_u are IID draws from a multinomial distribution parameterized by γ, where $\gamma_r = P(g_u = r)$ is the prior probability that a given node belongs to block r. Thus $g_u \sim \text{Multi}(\gamma)$. After it assigns nodes to blocks, each model generates the number of edges A_{uv} between each pair of nodes u and v by making an independent Poisson draw for each pair. In the ordinary stochastic block model, the means of these Poisson draws are specified by the $k \times k$ block affinity matrix ω, so $A_{uv}|g \sim \text{Poi}(\omega_{g_u g_v})$. If we could observe the block assignment g along with G, the "complete data" likelihood would be

$$P(G, g \,|\, \omega, \gamma) = \prod_u \gamma_{g_u} \prod_{u<v} \frac{\omega_{g_u g_v}^{A_{uv}} \, \mathrm{e}^{-\omega_{g_u g_v}}}{A_{uv}!} = \prod_r \gamma_r^{n_r} \prod_{r,s=1}^{k} \omega_{rs}^{m_{rs}/2} \, \mathrm{e}^{-\frac{1}{2} n_r n_s \omega_{rs}} \prod_{u<v} \frac{1}{A_{uv}!} \,. \tag{1}$$

Here n_r denotes the number of nodes in block r, and m_{rs} denotes the number of edges connecting block r to block s, or twice that number if $r = s$. The last term is constant in the parameters, and is identically 1 for simple graphs, so we

will discard it in what follows. The log-likelihood is then

$$\log P(G, g \,|\, \omega, \gamma) = \sum_r n_r \log \gamma_r + \frac{1}{2}\left(\sum_{r,s=1}^{k} m_{rs} \log \omega_{rs} - n_r n_r \omega_{rs}\right) . \tag{2}$$

Maximizing (2) over γ and ω gives

$$\widehat{\gamma}_r = \frac{n_r}{n} , \qquad \widehat{\omega}_{rs} = \frac{m_{rs}}{n_r n_s} \tag{3}$$

Of course, the block assignments g are not observed, but rather are what we most want to infer. We could try to infer g by maximizing (2) over ω, γ and g jointly; in terms borrowed from statistical physics, this amounts to finding the *ground state* $\hat{g}$ that minimizes the *energy* $-\log P(G, g \,|\, \omega, \gamma)$. When this $\hat{g}$ can be found, it recovers the correct g *exactly* if the graph is dense enough [5]. But if we wish to infer the parameters γ, ω, or to perform model selection, we are interested in the total probability that the block model generates the network at hand. This is

$$P(G \,|\, \omega, \gamma) = \sum_g P(G, g \,|\, \omega, \gamma) ,$$

where the sum is over all k^n possible block assignments. Again following the physics picture, this is the partition function of the Gibbs distribution of g, and its logarithm is (minus) the *free energy*.

As is usual with latent variable models, we can infer γ and ω using an EM algorithm [20], where the E step approximates the average over g with respect to the Gibbs distribution, and the M step estimates γ and ω in order to maximize that average. One approach to the E step would use a Monte Carlo Markov Chain (MCMC) algorithm to sample g from the Gibbs distribution. However, as we will see below, in order to determine γ and ω it suffices to estimate the marginal distributions of g_u of each u, and joint marginal distributions of (g_u, g_v) for each pair of nodes u, v [12, 17, 4]. As we show in §3, belief propagation efficiently approximates both the free energy $-\log P(G \,|\, \omega, \gamma)$ and these marginals, and for many networks it converges very rapidly. Other methods of approximating the E step are certainly possible, and could be used with our model-selection analysis.

2.2 The Degree-Corrected Block Model

As discussed above, in the SBM any two nodes in the same block have the same degree distribution. Moreover, their degrees are sums of independent Poisson variables, so this distribution is Poisson. As a consequence, the SBM "resists" putting nodes with very different degrees in the same block. This leads to problems with real networks where the degree distribution is highly skewed.

The degree-corrected (DC) model extends the SBM to allow for heterogeneity of degree within blocks. Nodes are assigned to blocks as before, but each node also gets an additional parameter θ_u, which scales the number of edges connecting it to other nodes. Thus

$$A_{uv} | g \sim \text{Poi}(\theta_u \theta_v \omega_{g_u g_v})$$

Varying the θ_u gives any desired degree sequence, at least in expectation. Since setting $\theta_u = 1$ for all u recovers the SBM, that model is nested inside the DC model, which is strictly more general.

The likelihood stays the same if we increase θ_u by some factor c for all nodes in block r, provided we also decrease ω_{rs} for all s by the same factor. Thus identification demands a constraint, and a convenient one forces θ_u to sum to the total degree within each block: $\sum_{u:g_u=r} \theta_u = \sum_{u:g_u=r} d_u$. We denote this total degree D_r. The complete-data likelihood of the DC model is then

$$\begin{aligned} P(G, g \,|\, \theta, \omega, \gamma) &= \prod_u \gamma_{g_u} \prod_{u<v} \frac{(\theta_u \, \theta_v \, \omega_{g_u g_v})^{A_{uv}}}{A_{uv}!} \exp(-\theta_u \, \theta_v \, \omega_{g_u g_v}) \\ &= \prod_r \gamma_r^{n_r} \prod_u \theta_u^{d_u} \prod_{rs} \omega_{rs}^{m_{rs}/2} \exp(-\frac{1}{2} D_r D_s \omega_{rs}) \prod_{u<v} \frac{1}{A_{uv}!} , \end{aligned} \tag{4}$$

where n_r and m_{rs} are as before. Again ignoring the constant term, the log-likelihood is

$$\log P(G, g \,|\, \theta, \omega, \gamma) = \sum_r n_r \log \gamma_r + \sum_u d_u \log \theta_u + \frac{1}{2}\left(\sum_{rs} m_{rs} \log \omega_{rs} - D_r D_s \omega_{rs}\right) . \tag{5}$$

Maximizing (5) yields the MLEs

$$\widehat{\theta}_u = d_u\,, \quad \widehat{\gamma}_r = \frac{n_r}{n}\,, \quad \widehat{\omega}_{rs} = \frac{m_{rs}}{D_r D_s}\,. \tag{6}$$

However, as with the ordinary SBM, we will estimate γ and ω not just for a ground state $\hat{g}$, but using belief propagation to find the marginal distributions for g_u and pairwise marginals for (g_u, g_v).

3 Belief Propagation and the Bethe Free Energy

We referred above to the use of belief propagation for computing free energies and marginal distributions of block assignments. Here we describe how belief propagation works for the degree-corrected block model, extending the treatment of the SBM in [10, 9]. The key idea [29] is that each node u sends a message to every other node v, indicating the marginal distribution of g_u if v were absent. We write $\mu_r^{u\to v}$ for the probability that u would be of type r in the absence of v. Then $\mu^{u\to v}$ gets updated in light of the messages u gets from the *other* nodes as follows. Let

$$f(\theta_u, \theta_v, \omega_{rs}, A_{uv}) = \frac{(\theta_u\theta_v\omega_{rs})^{A_{uv}}}{A_{uv}!}\exp(-\theta_u\theta_v\omega_{rs}) \tag{7}$$

denote the probability that A_{uv} takes its observed value assuming that $g_u = r$ and $g_v = s$. Then

$$\mu_r^{u\to v} = \frac{1}{Z^{u\to v}}\,\gamma_r \prod_{w\neq u,v}\sum_{s=1}^{k}\mu_s^{w\to u} f(\theta_w, \theta_u, \omega_{rs}, A_{wu})\,, \tag{8}$$

where $Z^{u\to v}$ is a normalization factor set so that $\sum_r \mu_r^{u\to v} = 1$. As usual in belief propagation, we assume here that the block assignments g_w of the other nodes are independent conditioned on g_u.

Note that each node sends messages to every other node, not just to its neighbors, since non-edges are also informative about g_u and g_v. Thus we have a Markov random field on a weighted complete graph, as opposed to just on the network itself. However, keeping track of n^2 messages is cumbersome. For sparse networks, we can restore scalability by noticing that, up to $O(1/n)$ terms, each node u sends the same message to all of its non-neighbors. That is, for any v such that $A_{uv} = 0$, we have $\mu_r^{u\to v} = \mu_r^u$ where

$$\mu_r^u = \frac{1}{Z^u}\,\gamma_r \prod_{w\neq u}\sum_{s=1}^{k}\mu_s^{w\to u} f(\theta_w, \theta_u, \omega_{rs}, A_{wu})\,. \tag{9}$$

This simplification reduces the number of messages to $O(n+m)$. We can then write

$$\mu_r^{u\to v} = \frac{1}{Z^{u\to v}}\,\gamma_r \prod_{w\neq v, A_{uw}\neq 0}\frac{\sum_{s=1}^{k}\mu_s^{w\to u} f(\theta_w, \theta_u, \omega_{rs}, A_{wu})}{\sum_{s=1}^{k}\mu_s^{w} f(\theta_w, \theta_u, \omega_{rs}, 0)} \times \prod_{w}\sum_{s=1}^{k}\mu_s^{w} f(\theta_w, \theta_u, \omega_{rs}, 0)\,.$$

Since the second product depends only on θ_u, we can compute it once for each degree in the network, and then update the messages for each u in $O(k^2 d_u)$ time. Thus, for fixed k, the total time it takes to update all the messages is $O(m + \ell n)$, where ℓ is the number of distinct degrees. As discussed in [9], for many networks only a constant number of updates are necessary in order to reach a fixed point, making the entire algorithm quite scalable.

The BP estimate of the joint marginals $\Pr[g_u = r, g_v = s]$ is $b_{rs}^{uv} \propto f(\theta_u, \theta_v, \omega_{rs}, A_{uv})\mu_r^{u\to v}\mu_s^{v\to u}$, normalized so that $\sum_{rs} b_{rs}^{uv} = 1$. The M step of the EM algorithm sets γ and ω analogously to (6),

$$\gamma_r = \frac{\bar{n}_r}{n} = \frac{\sum_u \mu_r^u}{n}\,, \quad \omega_{rs} = \frac{\overline{m}_{rs}}{\overline{D}_r\overline{D}_s} = \left(\sum_{u\neq v: A_{uv}\neq 0} A_{uv} b_{rs}^{uv}\right) \Big/ \left(\sum_u d_u \mu_r^u \sum_u d_u \mu_s^u\right)\,. \tag{10}$$

Belief propagation also lets us approximate the partial-data likelihood, i.e., the total probability summed over g that the model generates G. The *Bethe free energy* is the following approximation to the log partition function [30]:

$$\log P(G\,|\,\theta,\omega,\gamma) \approx \sum_u \log Z^u - \sum_{u\neq v, A_{uv}\neq 0}\log\left[\sum_{rs} f(\theta_u, \theta_v, \omega_{rs}, A_{uv})\mu_r^{u\to v}\mu_s^{v\to u}\right] + \frac{1}{2}\sum_{rs}\omega_{rs}\overline{D}_r\overline{D}_s\,. \tag{11}$$

We reiterate that while we use belief propagation in our numerical work, our results on model selection in the next section are quite indifferent as to *how* the likelihood is maximized, or how the free energy is computed.

4 Model Selection

When the degree distribution is relatively homogeneous within each block (e.g. [11, 15]), the ordinary stochastic block model is better than the degree-corrected model, since the extra parameters θ_u simply lead to over-fitting. On the other hand, when degree distributions within blocks are highly heterogeneous, DC is better. The challenge comes when each model offers a different partition; for instance, when the SBM divides blogs into high- and low-degree groups, and DC divides them according to political leanings. If we lack prior information about which model is a better account of the network, we need to use the data to pick a model, i.e., to do model selection [6].

From the machine-learning perspective, the natural impulse is to reach for multi-fold cross-validation. Unfortunately, because network data is globally dependent, there is as yet no good way to split a given into training and testing sets for cross-validation. Predicting missing links or tagging false positives are popular forms of leave-k-out cross-validation in the network literature [7, 14], but the latter does not converge on the true model even for IID data [6].

Instead, we approach this problem statistically, as one of hypothesis testing. Since the ordinary SBM is nested within the DC model, any given graph G must be at least as likely under the latter as under the former. Moreover, if the SBM really is the better model, the DC should converge to it, at least in the limit of large networks. Our null model $H_0 = \{\gamma, \omega\}$ then is the SBM, and the larger, nesting alternative $H_1 = \{\theta, \gamma, \omega\}$ is the DC model. The appropriate test statistic is the log-likelihood ratio,

$$\Lambda(G) \quad = \quad \log \frac{\sup_{H_1} \sum_g \prod_r \gamma_r^{n_r} \prod_u \theta_u^{d_u} \prod_{rs} \omega_{rs}^{m_{rs}/2} \exp(-\frac{1}{2} D_r D_s \omega_{rs})}{\sup_{H_0} \sum_g \prod_r \gamma_r^{n_r} \prod_{rs} \omega_{rs}^{m_{rs}/2} \exp(-\frac{1}{2} n_r n_s \omega_{rs})} \tag{12}$$

We reject the null model in favor of the more elaborate alternative when Λ exceeds some threshold. This threshold, in turn, is fixed by our desired error rate, and by the distribution of Λ when G is generated from the null model. When G is small, the null-model distribution of Λ can be found through parametric bootstrapping [8]: fitting H_0, generating new graphs $\tilde{G}$ from it, and evaluating $\Lambda(\tilde{G})$. When n is large, however, it would be helpful to replace bootstrapping with analytic calculations.

A classic result in asymptotic statistics [26] asserts that in hypothesis-testing problems like this, the large-sample null distribution of Λ is $2\Lambda(G) \sim \chi^2_\ell$, where ℓ is the number of constraints that must be imposed on H_1 to recover H_0. In this case we have $\ell = n - k$, as we must set all n of the θ_u to 1, while our identifiability convention $\sum_{u:g_u=r} \theta_u = D_r$ already imposed k constraints.

However, deriving the χ^2 distribution relies on a key assumption [26, 13]: namely, that the log-likelihood of both models is well-approximated by a quadratic function in the vicinity of its maximum, so that the parameter estimates have Gaussian distributions around the true model. The most common grounds for this assumption are central limit theorems for IID data, or more generally, being in a "large data limit." We will see that, for sparse networks, this assumption does not hold for the parameters θ_u. Nevertheless, with some work we are able to compute the mean and variance of Λ's null distribution. While we recover the classical χ^2 distribution in the the limit of large, dense graphs, there are significant corrections when the average degree of the graph is small. There corrections need to be taken into account in order to solve this model selection problem correctly.

To obtain theoretical estimates of the null distribution of Λ, we assume that the Gibbs distribution of both models is concentrated on the same block assignment g. This is a major assumption, but it is borne out by our experiments (Fig. 1 and 2), and the fact that under some conditions [5] the SBM recovers the underlying block assignment exactly. Under this assumption, while the free energy differs from the ground state energy by an entropy term, the free energy *difference* between the two models has the same distribution as the ground state energy difference. The MLE estimates for H_0 and H_1 are then given by (3) and (6) respectively. Substituting these into (12) gives Λ the form of a Kullback-Leibler divergence,

$$\Lambda(G) = \log \prod_u d_u^{d_u} \prod_{rs} \left(\frac{n_r n_s}{D_r D_s} \right)^{m_{rs}/2} = \log \prod_u \left(\frac{d_u}{\overline{d}_{g_u}} \right)^{d_u} = \sum_u d_u \log \frac{d_u}{\overline{d}_{g_u}} \,, \tag{13}$$

where $\overline{d}_{g_u} = D_{g_u}/n_u$ is the average degree of u's block. Note that $\overline{d}_r$ is the empirical mean, not the expected degree $\mu_r = \sum_s \gamma_s \omega_{rs}$ of the true underlying SBM.

We can understand the asymptotic null distribution of Λ by assuming that the d_u in each block r are IID and Poisson with expectation μ_r. This assumption is sound in the limit $n \to \infty$, since the correlations between node degrees are

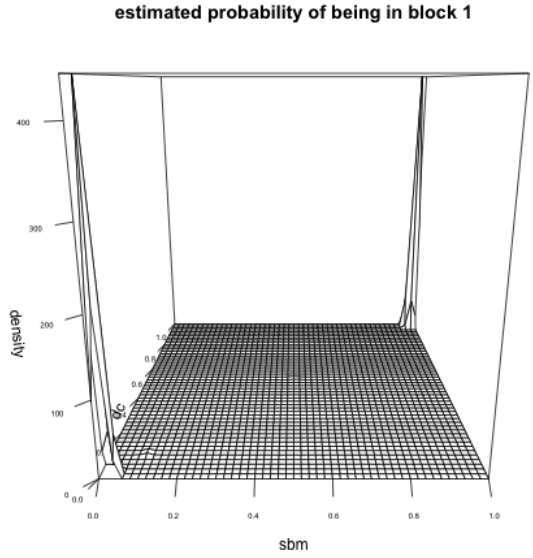


Figure 1: Joint density of posterior probabilities over block assignments, showing that the SBM and the DC are concentrated around the *same* ground state. The synthetic network had $n = 10^3$, $k = 2$, $\gamma_1 = \gamma_2 = 1/2$, $\mu_r = 11$, $\omega_{12}/\omega_{11} = \omega_{21}/\omega_{22} = 1/11$. The x and y axes are the marginal probabilities of being in block 1 according to the SBM and DC models.

$O(1/n)$. In that case, we can compute the expectation and variance of Λ analytically (see Appendix A). These results show how the behavior of Λ differs from naive χ^2 asymptotics, as well as revealing the limits where the naive results apply. Specifically, we have

$$\mathbb{E}[\Lambda] = \sum_r n_r f(\mu_r) - f(n_r \mu_r) \tag{14}$$

where if d is Poisson with mean μ,

$$f(\mu) = \mathbb{E}[d \log d] - \mu \log \mu = \sum_{d=1}^{\infty} \frac{\mathrm{e}^{-\mu} \mu^d}{d!} \, d \log d - \mu \log \mu \, . \tag{15}$$

In the limit $\mu \to \infty$, i.e., for dense graphs, both $f(\mu)$ and $f(n\mu)$ approach $1/2$, and (14) gives $\mathbb{E}[\Lambda] = (n-k)/2$ just as in the standard χ^2 analysis. However, when μ is finite, $f(\mu)$ differs significantly from $1/2$.

The variance of Λ is more complicated, but still calculable. The limiting variance per node is

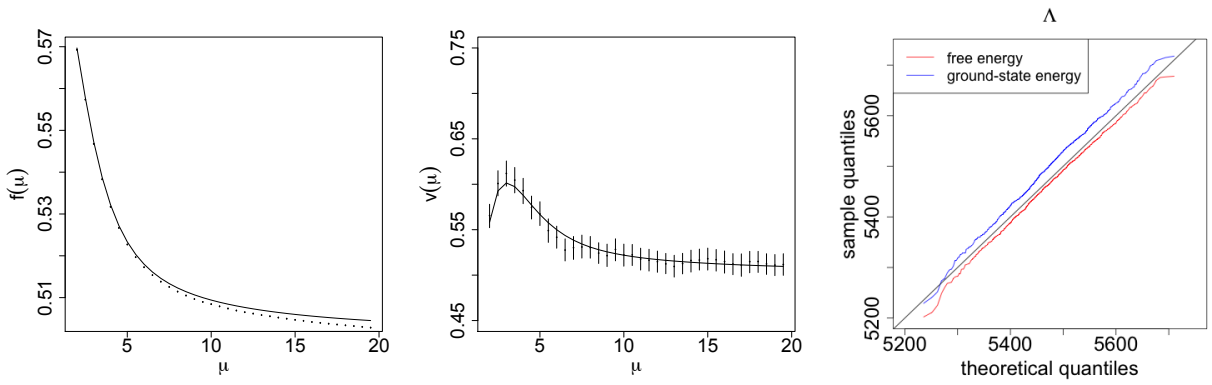
$$\lim_{n\to\infty} \frac{1}{n} \mathrm{Var}[\Lambda] = \sum_r \gamma_r v(\mu_r) \, , \tag{16}$$

where, again taking d to be Poisson with mean μ,

$$v(\mu) = \mu(1 + \log \mu)^2 + \mathrm{Var}[d \log d] - 2(1 + \log \mu) \, \mathrm{Cov}[d, d \log d] \, . \tag{17}$$

Since the variance of χ^2_ℓ is 2ℓ, the χ^2 analysis would predict $(1/n)\mathrm{Var}[\Lambda] = 1/2$. Indeed $v(\mu)$ approaches $1/2$ in the limit $\mu \to \infty$, but like $f(\mu)$ it differs significantly from $1/2$ for finite μ. Plots of $f(\mu)$ and $v(\mu)$, and leading corrections to the classical asymptotics, are given in Appendix A.

Why exactly does the null distribution of Λ differ from the usual χ^2 distribution? The reason is that the parameters θ_u are not in the large data limit. We have one observation for each node, i.e., its degree d_u. If a Poisson distribution has small mean, its shape differs significantly from a Gaussian, and so does the posterior distribution of the mean based on a single sample. In particular, $P(\theta \mid d)$ follows a Gamma distribution, if the prior on θ is uninformative [32]. When



(a) (b) (c)

Figure 2: (a) $f(\mu)$ from (15), the expected log-likelihood difference per node, compared to simulation results; (b) the asymptotic variance of the log-likelihood difference per node, from (17), with simulation results; (c) QQ plots comparing the distribution of log-likelihood differences from 10^4 synthetic networks with $\mu = 3$ to a Gaussian with the theoretical mean and variance, showing that the free energy difference and the ground state energy difference have similar distributions. All simulations used $n = 10^4$, $k = 2$, $\gamma_1 = \gamma_2 = 1/2$, and $\omega_{12}/\omega_{11} = 0.15$, $\omega_{11}/\omega_{22} = 1$; in (a) and (b), each point is the average over 10^3 networks, including 95% bootstrap confidence intervals.

the degrees are large, both the sample distribution and the posterior become Gaussian, and the χ^2 analysis takes over; but when they are small, the geometry is simply different, causing $f(\mu)$ and $v(\mu)$ to differ from $1/2$.

As shown in Fig. 2, experiments on synthetic networks generated from the SBM show that the mean and variance of Λ are very well fit by our theoretical results. We have not attempted to compute higher moments of Λ. However, if we assume that d_u are independent, then the central limit theorem applies, and Λ follows a Gaussian distribution in the limit of large n. Quantile plots from the same experiments (Fig. 2(c)) show that a Gaussian with mean and variance given by (14) and (16) is indeed a good fit. Moreover, the free energy difference and the ground state energy difference have similar distributions, as implied by our assumption that both Gibbs distributions are concentrated around the ground state. Interestingly, in Fig. 2(c), the degree is low enough that this concentration must be imperfect, but our theory still holds remarkably well.

5 Results on real world networks

We have derived the theoretical null distribution of Λ, and backed up our calculations with simulations. We now apply our theory to the real world, considering two examples studied in [16].

5.1 Zachary's karate club

This is a social network consisting of 34 members of a karate club, where undirected edges represent friendships [31]. The club split into two factions, one centered around the instructor and the other around the club president. The network is thus made up of two assortative blocks, each with a high degree hub and lower-degree peripheral nodes.

The authors of [16] compared the perfomance of SBM and DC on this network, and heavily favored DC over SBM because the former leads to a community structure agreeing with the ground truth. Our test, however, shows that the evidence is not strong enough to reject the null SBM model with any great confidence. As shown in Fig. 3(a), the distribution of Λ from bootstrap experiments is fit reasonably will by a Gaussian with our predicted mean and variance. The observed $\Lambda = 20.7$ has a p-value of 0.19 according to the theoretical Gaussian, and 0.15 according to the bootstrap distribution. Thus a prudent statistician would think twice before embracing the additional n parameters of DC. Indeed, in a study of active learning, the authors of [18] found that SBM labels most of the nodes correctly if

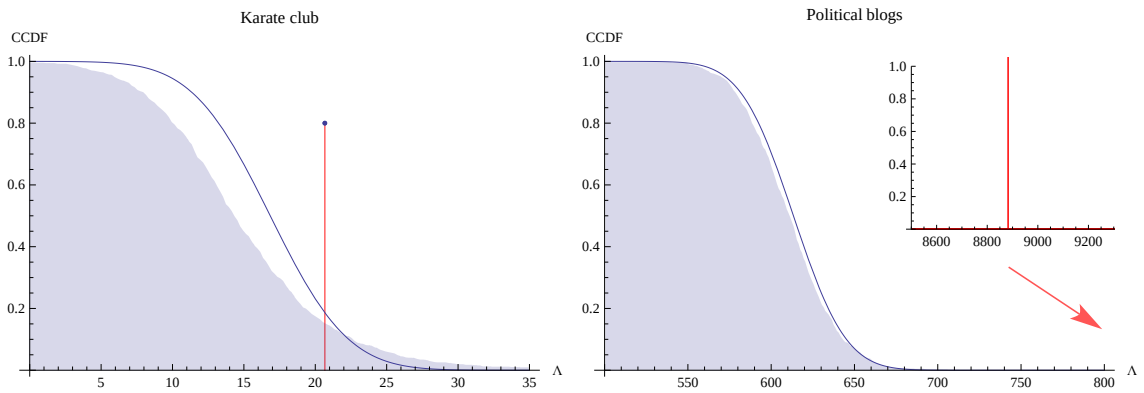


Figure 3: Hypothesis testing of real world networks. (a): Zachary's karate club [31], where $n = 34$. The CCDF (complementary cumulative distribution) of the log-likelihood ratio Λ under the null model is estimated using bootstrapping (shaded), and is fit reasonably well by the CCDF of a Gaussian (curve) with our theoretically predicted mean and variance. The observed $\Lambda = 20.7$ (marked with the red line) has p-values of 0.15 and 0.19 according to the bootstrap and theoretical distributions respectively. (b): A network of political blogs [1] where $n = 1222$. The bootstrap distribution (shaded) is very well fit by a Gaussian (curve) with our predicted mean and variance. The actual log-likelihood ratio is so far in the tail (see inset) that its p-value is effectively zero. Thus for the blog network, we can decisively reject the ordinary block model in favor of the degree-corrected model, while for the karate club, the evidence is less clear.

we fix the block assignment of the instructor and the president to 1 and 2 respectively. This implies that the degree inhomogeneity is not too extreme, and that only a handful of nodes are responsible for the better performance of DC.

5.2 Political blogs

The second example is a network of political blogs in the US assembled by Adamic and Glance [1]. As in [16], we focus on the giant component, which consists of 1222 blogs and 19087 links between them, as captured on a single day in 2005. The blogs have known political leanings, and were labeled as either liberal or conservative. The network is assortative and has a highly right-skewed degree distribution within each block.

In its agreement with ground truth, DC substantially outperforms SBM, as observed in [16]. This time around, our hypothesis testing procedure completely agrees with their choice of model. As shown in Fig. 3(b), the bootstrap distribution of Λ is very well fit by a Gaussian with our theoretical prediction of the mean and variance. The observed log-likelihood ratio $\Lambda = 8883$ is 330 standard deviations above the mean. It is essentially impossible to produce such extreme results through mere fluctuations under the null model. Thus, for this network, introducing n extra parameters to capture the degree heterogeneity, and rejecting SBM in favor of DC, is fully justified.

6 Conclusion

We have presented a mathematically principled procedure for determining whether the degree-corrected block model is justified over the ordinary stochastic block model. We found that for sparse networks, the distribution of log-likelihood ratios differs significantly from the naive χ^2 analysis, and showed how to compute its mean and variance exactly in the large-n limit where node degrees are essentially independent and Poisson. We confirmed our calculations with experiments on synthetic networks, and applied our procedure to two real-world networks; one where the ordinary block model can be decisively rejected, and another where the evidence is less clear. We hope that similar approaches will let us choose between competing generative models for network data, and in particular between other variants of the block model such as those in [32].

Acknowledgments XY, CM, and YZ were supported by the McDonnell Foundation and by the DARPA GRAPHS program. JJ was supported by the NSF as a Research Experience Undergraduate intern at the Santa Fe Institute.

FK, LZ, and PZ were supported by the Projet DIM "problématique transversales aux systèmes complexes" of the Institut des Systèmes Complexes, Paris Ìle-de-France (ISC-PIF). CRS was supported by grants from the NIH (#2 R01 NS047493) and INET (IN01100005). We are grateful to Mark Newman for helpful conversations.

References

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 US election: divided they blog. *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.

[2] E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing. Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, 9:1981–2014, 2008.

[3] Stefano Allesina and Mercedes Pascual. Food web models: a plea for groups. *Ecology letters*, 12:652–662, July 2009.

[4] M. J. Beal and Z. Ghahramani. Variational Bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis*, 1:793–832, 2006.

[5] Peter J. Bickel and Aiyou Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences (USA)*, 106:21068–21073, 2009.

[6] Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Cambridge University Press, Cambridge, England, 2008.

[7] Aaron Clauset, Cristopher Moore, and Mark E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.

[8] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Applications*. Cambridge University Press, Cambridge, England, 1997.

[9] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84, 2011.

[10] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Inference and Phase Transitions in the Detection of Modules in Sparse Networks. *Physical Review Letters*, 107, 2011.

[11] S.E. Fienberg and S. Wasserman. Categorical data analysis of single sociometric relations. *Sociological Methodology*, pages 156–192, 1981.

[12] N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society B*, 70:589, 2008.

[13] Charles J. Geyer. Le Cam made simple: Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity. Technical Report 643, School of Statistics, University of Minnesota, 2005.

[14] R. Guimera and M. Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci. U. S. A.*, 106:22073–22078, 2009.

[15] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5:109–137, 1983.

[16] Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83, 2011.

[17] S. Lin, B. Sturmfels, and Z. Xu. Marginal likelihood integrals for mixtures of independence models. *The Journal of Machine Learning Research*, 10:1611–1631, 2009.

[18] Cristopher Moore, Xiaoran Yan, Yaojia Zhu, Jean-Baptiste Rouquier, and Terran Lane. Active learning for node classification in assortative and disassortative networks. In *Proc. 17th KDD*, pages 841–849, 2011.

[19] M. Mørup and L.K. Hansen. Learning latent structure in complex networks. *NIPS Workshop on Analyzing Networks and Learning with Graphs*, 2009.

[20] Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 355–368, Dordrecht, 1998. Kluwer Academic.

[21] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89:208701, 2002.

[22] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67:026126, 2003.

[23] Patrick O. Perry and Patrick J. Wolfe. Null models for network data. *Biometrika*, submitted, 2012.

[24] Mason A. Porter, Jukka-Pekka Onnela, and Peter J. Mucha. Communities in networks. *Notices of the American Mathematical Society*, 56:1082–1097 and 1164–1166, 2009.

[25] Jörg Reichardt, Roberto Alamino, and David Saad. The interplay between microscopic and mesoscopic structures in complex networks. *PLoS ONE*, 6:e21282, 2011.

[26] Mark J. Schervish. *Theory of Statistics*. Springer-Verlag, Berlin, 1995.

[27] T.A. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14:75–100, 1997.

[28] S. Wasserman and C. Anderson. Stochastic a posteriori blockmodels: Construction and assessment. *Social Networks*, 9:1–36, 1987.

[29] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. In Gerhard Lakemeyer and Bernhard Nebel, editors, *Exploring Artificial Intelligence in the New Millennium [IJCAI 2001]*, pages 239–269, San Francisco, 2003. Morgan Kaufmann.

[30] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Constructing Free-Energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51:2282–2312, 2005.

[31] W. W Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

[32] Yaojia Zhu, Xiaoran Yan, and Cristopher Moore. Generating and inferring communities with inhomogeneous degree distributions. 2012.

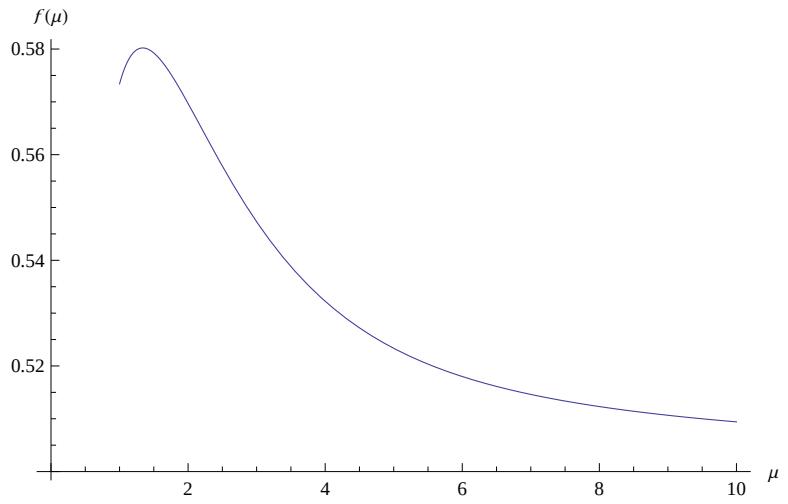


Figure 4: The function $f(\mu)$ defined in (19), or equivalently the expected log-likelihood difference divided by n. We compare this with experiment in Fig. 2(a).

A Behavior of Λ under the null hypothesis

For simplicity we focus on one group with expected degree μ. Assuming independence between the groups will then recover the expressions (14) and (16) where the mean and variance of Λ is a weighted sum over groups. We have

$$\begin{aligned}\Lambda &= \sum_{i=1}^{n} d_i \log \frac{d_i}{\overline{d}} \\ &= \sum_i d_i \log d_i - \left(\sum_i d_i\right) \log \left(\sum_i d_i\right) + \left(\sum_i d_i\right) \log n \,, \end{aligned} \tag{18}$$

where $\overline{d} = (1/n) \sum_i d_i$ is the sample mean. We wish to compute the mean and expectation of $\log L$ if the data is generated by the null model.

If d is Poisson-distributed with mean μ, let $f(\mu)$ denote the difference between the expectation of $d \log d$ and its most likely value $\mu \log \mu$:

$$f(\mu) = \left(\sum_{d=1}^{\infty} \frac{\mathrm{e}^{-\mu} \mu^d}{d!} \, d \log d \right) - \mu \log \mu \,. \tag{19}$$

Assume that the d_i are independent and Poisson with mean μ; this is reasonable in a large sparse graph, since the correlations between degrees of different nodes is $O(1/n)$. Then $\sum_i d_i$ is Poisson with mean $n\mu$, and (18) gives

$$\mathbb{E}[\Lambda] = n f(\mu) - f(n\mu) \,. \tag{20}$$

To understand this asymptotically, note that $f(\mu)$ converges to $1/2$ when μ is large. Thus in the limit of large n,

$$\mathbb{E}[\Lambda] = n f(\mu) - \frac{1}{2} \,.$$

When μ is large, this gives $\mathbb{E}[\Lambda] = (n-1)/2$, just as χ^2 hypothesis testing would suggest. However, as Fig. 4 shows, $f(\mu)$ deviates significantly from $1/2$ for finite μ. We can obtain the leading corrections as a power series in $1/\mu$ by approximating (19) with the Taylor series of $d \log d$ around $d = \mu$, giving

$$f(\mu) = \frac{1}{2} + \frac{1}{12\mu} + \frac{1}{12\mu^2} + O(1/\mu^3) \,.$$

Computing the variance is harder, but still possible. It will be convenient to define several functions. If d is Poisson with mean μ, let $\phi(\mu)$ denote the variance of $d \log d$:

$$\begin{aligned}\phi(\mu) = \operatorname{Var}[d \log d] &= \mathbb{E}\big[(d \log d)^2\big] - \mathbb{E}[d \log d]^2 \\ &= \sum_{d=0}^{\infty} \frac{\mathrm{e}^{-\mu}\mu^d}{d!}(d \log d)^2 - \left(f(\mu) + \mu \log \mu\right)^2 \,. \end{aligned} \tag{21}$$

We will also use

$$\begin{aligned}c(\mu) = \operatorname{Cov}[d, d \log d] &= \mathbb{E}\big[d^2 \log d\big] - \mu \mathbb{E}[d \log d] \\ &= \sum_{d=1}^{\infty} \frac{\mathrm{e}^{-\mu}\mu^d}{d!}\, d^2 \log d - \mu \left(f(\mu) + \mu \log \mu\right) \,. \end{aligned} \tag{22}$$

Finally, let $\lambda \geq \mu$, and let d and u be independent and Poisson with mean μ and $\lambda - \mu$ respectively. Then let

$$\begin{aligned} r(\mu, \lambda) &= \operatorname{Cov}[d \log d, (d+u)\log(d+u)] \\ &= \mathbb{E}[(d \log d)((d+u)\log(d+u))] - \mathbb{E}[d \log d]\, \mathbb{E}[(d+u)\log(d+u)] \\ &= \sum_{d,u=1}^{\infty} \frac{\mathrm{e}^{-\lambda}\mu^d(\lambda-\mu)^u}{d!u!}(d \log d)((d+u)\log(d+u)) \\ &\quad - \left(f(\mu) + \mu \log \mu\right)\left(f(\lambda) + \lambda \log \lambda\right) \,, \end{aligned} \tag{23}$$

where we used the fact that $d + u$ is Poisson with mean λ.

Then again assuming that the d_i are independent, we have the following terms and cross-terms for the variance of (18):

$$\begin{aligned} \operatorname{Var}\left[\sum_i d_i \log d_i\right] &= n\phi(\mu) \\ \operatorname{Var}\left[\left(\sum_i d_i\right) \log\left(\sum_i d_i\right)\right] &= \phi(n\mu) \\ \operatorname{Var}\left[\sum_i d_i\right] &= n\mu \\ \operatorname{Cov}\left[\sum_i d_i \log d_i,\, \left(\sum_i d_i\right) \log\left(\sum_i d_i\right)\right] &= nr(\mu, n\mu) \\ \operatorname{Cov}\left[\sum_i d_i \log d_i,\, \sum_i d_i\right] &= nc(\mu) \\ \operatorname{Cov}\left[\left(\sum_i d_i\right) \log\left(\sum_i d_i\right),\, \sum_i d_i\right] &= c(n\mu) \end{aligned}$$

Putting this all together, we have

$$\operatorname{Var}[\Lambda] = n\phi(\mu) + \phi(n\mu) + n\mu \log^2 n - 2nr(\mu, n\mu) + 2\big(nc(\mu) - c(n\mu)\big) \log n \,. \tag{24}$$

In the limit of large μ, using Taylor series to expand the summands of (21) and (22) gives the following simplifications:

$$\begin{aligned} \phi(\mu) &= \mu \log^2 \mu + 2\mu \log \mu + \mu + \frac{1}{2} + O\left(\frac{\log \mu}{\mu}\right) \\ c(\mu) &= \mu \log \mu + \mu + O(1/\mu) \,. \end{aligned}$$

Also, when $\lambda \gg \mu$ and $\mu = O(1)$, using $\log(d+u) \approx \log u + d/u$ lets us separate the double sum in (23), giving

$$\begin{aligned} r(\mu, \lambda) = \mathbb{E}\big[d^2 \log d\big]\,(1 + \log \lambda) + \mathbb{E}[d \log d]\, \mathbb{E}[u \log u] \\ - \mathbb{E}[d \log d]\, \mathbb{E}[(d+u)\log(d+u)]) + O(1/\lambda) \,. \end{aligned}$$

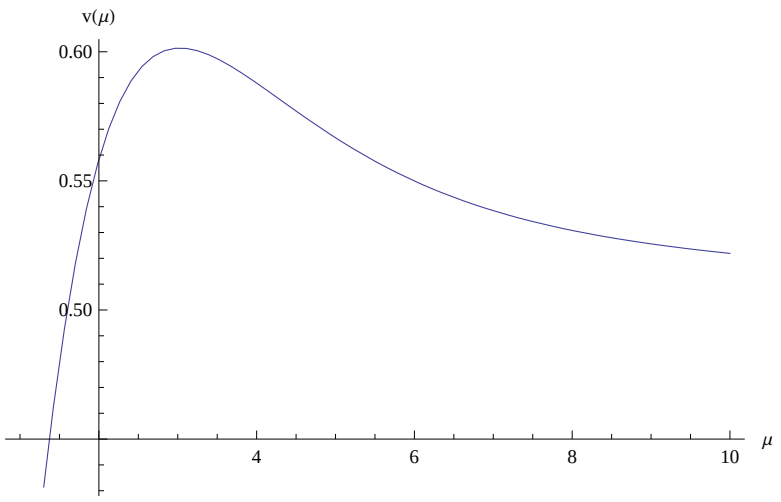


Figure 5: The asymptotic variance of the log-likelihood difference, divided by n, given in (25). We compare this with experiment in Fig. 2(b).

In particular, setting $\lambda = n\mu$ gives

$$r(\mu, n\mu) = c(\mu)(1 + \log n\mu) + O(1/n)\,.$$

Finally, keeping $O(n)$ terms in (24) and defining $v(\mu)$ as in (16) gives

$$v(\mu) = \lim_{n\to\infty} \frac{1}{n}\mathrm{Var}[\Lambda] = \phi(\mu) + \mu(1 + \log\mu)^2 - 2c(\mu)(1 + \log\mu)\,. \tag{25}$$

Using the definitions of ϕ and c, we can write this more explicitly as (where Var and Cov denote the variance and covariance in the Poisson distribution with mean μ)

$$\begin{aligned} v(\mu) &= \mu(1 + \log\mu)^2 + \mathrm{Var}[d\log d] - 2(1 + \log\mu)\,\mathrm{Cov}[d, d\log d] \\ &= \mu(1 + \log\mu)^2 \\ &\quad + \sum_{d=1}^{\infty} \frac{\mathrm{e}^{-\mu}\mu^d}{d!}(d\log d)\big(d\log d - 2(1 + \log\mu)(d - \mu)\big) \\ &\quad - \left(\sum_{d=1}^{\infty} \frac{\mathrm{e}^{-\mu}\mu^d}{d!}\, d\log d\right)^2 . \end{aligned} \tag{26}$$

We plot this function in Fig. 5. It converges to $1/2$ in the limit of large μ, but it is significantly larger for finite μ.