Algorithms **2012**, *5*, 629-635; doi:10.3390/a5040629

OPEN ACCESS

algorithms

ISSN 1999-4893

www.mdpi.com/journal/algorithms

Article

Testing Goodness of Fit of Random Graph Models

Villő Csiszár [1], Péter Hussami [2], János Komlós [3], Tamás F. Móri [1], Lídia Rejtő [2,4,]* and Gábor Tusnády [2]

[1] Department of Probability Theory and Statistics, Eötvös Loránd University, Budapest, Hungary 1053; E-Mails: villo@ludens.elte.hu (V.C.); moritamas@ludens.elte.hu (T.F.M.)

[2] Alfréd Rényi Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary 1053; E-Mails: haprim@yahoo.com (P.H.); tusnady.gabor@renyi.mta.hu (G.T.)

[3] Department of Mathematics, Rutgers University, New Brunswick, NJ 08901, USA; E-Mail:komlos@math.rutgers.edu (J.K.)

[4] Statistics Program, University of Delaware, Newark, DE 19716, USA

* Author to whom correspondence should be addressed; E-Mail: rejto@udel.edu;
Tel.: +1-302-831-8034; Fax: +1-302-831-6243.

Received: 7 May 2012; in revised form: 8 November 2012 / Accepted: 30 November 2012 / Published: 6 December 2012

Abstract: Random graphs are matrices with independent 0–1 elements with probabilities determined by a small number of parameters. One of the oldest models is the Rasch model where the odds are ratios of positive numbers scaling the rows and columns. Later Persi Diaconis with his coworkers rediscovered the model for symmetric matrices and called the model beta. Here we give goodness-of-fit tests for the model and extend the model to a version of the block model introduced by Holland, Laskey and Leinhard.

Keywords: random graph; maximum likelihood; rank entropy

1. Introduction

Let n be a positive integer, $1 \leq i, j \leq n$, and $\varepsilon(i,j)$ independent random variables such that $\varepsilon(i,j) = \varepsilon(j,i)$ and $\varepsilon(i,i) = 0$, furthermore

$$P(\varepsilon(i,j) = 1) = p_{i,j} = p + p_i + p_j, \quad 1 \leq i < j \leq n \tag{1}$$

where the sum of the p_i-s is zero. The least square estimate $\hat{p}$ of p is the average of the epsilons, and the least square estimate of p_i is the average of the differences $\varepsilon(i,j) - \hat{p}$. The modification of the model for non-symmetric matrices is straightforward, and in that case the statistical inference is practically a two-way analysis of variance. Perhaps this is the simplest random graph model but it shares the inconvenient property of many other random graph models that it is hard to ensure that edge probabilities remain in the interval $(0, 1)$. If we use the odds

$$r_{i,j} = \frac{p_{i,j}}{1 - p_{i,j}} \tag{2}$$

instead of the probabilities, then it is enough to ensure the positivity of $r_{i,j}$-s. This is the case in the model introduced by George Rasch [1]. Historically the odds were defined as the ratios of scaling factors for rows and columns but we prefer the multiplicative form

$$r_{i,j} = \beta_i \gamma_j \tag{3}$$

for non-symmetric and

$$r_{i,j} = \beta_i \beta_j \tag{4}$$

for symmetric case. Statistical investigation of the model started with Andersen [2] (see also [3–5]) and later Persi Diaconis with his coworkers rediscovered the model and introduced the name *beta-model* for its parameter. The model has many attractive properties (see in [6–11]):

- degree sequences are sufficient statistics;
- the model covers practically all possible expected degree sequence;
- the conditional distribution of the graphs on condition of a prescribed degree sequence is uniform on the set of all graphs with the given degree sequences.

Statistical inference emerged from Gaussian distribution and later was extended to random variables in Euclidean spaces but the statistical inference on discrete structures is rather sparse [12–16]. Mathematical investigation of graphs has its own history. Nowadays instead of graphs, we are speaking of networks [17] where the most investigated model is the stochastic block model introduced by Holland, Laskey and Leinhard ([18]). Here the vertices are labeled by small numbers or colors and edge probabilities depend only on the labels [19,20]. With an eye on preferential attachment where degree sequences follow scale-free power-law, the block model was criticized because it has moderated flexibility on degree sequences. Chung, Lu, and Vu [21] introduced a model with independent vertices. Chaughuri, Chung and Tsiatas [22] introduced the *planted partition model* (see also [23]). Karrer and Newman [24] proposed another extension of the block model. A natural extension of these models is the unification of the beta and block models:

$$r_{i,j} = b(i, c(j))b(j, c(i)) \tag{5}$$

where $b(.,.)$ is a positive matrix with n rows and k columns, and $c(i)$ is the label of the i-th vertex *i.e.*, it is an integer between 1 and k. We call the model *k-beta model.* The estimation of the labels in block models is possible by the spectral method [25]. It is generally believed that eigenvectors and eigenvalues of the matrix $\varepsilon(i, j)$ tells everything of the structure of the graph [22,26–30], while there are many attempts to provide more flexible models [31,32].

2. Goodness-of-Fit

We cannot test edge-independence on a single graph. While i.i.d. sample is common in statistical inference, in case of graphs the sample generally means a copy of a graph. Perhaps the number one question in statistical inference is the following. Let

$$p_1, \ldots, p_n \tag{6}$$

be an arbitrary given sequence of probabilities, and

$$\varepsilon_1, \ldots, \varepsilon_n \tag{7}$$

be independent $0-1$ variables such that $P(\varepsilon_i = 1) = p_i$. Can we test the model? A randomized answer is the following. Let

$$u_1, \ldots, u_n \tag{8}$$

be independent and uniformly distributed in $(0, 1)$. Then

$$x_i = p_i u_i \varepsilon_i + (1 - \varepsilon_i)(p_i + (1 - p_i)u_i), \quad i = 1, \ldots, n \tag{9}$$

are independent and uniformly distributed in $(0, 1)$, which we can test. Another more practical solution is ordering the pairs (p_i, ε_i) according to the p_i-s in increasing order and compare their partial sums. Or we can clump them into blocks of small number and compare again the sums. All these possibilities hold for graphs with estimated edge probabilities. Let us partition the edges of the complete graph according to the blocks formed with respect to the edge probabilities. In each portion the edge probabilities are close to each other whence the $\varepsilon_{i,j}$-s corresponding to that portion behave like a pure random graph, which we again can test, e.g., by their sums on subsets of vertices.

Diaconis with coworkers [9,33] propose for testing the beta model the following general procedure. Let us choose any graph statistic and determine it on our graph. Let us generate as many graphs as we can with the same degree sequence that the investigated graph has according to the uniform distribution, and let us calculate the chosen statistics. If the value of the sample graph is inside the generated numbers, we accept the beta model, otherwise reject it. One can ask, are there any effect of the choice on the power of the test?

We have found by computer simulations that graphs generated by beta model have only one eigenvalue proportional with n, and all the others are of order $\sqrt{n}$. We think that it is a characteristic property of beta graphs. One wonders that

- if beta model covers all possible degree sequences
- the conditional distribution is uniform over graphs sharing the same degree sequence

then how is it possible that graph behaves differently from typical graphs generated by beta model? Of course there are graphs having many large eigenvalues. But where are they coming from once beta model can generate all the graphs? A possible solution of the catch is the following.

Let us generate a meta-graph from graphs sharing the same degree sequence. Let us say that neighborhood in this meta-graph is given by one single swap. If we have four vertices A, B, C, D in a graph such that AC and BD are edges but AD and BC are not, then changing existence into non-existence among these edges we form a new graph with the same degree sequence. The degree of a graph in this meta-graph goes parallel with the second largest eigenvalue: typical beta model graphs have minimal degree and any increase in their degree results in a more complicated eigenvalue structure. Perhaps the degree in the meta-graph is the most characteristic statistic for beta model.

3. The k-Beta Model

The maximum likelihood equations for the parameters $b(.,.)$ in (5) say that the expected values of degrees *inside* all the subgraph with a given pair of labels should be the same us in the given graph. This is the case when the labels are known. With unknown labels we can form a two-level optimization: for each label set, first determine the parameters $b(.,.)$, then change a small number of labels and repeat the calculation of the parameters. But the procedure is slow even for graphs of moderate size. Spectral methods available for block models fail for coloring k-beta models because the model lose the well-pronounced checkerboard character of block models. It is the ANOVA that offers an applicable algorithm. For any set C of labels $c(.)$, let us calculate the statistic

$$Q(C) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} (\varepsilon(i,j) - u(c(i),c(j)) - v(i,c(j)) - v(j,c(i)))^2 \tag{10}$$

where

$$u(s,t) = \frac{\sum_{c(i)=s} \sum_{c(j)=t} \varepsilon(i,j)}{\sum_{c(i)=s} \sum_{c(j)=t} 1} \tag{11}$$

and

$$v(i,t) = \frac{\sum_{c(j)=t} (\varepsilon(i,j) - u((c(i),t))}{\sum_{c(j)=t} 1} \tag{12}$$

$Q(C)$ is the sum of two way ANOVA sum of squares calculated independently for subgraphs defined for pairs of labels. Starting from a uniform random set C of labels on the vertices and perturbing small number of labels in the individual steps, a simple greedy optimization results in a good set of labels, which is close to the original (true) labels.

For evaluating the character of a random graph, we use the number

$$\exp(-\frac{\sum_{i=2}^{n} \sum_{j=1}^{i-1} (p(i,j) \log p(i,j) + (1-p(i,j)) \log(1-p(i,j))}{n(n-1)/2}) \tag{13}$$

We call it *delogarithmed average entropy* or DAE. This is a number between 1 and 2. If it is close to one, the graph is almost deterministic: the probabilities are close to 0 or 1. In checkerboard block models it means that empty and full subgraphs are amalgamated together. If DAE is close to 2, then the graph has no structure at all. DAE depends on edge density, too. The above tendency is valid for edge density $\frac{1}{2}$, for other edge densities the cut point is closer to 1. According to our experience, if DAE is smaller than 1.9 while edge density is half, then we are able to reconstruct the original labels. For these graphs, the number of non-trivial eigenvalues is $2k-1$, thus the spectrum determines the number of different labels.

The k-beta model has a sister model

$$r_{i,j} = \sum_{s=1}^{k} b(i,s) b(j,s) \tag{14}$$

which we call *small odds rank* model. Strictly speaking, we ought to redefine the diagonal of odds matrix, but perhaps the name is permissible without doing so. The maximum likelihood estimation of parameters in small odds rank models is straightforward and the block structure is detectable in the estimated parameters. Actually the block model is in the intersection of k-beta and small odds rank models, thus if there is any block structure in the graph, it is detectable even in fitting k-beta model to the graph. But if there is no block structure and we are trying to use ANOVA coloring for a small odds rank graph, then the algorithm is no longer stable—it results in different local minima in each run.

Acknowledgements

The authors thank the referees for several valuable remarks. Tamás F. Móri's research was supported by OTKA grant 12574.

References

1. Rasch, G. *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*; Nielsen & Lydiche: Oxford, UK, 1960.
2. Andersen, E.B. Sufficient statistics and latent trait models. *Psychomretrika* **1977**, *42*, 69–81.
3. Linacre, J.M. Predicting responses from Rasch measures. *J. Appl. Meas.* **2010**, *11* , 1–10.
4. Ponicny, I. Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika* **2001**, *66*, 437–460.
5. Verhelst, N. Testing the unidimensionality assumption of the Rasch model. *Meth. Psychol. Res. Online* **2001**, *6*, 231–271.
6. Bickel, P.J.; Chen, A. A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 21068–21073.
7. Barvinok, A.; Hartigan, J.A. An asymptotic formula for the number of non-negative integer matrices with prescribed row and column sums. Available online: http://arxiv.org/abs/0910.2477 (accessed on 5 April 2010).
8. Barvinok, A.; Hartigan, J.A. The number of graphs and a random graph with a given degree sequence. Available online: http://arxiv.org/abs/1003.0356 (accessed on 2 December 2011).
9. Blitzstein, J.; Diaconis, P. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *J. Int. Math.* **2010**, *6*, 489–522.
10. Chatterjee, S.; Diaconis, P.; Sly, A. Random graphs with a given degree sequence. *Ann. Stat.* **2010**, *21*, 1400–1435.
11. Ogawa, M.; Hara, H.; Takemura, A. Graver basis for an undirected graph and its application to testing the beta model of random graphs. *Ann. Inst. Stat. Mat.* **2012**, doi:10.1007/s10463-012-0367-8.
12. Bolla, M.; Tusnády, G. Spectra and optimal partitions of weighted graphs. *Discret. Math.* **1994**, *128*, 1–20.

13. Csiszár, V.; Rejtő, L.; Tusnády, G. Statistical Inference on Random Structures. In *Horizon of Combinatorics*; Győri, E., Katona, G.O.H., Lovász, L., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 37–67.
14. Csiszár, V.; Hussami, P.; Komlós, J.; Móri, T.; Rejtő, L.; Tusnády, G. When the degree sequence is a sufficient statistic. *Acta Math. Hung.* **2011**, *134*, 45–53.
15. Hussami, P. Statistical Inference on Random Graphs. Ph.D. Thesis, Central European University, Budapest, Hungary, 2010.
16. Nepusz, T.; Négyessy, L.; Tusnády, G.; Bazsó, F. Reconstructing cortical networks: Case of directed graphs with high level of reciprocity. *Dyn. Syst. Appl.* **2009**, *18*, 335–362.
17. Newman, M.; Barabási, A.-L.; Watts, D. The Structure and Dynamics of Networks. In *Princeton Studies in Complexity*; Princeton University Press: Princeton, NJ, USA, 2007.
18. Holland, P.; Laskey, K.B.; Leinhardt, S. Stochastic blockmodels: Some first steps. *J. Am. Stat. Assoc.* **1981**, *76*, 33–50.
19. Bickel, P.; Choi, D.; Chang, X.; Zhang, H. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Stat.* **2012**, submitted for publication.
20. Flynn, C.J.; Perry, P.O. Consistent biclustering. Available online: arXiv:1206.6927v1 [stat.ME] (accessed on 29 June 2012).
21. Chung, F.; Lu, L.; Vu, V. Spectra of random graphs with given expected degrees. *Proc. Natl. Acad. Sci. USA* **2003**, *27*, 6313–6318.
22. Chaudhuri, K.; Chung, F.; Tsiatas, A. Spectral clustering of graphs with general degrees in the extended planted partition model. *J. Mach. Learn. Res.* **2012**, *1*, 1–23.
23. Mossel, E.; Neeman, J.; Sly, A. Reconstruction and estimation in the planted partition model. Available online: arXiv:1202.1499v4 [math.PR] (accessed on 22 August 2012).
24. Karrer, B.; Newman, M.E.J. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **2011**, *83*, 016107.
25. Rohe, K.; Chatterjee, S.; Yu, B. Spectral clustrering and high-dimensional stochastic block model. *Ann. Stat.* **2011**, *39*, 1878–1915.
26. Chung, F. *Spectral Graph Theory*; American Mathematical Society: Providence, RI, USA, 1997.
27. Chung, F.; Lu, L. *Complex Graphs and Networks*; American Mathematical Society: Boston, MA, USA, 2006.
28. Lovász, L. Very large graphs. Available online: arXiv:0902.0132 (accessed on 1 February 2009).
29. Lu, L.; Peng, X. Spectra of edge-independent random graphs. Available online: arXiv:1204.6207v1 [math.CO] (accessed on 27 April 2012).
30. Nadakuditi, R.R.; Newman, M.E.J. Spectra of random graphs with arbitrary expected degrees. Available online: arXiv:1208.1275v1 [cs.SI] (accessed on 6 August 2012).
31. Chatterjee, S.; Diaconis, P. Estimating and understanding exponential random graph models. Available online: arXiv:1102.2650v3[math.PR] (accessed on 6 April 2011).
32. Palla, G.; Lovász, L.; Vicsek, T. Multifractal network generator. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 7641–7645.

33. Chen, Y.; Diaconis, P.; Holmes, S.P.; Liu, J.S. Sequential Monte Carlo methods for statistical analysis of tables. *J. Am. Stat. Assoc.* **2005**, *100*, 109–120.