Journal of Statistical Planning and Inference 141 (2011) 1041–1046

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A comparative power analysis of the maximum degree and size invariants for random graph inference

Andrey Rukhin [a], Carey E. Priebe [b,*]

[a] Naval Surface Warfare Center, Dahlgren Division, Dahlgren, VA 22448, United States
[b] Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218-2682, United States

ARTICLE INFO

Article history:
Received 21 April 2009
Received in revised form
8 August 2010
Accepted 11 September 2010
Available online 18 September 2010

Keywords:
Erdos–Renyi random graphs
Statistical inference
Comparative power analysis

ABSTRACT

Let $p,s \in (0,1]$ with $s > p$, let $m,n \in \mathbb{N}$ with $1 < m < n$, and define $V=\{1,\ldots,n\}$. Let $ER(n,p)$ denote the random graph model on V where each edge is independently included in the graph with probability p. Let $\kappa(n,p,m,s)$ denote the random graph model on V where each edge among the m vertices $\{1,\ldots,m\}$ is independently included in the graph with probability s and all other edges are independently included with probability p. We view graphs from the $ER(n,p)$ model as "homogeneous": the probability of the presence of an edge is the same throughout such a graph. On the other hand, we view a graph generated by the κ model as "anomalous"; such a graph possesses increased edge probability among a certain subset of its vertices.

Our inference setting is to determine whether an observed graph G is "homogeneous" (with some known p) or "anomalous". In this article, we analyze the statistical power β of the size invariant $|E(G)|$ (the number of edges in the graph) and the maximum degree invariant $\Delta(G)$ in detecting such anomalies. In particular, we demonstrate an interesting phenomenon when comparing the powers of these statistics: the limit theory can be at odds with the finite-sample evidence even for astronomically large graphs. For example, under certain values of p,s and $m=m(n)$, we show that the maximum degree statistic is more powerful ($\beta_\Delta > \beta_{|E|}$) for $n \leqslant 10^{24}$ while $\lim_{n\to\infty} \beta_\Delta/\beta_{|E|} < 1$.

1. Introduction

In this article, we present a comparative analysis of power characteristics for two statistics within the context of detecting anomalies in random graphs. If a randomly generated graph $G = (V,E)$ over a vertex set $V = [n] = \{1,\ldots,n\}$ is observed for some $n \in \mathbb{N}$, the null hypothesis H_0 of interest is, in general, some version of "homogeneity." The simplest null in this class is the Erdös–Rényi(n,p) model (hereafter $ER(n,p)$) wherein each of the $\binom{n}{2}$ possible edges are included independently in the graph with a fixed probability p. Our corresponding "anomaly" alternative hypothesis H_1 is that the observed graph includes an anomalous subset $K = \{i_1,\ldots,i_m\}$ of m vertices ($m=o(n)$) where the edges are again independently included in the graph — however, an edge (i,j) with $i,j \in K$ is present with probability s where $s > p$, and each of the remaining $\binom{n}{2}-\binom{m}{2}$ edges are present with probability p. We denote our alternative random graph model by $\kappa(n,p,m,s)$. Such stochastic blockmodels have been proposed in Wasserman and Anderson (1987) and considered in an inferential setting in e.g. Nowicki and Snijders (2001).

* Corresponding author.
E-mail address: cep@jhu.edu (C.E. Priebe).

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.09.013

We investigate the distribution theory of the size $|E| = |E(G)|$ and maximum degree $\Delta = \Delta(G)$ statistics under both our null and alternative hypotheses. In both cases, the test H_0 versus H_1 rejects for large values of the statistic as $s > p$. The limiting distributions $(n \to \infty)$ of these statistics are available. Under H_0, the invariant $|E|$ is binomial and, properly normalized, asymptotically normal. The invariant Δ is asymptotically Gumbel (see Bollobás, 2001). Under H_1, the invariant $|E|$ is the convolution of two independent binomials. For an appropriate choice of $m = m(n)$, and the invariant Δ is again asymptotically Gumbel (see Rukhin, 2009, and Fig. 1 for an example).

The objective of this paper is to compare the powers $\beta_{|E|}$ and β_Δ of these two statistics under our alternative. Fig. 2 provides comparative Monte Carlo power estimates for $ER(2000,0.1)$ versus $\kappa(2000,0.1,m,s)$ for a range of m and s; for instance, when $m = 45 \approx \sqrt{2000}$ and $s = 1$, the maximum degree statistic is apparently more powerful $(\hat{\beta}_\Delta - \hat{\beta}_{|E|} \approx 0.3)$. When the number of vertices is large in our inferential setting, one often assumes that the finite-sample behavior of these statistics does not differ significantly from the limiting behavior; thus, one trusts that the limiting distribution theory provides reasonable estimates for the power for finite-sample testing. However, in our setting, we demonstrate that the asymptotic theory is misleading; our result demonstrates that while $|E|$ is asymptotically more powerful than Δ for $m = \Theta(\sqrt{n})$, the number of vertices must be astronomically large before the limiting theory agrees with finite-sample behavior.

1.1. Notation

Let $s,p \in (0,1]$ with $s > p$, $q = 1-p$ and $r = 1-s$. Let n be a positive integer, and let $m = m(n)$ be such that $m = o(n)$.

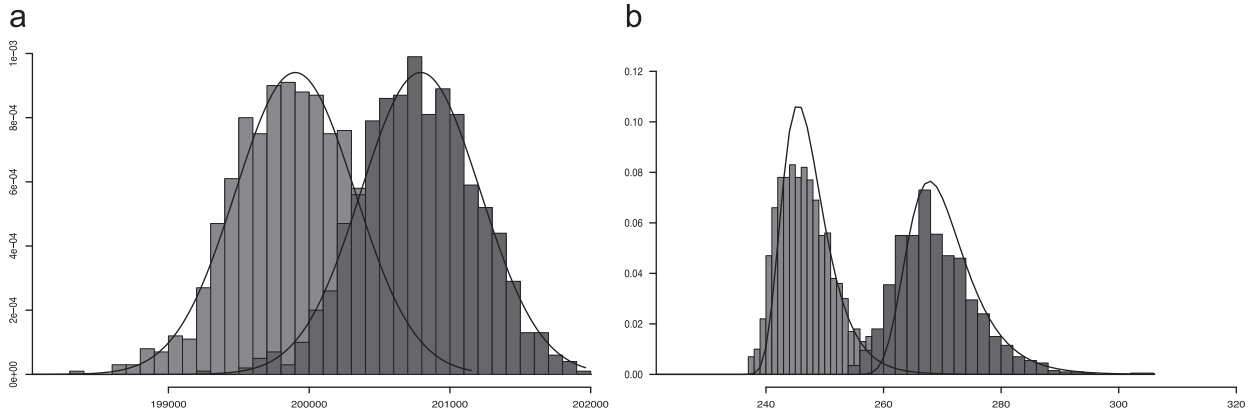

Fig. 1. Histograms from 1000 Monte Carlo replicates of both (a) the size statistic $|E|$ and (b) the maximum degree statistic Δ under H_0:$ER(2000,0.1)$ in light gray and $H_1 : \kappa(2000,0.1,45,1)$ in dark gray. The large sample density approximations from the limiting distributions are included (solid curves) for both statistics and both hypotheses (see Rukhin, 2009, for the limit theory under our hypotheses).

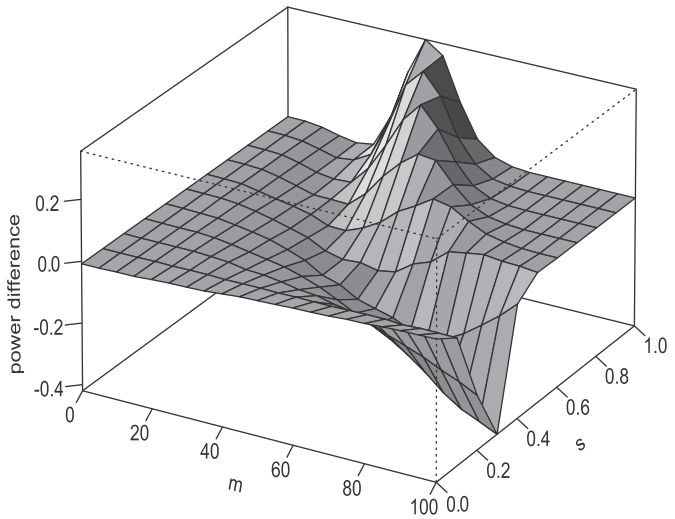

Fig. 2. Power difference surface plot $\hat{\beta}_n^{\Delta} - \hat{\beta}_n^{|E|}$ for $ER(2000,0.1)$ versus $\kappa(2000,0.1,m,s)$ at test size $\alpha = 0.05$ for a range of m and s, based on 1000 Monte Carlo replicates. We see that for large m and small s the size statistic $|E|$ is more powerful while for small m and large s (e.g., $m = 45$ and $s = 1$) the maximum degree statistic Δ is more powerful. In particular, $\hat{\beta}_n^{\Delta} - \hat{\beta}_n^{|E|} = 0.987 - 0.669 = 0.318$ at $m = 45$ and $s = 1$. (These power differences are statistically significant.)

Under hypothesis i ($i \in \{0,1\}$), let $\mathbb{P}_i(A)$ denote the probability of an event A. Also, for any random variable T, let $\mu_i = \mu_i(T)$ and $\sigma_i = \sigma_i(T)$ denote the mean and standard deviation, respectively, of T under hypothesis i. Also, we will let $c_{\alpha,n} = c^T_{\alpha,n}$ denote the α–level critical value of T (which is a function of n), and we will let $\beta^T_n = \mathbb{P}_1(T > c_{\alpha,n})$ denote the power of T. Let Z denote a random variable with the standard normal distribution.

We will use the following notation for asymptotics. Let $f = f(n)$ and $g = g(n)$ denote real-valued functions. We write

$$f = \mathrm{O}(g)$$

when there exists some N_1 and positive constant c_1 so that

$$n > N_1 \Longrightarrow 0 \le f(n) \le c_1 g(n).$$

Similarly, we will write

$$f = \Omega(n)$$

when there exists some N_2 and positive constant c_2 so that

$$n > N_2 \Longrightarrow f(n) \ge c_2 g(n) \ge 0.$$

We say $f = \Theta(g)$ if both $f = \mathrm{O}(g)$ and $f = \Omega(g)$. We will also write

$$f = \mathrm{o}(g)$$

when for any $\varepsilon > 0$, there exists N_ε so that

$$0 \le f(n) < \varepsilon g(n)$$

for all $n > N_\varepsilon$. Similarly, we will also write

$$f = \omega(g)$$

when $g = \mathrm{o}(f)$.

2. On the asymptotic power analysis of $|E|$

In this section, we analyze the asymptotic properties of $|E|$. In particular, we will show that $|E|$ possesses non-degenerate limiting power ($\lim_{n\to\infty}\beta_{|E|} \in (\alpha,1]$) when $m = \Omega(\sqrt{n})$. To this end, we have that

$$|E| = \begin{cases} X & \text{under } H_0; \\ U + W & \text{under } H_1. \end{cases}$$

where $X \sim Bin(\binom{n}{2}, p)$, $U \sim Bin(\binom{m}{2}, s)$ and $W \sim Bin(\binom{n}{2} - \binom{m}{2}, p)$. Thus,

$$\mu(|E|) = \begin{cases} \binom{n}{2} p & \text{under } H_0; \\ \binom{m}{2} s + \left[\binom{n}{2} - \binom{m}{2}\right] p & \text{under } H_1 \end{cases}$$

and

$$\sigma^2(|E|) = \begin{cases} \binom{n}{2} pq & \text{under } H_0; \\ \binom{m}{2} sr + \left[\binom{n}{2} - \binom{m}{2}\right] pq & \text{under } H_1 \end{cases}$$

We will now prove the following:

Theorem 2.1. $m(n) = \Omega(\sqrt{n}) \Longrightarrow \lim \beta_{|E|} > \alpha$.

Proof. The power of $|E|$ is computed below:

$$\beta^{|E|}_n = \mathbb{P}_1(|E| \ge c_{\alpha,n}) = \mathbb{P}_1\left(\frac{|E| - \mu_1}{\sigma_1} \ge \frac{\sigma_0}{\sigma_1}\left(\frac{c_{\alpha,n} - \mu_0}{\sigma_0}\right) + \frac{\mu_0 - \mu_1}{\sigma_1}\right).$$

As $n \to \infty$, the ratio

$$\frac{\sigma_0}{\sigma_1} = \left[\frac{pqn(n-1)}{pq(n-m)(n+m-1) + rsm(m-1)}\right]^{1/2} \to 1$$

since $m=o(n)$ and

$$\frac{c_{\alpha,n}-\mu_0}{\sigma_0}\to z_\alpha$$

where z_α is the $(1-\alpha)-$quantile of the standard normal distribution.

The expression

$$\mu_0-\mu_1=\binom{n}{2}p-\left[\binom{m}{2}s+\left(\binom{n}{2}-\binom{m}{2}\right)p\right]=-(s-p)\binom{m}{2}.$$

When $m(n)=\lambda\sqrt{n}$ for fixed $\lambda>0$, we have

$$\frac{\mu_0-\mu_1}{\sigma_1}\to-\lambda^2\left(\frac{s-p}{\sqrt{2pq}}\right).$$

We can express $|E|$ as the sum of independent Bernoulli random variables $|E|=\sum_{i<j}A_{i,j}$ where

$$A_{i,j}\sim Bern(p_{i,j})$$

and

$$p_{i,j}=\begin{cases}s & 1\le i<j\le m,\\ p & \text{otherwise.}\end{cases}\tag{1}$$

We have that $\sum_{i<j}E[|A_{i,j}|^3]/\sigma^3(|E|)\to0$ as $n\to\infty$ and according to Liapunov's Central Limit Theorem (see Chung, 1974) the limiting power of $|E|$ under this condition satisfies the inequality

$$\mathbb{P}\left(Z\ge z_\alpha-\lambda^2\left(\frac{s-p}{\sqrt{2pq}}\right)\right)>\alpha.$$

When $m=\omega(\sqrt{n})$,

$$\frac{\mu_0-\mu_1}{\sigma_1}\to-\infty$$

and $\beta_n^{|E|}\to1$. □

3. On the asymptotic distribution of Δ under H_1

Before analyzing the maximum degree invariant under the alternative, we summarize some results from Bollobás (2001). Let $y>0$ be a fixed real number, and define $x=x(n,y)$ to be the y/n-critical value approximation for the standard normal distribution:

$$x(n,y)=\sqrt{2\log n}\left(1-\frac{\log\log n+\log4\pi}{4\log n}+\frac{y}{2\log n}\right).$$

From Bollobás (2001), we have

$$\lim_{n\to\infty}\mathbb{P}_0(\Delta<pn+x(n,y)\sqrt{npq})\to e^{-e^{-y}}.$$

That is, under H_0, the maximum degree has an asymptotic Gumbel distribution.

In this section, we will show that the limiting power of the maximum degree statistic degenerates to α when $m=\Theta(\sqrt{n})$. Thus, under this condition on m, one might conclude that the size statistic is a more powerful test statistic in our inferential setting when n is "large" and $m=\Theta(\sqrt{n})$. However, for certain choices of s and p, our numerical evidence demonstrates that the opposite is true for graphs of order $n\le10^{24}$.

We begin by showing the following.

Theorem 3.1. $m=O(\sqrt{n})\Longrightarrow\lim\beta_\Delta=\alpha$.

Proof. We express $V=[m]\cup([n]\backslash[m])$ under the $\kappa(n,p,m,s)$ alternative, where $[m]$ denotes the signal component. If we let D_j denote the degree of vertex j, then

$$\mathbb{P}_1(\Delta>c_{\alpha,n})=\mathbb{P}_1\left(\bigcup_{j\in[n]}\{D_j>c_{\alpha,n}\}\right)\le\mathbb{P}_1\left(\bigcup_{j\in[m]}\{D_j>c_{\alpha,n}\}\right)+\mathbb{P}_1\left(\bigcup_{j\in[n]-[m]}\{D_j>c_{\alpha,n}\}\right)\le mQ_1+\mathbb{P}_1\left(\bigcup_{j\in[n]-[m]}\{D_j>c_{\alpha,n}\}\right)$$

where

$$Q_1 = \mathbb{P}(Bin(n-m,p)+m > c_{\alpha,n}).$$

If we define D_j to be the degree distribution of vertex j, then

$$\limsup \mathbb{P}_1\left(\bigcup_{j\in[n]-[m]}\{D_j > c_{\alpha,n}\}\right) \le \alpha$$

as the degree distributions of the $n-m$ vertices in the null component of the graph are unchanged in the κ model.

What remains to be shown is that $mQ_1 \to 0$. To this end, let $X \sim Bin(n-m,p)$ with mean $\mu_X = (n-m)p$ and variance $\sigma_X^2 = (n-m)pq$. Also, let $\gamma = -\log(-\log\alpha)$. We have that

$$mQ_1 = m\mathbb{P}(X+m > c_{\alpha,n}) = m\mathbb{P}\left(\frac{X-\mu_X}{\sigma_X} > \frac{c_{\alpha,n}-m-\mu_X}{\sigma_X}\right) \sim m(1+o(1))\mathbb{P}\left(Z > \frac{c_{\alpha,n}-m-\mu_X}{\sigma_X}\right).$$

As $m+\mu_X = pn+qm$, we have

$$c_{\alpha,n}-m-\mu_X = x(n,\gamma)\sqrt{npq}-qm.$$

Let $m = \lambda\sqrt{n}$ for some fixed $\lambda > 0$. We express the final term in this equality as

$$x(n,\gamma)\sqrt{npq}-\Theta(q\lambda\sqrt{n}) \sim \sqrt{npq}\left[x(n,\gamma)-\lambda\sqrt{\frac{q}{p}}\right].$$

As $\sigma_X = \sqrt{(n-m)pq}$,

$$\frac{c_{\alpha,n}-m-\mu_X}{\sigma_X} \sim \frac{\sqrt{npq}\left[x(n,\gamma)-\lambda\sqrt{\frac{q}{p}}\right]}{\sigma_X} \ge x(n,\gamma)-\lambda\sqrt{\frac{q}{p}}$$

and thus

$$\lim\mathbb{P}\left(Z > \frac{c_{\alpha,n}-m-\mu_X}{\sigma_X}\right) \le \lim\mathbb{P}\left(Z > x(n,\gamma)-\lambda\sqrt{\frac{q}{p}}\right).$$

Appealing to the relation

$$\mathbb{P}\{Z > y\} \sim \frac{e^{-y^2/2}}{y\sqrt{2\pi}}$$

when $y \to \infty$, we have $(e^{-x(n,\gamma)^2/2}/x(n,\gamma)\sqrt{2\pi}) \sim \gamma/n$. When $m = \lambda\sqrt{n}$, we get

$$mQ_1 \le \rho,$$

where $\rho = \rho(\lambda,p,n,\alpha) := C\exp\{x(n,\gamma)\lambda\sqrt{q/p}\}/\sqrt{n}$ for some constant C. As $x(n,\gamma) = \Theta(\sqrt{2\log n})$, elementary analysis shows that

$$\lim_{n\to\infty}\rho = 0$$

as desired. □

Combining Theorems 2.1 and 3.1, we conclude that

$$\lim\frac{\beta_\Delta}{\beta_{|E|}} < 1$$

when $m = \Theta(n)$.

The Method of Falling Moments argument given in Bollobás (2001) demonstrates the role of the function ρ when computing the power of Δ. If Y_k denotes the number of vertices with degree at least k in the graph, then Bollobás shows (under our null) that Y_k has a limiting Poisson(λ_0) distribution with

$$\lambda_0 = n\mathbb{P}(Bin(n-1,p) > k).$$

Using the above analysis, it is a relatively straightforward exercise to repeat the argument in Bollobás (2001) to deduce that, under our alternative with $m = \lambda\sqrt{n}$, the number of vertices with degree at least $c_{\alpha,n}$ has a limiting Poisson(λ_1) distribution with

$$\lambda_1 = \lim_{n\to\infty}[\rho+(n-m)\mathbb{P}(Bin(n-1,p) > c_{\alpha,n})].$$

As the event $\{\Delta < c_{\alpha,n}\}$ is equivalent to the event $\{Y_{c_{\alpha,n}} = 0\}$, it follows that a large value of ρ results in $\beta_n^\Delta \approx 1$. As an example, we have $\rho(1,0.1,n,0.05) \gg 1$ for $n \le 10^{24}$ (see Fig. 3).

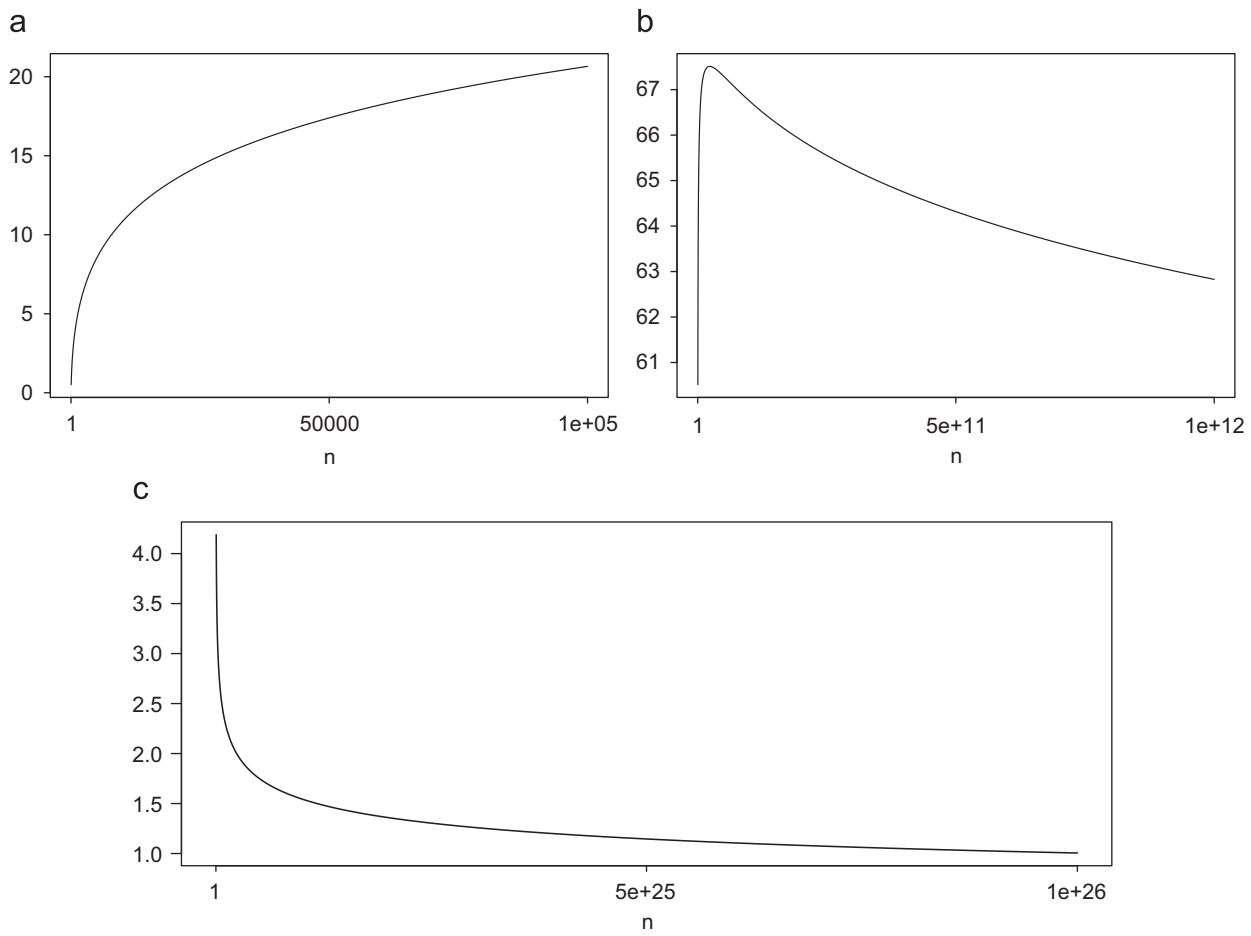

Fig. 3. Plots of $\rho(1,0.1,n,0.05)$ over various ranges of n. (a) $1 \le n \le 10^5$. (b) $1 \le n \le 10^{12}$. (c) $1 \le n \le 10^{26}$.

4. Conclusion

We have demonstrated that a comparison of test statistics based on limiting power can be misleading for graph inference. In particular, limiting results show that $|E|$ is asymptotically more powerful than Δ for $H_0 : ER(n,p)$ versus $H_1 : \kappa(n,p,m,s)$, but the opposite is in fact true for all but astronomically large graphs.

References

Bollobás, B., 2001. Random Graphs. Cambridge University Press.
Chung, K.L., 1974. A Course in Probability Theory. Academic Press, New York.
Nowicki, K., Snijders, T.A.B., 2001. Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association 96 (455), 1077.
Rukhin, A., 2009. Asymptotic analysis of various statistics for random graph inference. Ph.D. Dissertation, The Johns Hopkins University, Department of Applied Mathematics and Statistics, May.
Wasserman, S., Anderson, C., 1987. Stochastic a posteriori blockmodels: construction and assessment. Social Networks (9), 1–36.