Statistical Inference on Random Graphs: Comparative Power Analyses via Monte Carlo

Henry Pao[1], Glen A. Coppersmith[2], and Carey E. Priebe[1]

Johns Hopkins University

[1]Department of Applied Mathematics and Statistics

[2]Human Language Technology Center of Excellence

July 4, 2010

(JCGS final revision)

Abstract

We present a comparative power analysis, via Monte Carlo, of various graph invariants used as statistics for testing graph homogeneity versus a "chatter" alternative – the existence of a local region of excessive activity. Our results indicate that statistical inference on random graphs, even in a relatively simple setting, can be decidedly non-trivial. We find that none of the graph invariants considered is uniformly most powerful throughout our space of alternatives.

1 Introduction

Graphs are useful for representing a wide range of natural phenomenon. Thus, detecting anomalies within graphs may provide information relevant to making inferences for a variety of applications, such as corporate email traffic analysis (Priebe et al. 2005), examinations of turn-taking behavior (Grothendieck et al. 2008), entity extraction from text (Doddington et al. 2004), peer to peer application analysis (Sen et al. 2004), or analysis of social networks (Leenders 1995). Specifically, we are interested in being able to infer when a graph has a local region of excessive connectivity. In order to detect such changes we consider seven graph invariants: size, maximum degree, maximum average degree, scan statistic, number of triangles, clustering coefficient, and average path length. We design an inferential setting in which we evaluate the statistical power of these various graph invariants for detecting anomalies.

Our inference task is to differentiate homogeneous graphs from heterogenous graphs. Specifically, our null hypothesis (H_0) is that all vertices have the same probability of connection. The alternative hypothesis (H_A) is that there exists a subset of vertices that are (probabilistically) more highly inter-connected than the rest of the graph. In both cases, we consider the simple scenario in which all edges are mutually independent. For example, if the vertices represent the senders and recipients of email (actors) and each edge represents an email between two actors, then H_0 states that each actor communicates with each other actor with equal probability while H_A states that there is a subset of actors which share excessive email communication amongst each other – there is increased "chatter" among these vertices. See, for instance, Newman (2003) and Newman et al. (2006) for a general discussion of related applications.

Our results indicate that no invariant among those considered herein is universally most powerful for detecting increased local chatter.

1.1 Random Graphs

To model these phenomena and the effectiveness of the graph invariants for detecting anomalous behavior, we use undirected graphs $G \in \mathcal{G}_n$, the collection of all graphs on the n vertices $V = \{1, ..., n\}$. We denote the vertex set $V = V(G)$ and edge set $E = E(G)$; thus $G = (V, E)$. To denote edges in E, we use the notation e_{uv} for $u, v \in V$ (it is said that vertices u and v are adjacent). We will not consider weighted or parallel edges. We will not consider loops, so if $e_{uv} \in E$ then $u \neq v$. Our graphs are undirected, so there is no distinction between e_{uv} and e_{vu}.

1.2 Null Hypothesis

Our null hypothesis (H_0) is that the observed graph is drawn from an Erdős-Rényi (ER) random graph model (Bollobás 2001); that is, each of the $\binom{n}{2}$ possible edges exists independently with a given probability

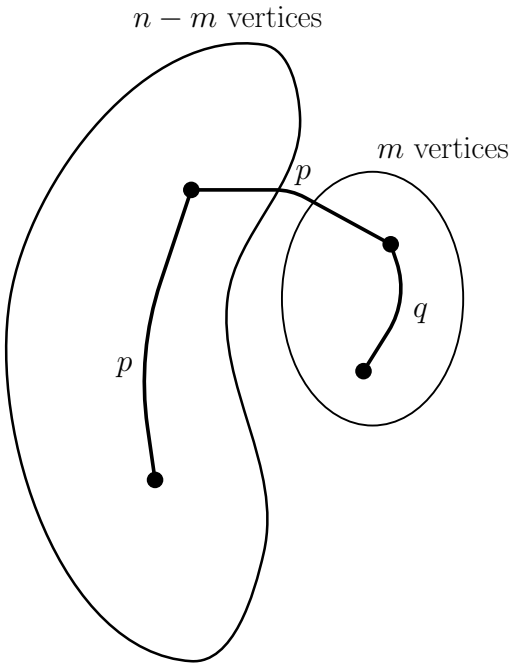

Figure 1: H_A: The "kidney and egg" graph, $\kappa(n, p, m, q)$. The small "egg" represents the m vertices (V_A) that exhibit chatter (each edge occurring with probability q). The "kidney" is the population of $n - m$ vertices which are not exhibiting chatter (each edge occurring with probability $p < q$). Edges between a vertex in the kidney and a vertex in the egg occur with probability p.

$p \in [0, 1)$. Again, $V = \{1, ..., n\}$, so H_0: $ER(n, p)$.

1.3 Alternative Hypothesis

Our alternative hypothesis (H_A) – local chatter – is the κ random graph model, $\kappa(n, p, m, q)$. Again, $V = \{1, ..., n\}$. A subset of m vertices ($V_A \subset V$, $|V_A| = m$, $m \in \{2, \ldots, n\}$) are connected with probability q where $q > p$. The remaining $n - m$ vertices are connected with probability p, just like the entire graph under H_0, to represent the portion of the population not "chattering". Edges between a vertex in V_A and a vertex in $V \setminus V_A$ occur with probability p. Again, all edges are independent of one another. We parameterize our alternative by $\theta \in \Theta_A = \{2, \ldots, n\} \times (p, 1]$. This κ graph is referred to as the "kidney and egg" graph as depicted in Figure 1. So, H_A: $\kappa(n, p, m, q)$ with $(m, q) \in \Theta_A$.

Our comparative power investigation consists of quantifying the ability of various graph invariants to distinguish a homogeneous $ER(n, p)$ graph from our local "chatter" alternative $\kappa(n, p, m, q)$.

1.4 Graph Preliminaries

Graph theory preliminaries are available in many textbooks; see, for instance, West (2001). We present here the basics required for our analysis.

1.4.1 Size and Order

The size of a graph G, denoted $\text{size}\,(G) = |E(G)|$, is the number of edges. Likewise, the order of a graph G, denoted $\text{order}\,(G) = |V(G)|$, is the number of vertices.

1.4.2 Distance

The distance between any two vertices is measured by the minimum number of edges required to traverse between them. For vertices u and v, this is denoted by $l(u, v)$. If $e_{uv} \in E(G)$ then $l(u, v) = 1$; $l(u, u) = 0$ for all u; and if no path exists between u and v then $l(u, v) = \infty$.

1.4.3 Degree

The degree of vertex v, denoted $d(v)$, is the number of edges incident to v. Since we allow only a single edge between two vertices and no loops, $d(v)$ is also the number of vertices connected to v.

1.4.4 Adjacency Matrix

The adjacency matrix $A = A(G)$ is an $n \times n$ symmetric, hollow (zeros on the diagonal), binary matrix where $n = \text{order}\,(G)$. If a_{uv} denotes the $(u, v)^{th}$ element of A, then $a_{uv} = 1$ if and only if $e_{uv} \in E(G)$.

1.4.5 Induced Subgraphs

Given a collection of vertices $V' \subset V$, the induced subgraph $\Omega(V'; G)$ is defined to be the graph (V', E') where $e_{uv} \in E'$ if and only if $u \in V'$, $v \in V'$ and $e_{uv} \in E$. In essence, $\Omega(V'; G)$ is the collection of V' vertices and the edges from G connecting any pair of those vertices.

1.4.6 Neighborhoods

To study local activity of a graph, we use neighborhoods to provide a notion of locality. The k^{th} order neighborhood of $v \in V$ is defined as $N_k[v; G] = \{u \in V(G) : l(v, u) \leq k\}$.

2 Graph Invariants

We examine the power of seven graph invariants, acting as test statistics, to detect excessive local activity. The statistics we examine are: size, maximum degree, maximum average degree (both via a greedy approxi-

mation and via an eigenvalue approximation), scan statistic, number of triangles, clustering coefficient, and average path length. In all cases, a large value of the invariant is evidence in favor of H_A.

2.1 Definitions

2.1.1 Size

The size of a graph is the number of edges in the graph, given by

$$\text{size}\,(G) = |E(G)|. \tag{1}$$

This simplest graph invariant is a global measure of activity; as such, it would not be expected to have good power characteristics against $\kappa(n, p, m, q)$ for small values of m.

2.1.2 Maximum Degree

The maximum degree $\delta(G)$ is given by

$$\delta(G) = \max_{v \in V} d(v) \tag{2}$$

and is the simplest local graph invariant.

2.1.3 Maximum Average Degree

The maximum average degree over all subgraphs of G is denoted $\text{MAD}(G)$. If $d(v)$ is the degree of vertex v, then the average degree of a graph G is given by

$$\bar{d}(G) = \frac{1}{|V|} \sum_{v \in V} d(v) = \frac{2\,\text{size}(G)}{\text{order}(G)}. \tag{3}$$

Thus the maximum average degree invariant is given by

$$\text{MAD}(G) = \max_{\Omega \subset G} \bar{d}(\Omega) \tag{4}$$

where the maximum is over all induced subgraphs of G. (Notice that it suffices to consider only induced subgraphs, since any subgraph with fewer edges than its related induced subgraph will have a lower average degree.)

Since this invariant is difficult to compute exactly, we resort to consideration of two approximations to the maximum average degree, $\text{MAD}_g(G)$ and $\text{MAD}_e(G)$, such that

$$\text{MAD}_g(G) \leq \text{MAD}(G) \leq \text{MAD}_e(G). \tag{5}$$

Greedy MAD

We consider a primitive greedy algorithm $\text{MAD}_g(G)$ to estimate the maximum average degree of a graph. The algorithm iteratively removes a vertex with the smallest degree and calculates the average degree of the remaining induced subgraph. After removing all vertices, the largest average degree encountered is returned. This provides an approximation for the maximum average degree that is easy to implement (Ullman and Scheinerman 1997) and is a lower bound for $\text{MAD}(G)$.

Maximum Eigenvalue MAD

$\text{MAD}(G)$ is bounded above by the largest eigenvalue of the adjacency matrix, denoted $\text{MAD}_e(G)$. As the Rayleigh-Ritz Theorem (Horn and Johnson 1985) states, if A is Hermitian, $\lambda_{\max}$ is the largest eigenvalue, and $\lambda_{\min}$ is the smallest eigenvalue, then

$$\lambda_{\min} x^T x \leq x^T A x \leq \lambda_{\max} x^T x \text{ for all } x \in \mathbb{R}^n, \tag{6}$$

$$\lambda_{\max} = \max_{x \neq 0} \frac{x^T A x}{x^T x}, \tag{7}$$

$$\lambda_{\min} = \min_{x \neq 0} \frac{x^T A x}{x^T x}. \tag{8}$$

Now let $A = [a_{ij}]$ be the adjacency matrix for graph G, and consider $x = [x_i]$ to be any nonzero binary vector. Then $\frac{x^T A x}{x^T x}$ is the average degree of an induced subgraph, where the i^{th} vertex is present if the i^{th} element of x is 1 ($x_i = 1$). Note that MAD is also of this form, implying

$$\text{MAD}(G) \leq \lambda_{\max}, \tag{9}$$

so $\text{MAD}_e(G) \equiv \lambda_{\max} \geq \text{MAD}(G)$.

MAD **Comparison**

When comparing the power of MAD_g and MAD_e using the inference problem described in Section 3, the eigenvalue method appears to be strictly better at detecting increased local activity than the greedy method, as evaluated by Monte Carlo experiments. Henceforth, we shall consider only MAD_e and disregard MAD_g. Figure 2 shows the superior performance on our inference problem of MAD_e for $n = 1000$, $p = 0.1$.

2.1.4 Scan Statistic

Scan statistics (Priebe et al. 2005) are graph invariants based on local neighborhoods of the graph. We will consider the scan statistic $S_k(G)$ to be the maximum number of edges over all k^{th} order neighborhoods. We

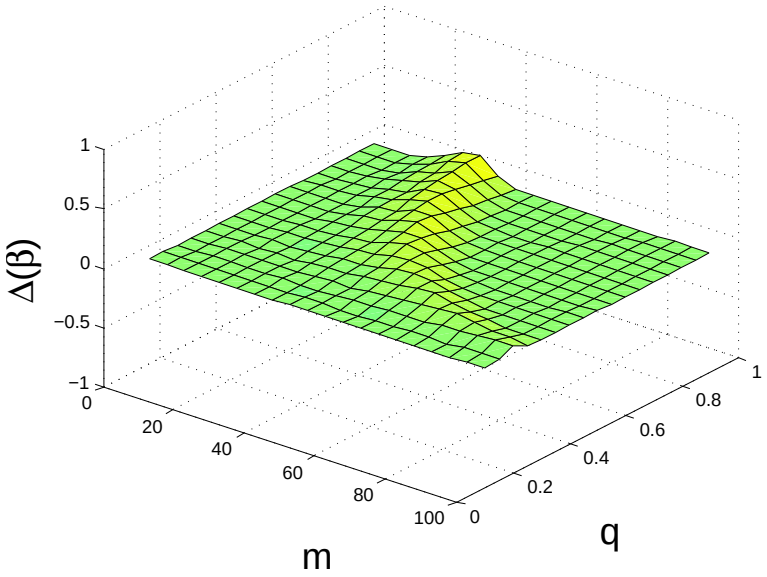

Figure 2: Statistical power difference surface for $\text{MAD}_e - \text{MAD}_g$ with $n = 1000$ and $p = 0.1$ over a range of $(m, q) \in \Theta_A$, via Monte Carlo. MAD_e dominates. (See Section 3 for details of the inference problem and Monte Carlo experiment. This surface is based on Figure 10 (c) vs. (d).)

will consider only $k = 1$, so $S_1(G)$ is given by

$$S_1(G) = \max_{v \in V} \text{size}(\Omega(N_1[v; G])). \tag{10}$$

Notice that $S_1(G)$ considers only a subset of cardinality n of all induced neighborhoods. The locality statistic $\text{size}(\Omega(N_1[v; G]))$ is an extension of degree $d(v)$ which also counts cross-talk among a vertex's neighbors, which suggests the scan statistic S_1 as more appropriate than maximum degree δ for our local chatter alternative; see Rukhin and Priebe (2009) for an analytic investigation of this claim via asymptotics. For $k > 1$, $S_k(G)$ is conjectured to be valuable against more elaborate alternatives than the H_A considered herein, but will not be considered further in this paper.

2.1.5 Number of Triangles

We consider the total number of triangles in G. If A is the adjacency matrix for graph G, then the number of triangles is given by

$$\tau(G) = \frac{\text{trace}(A^3)}{6}. \tag{11}$$

(The v^{th} diagonal element of A^3 counts the number of paths of length k from v back to itself. This counts triangles – in fact, it over-counts by a factor of six, since each triangle has three vertices and each vertex can traverse the triangle-path two different ways.)

2.1.6 Clustering Coefficient

We consider the global clustering coefficient in G. Consider all induced subgraphs H of G with order$(H) = 3$ and size$(H) \geq 2$; each such subgraph is either a triangle or an angle. Let angles(G) be the total number of (non-triangle) angle induced subgraphs in G. Then the global clustering coefficient is given by

$$\text{CC} = \frac{\tau(G)}{\tau(G) + \text{angles}(G)}. \tag{12}$$

2.1.7 Average Path Length

We consider the average path length in G. We define

$$\text{APL} = -\frac{\sum_{u,v} l(u,v)}{n(n-1)}. \tag{13}$$

The negative sign in this definition allows *large* values of the invariant to provide evidence in favor of H_A, for compatibility with all the other invariants under consideration. If no path exists between u and v, we use $l(u,v) = 2 \max l(u,v)$, where the maximum is taken over all pairs of vertices that have an existing path between them. (Generally, the distance between two nodes for which no path exists is defined as ∞; this modification is necessary to make the average path length a meaningful test statistic in (possibly) disconnected graphs.)

2.2 Distributions

2.2.1 Size

Under $ER(n,p)$, size$(G) \sim$ Binomial $\left(\binom{n}{2}, p\right)$ so

$$P[\text{size}(G) = x] = \binom{\binom{n}{2}}{x} p^x (1-p)^{\binom{n}{2}-x} \tag{14}$$

for $x = 0, 1, \ldots, \binom{n}{2}$.

Our modified $\kappa(n,p,m,q)$ also has a probability mass function that is readily available: size(κ) is the sum of independent binomials, Binomial$(\binom{m}{2}, q)$ for the egg and Binomial$(\binom{n}{2} - \binom{m}{2}, p)$ for the rest of the graph. Thus

$$P[\text{size}(\kappa) = x] = \sum_{i=0}^{x} \left(\binom{\binom{m}{2}}{x-i} q^{x-i} (1-q)^{\binom{m}{2}-x+i} \binom{\binom{n}{2} - \binom{m}{2}}{i} p^{i} (1-p)^{\binom{n}{2}-\binom{m}{2}-i} \right) \tag{15}$$

for $x = 0, 1, \ldots, \binom{n}{2}$.

Limiting distributions based on normal approximation are also readily available.

Figure 3 presents Monte Carlo simulation results for size(G) for $H_0 : ER(n = 1000, p = 0.1)$ vs. $H_A : \kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$.

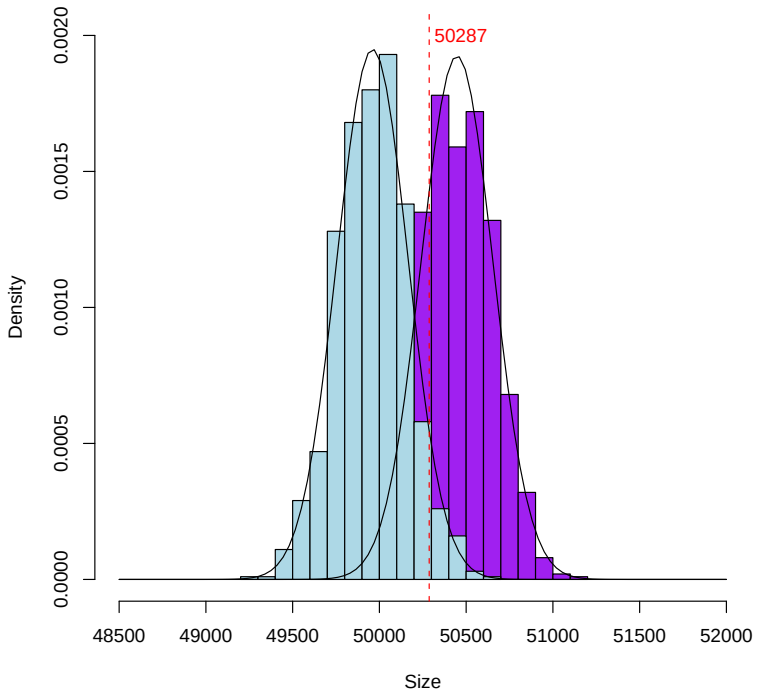

Figure 3: Monte Carlo simulation ($R = 1000$ replicates) for size(G) for Erdős-Rényi ($n = 1000, p = 0.1$) in blue and $\kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$ in purple, with their theoretical probability mass functions overlayed. The critical value is denoted by the vertical dotted line ($\alpha = 0.05$). The Monte Carlo power estimate is $\hat{\beta} = 0.775$. Exact calculation shows that the true power for this case is $\beta = 0.780$.

2.2.2 Maximum Degree

The exact probability mass function of the maximum degree $\delta(G)$ of an Erdős-Rényi random graph G is not available. However there is a limit result with $n \to \infty$. The limiting distribution is Gumbel (Bollobás 2001);

$$a = pn + \sqrt{2p(1-p)n \log n}\left(1 - \frac{\log\log n}{4\log n} - \frac{\log(2\sqrt{\pi})}{2\log n}\right), \tag{16}$$

$$b = \frac{\sqrt{2p(1-p)n\log n}}{2\log n}, \tag{17}$$

$$f_{\delta(G)}(d) \to \frac{1}{b}\exp\left[\frac{d-a}{b} - \exp\left(\frac{d-a}{b}\right)\right]. \tag{18}$$

A Gumbel approximation is also available under H_A (Rukhin 2008).

Figure 4 presents Monte Carlo simulation results for $\delta(G)$ for $H_0 : ER(n = 1000, p = 0.1)$ vs. $H_A : \kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$.

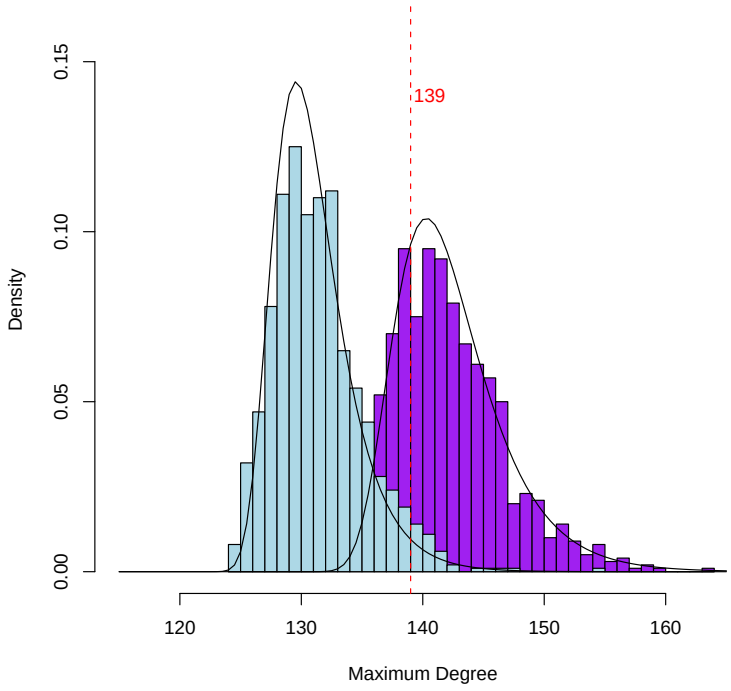

Figure 4: Monte Carlo simulation ($R = 1000$ replicates) for maximum degree $\delta(G)$ for Erdős-Rényi ($n = 1000$, $p = 0.1$) in blue, and $\kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$ random graphs in purple, with the theoretical null Gumbel probability density function overlayed. The critical value is denoted by the vertical dotted line ($\alpha = 0.05$). The Monte Carlo power estimate is $\hat{\beta} = 0.793$. Exact calculations shows that the true power for this case is $\beta = 0.715$.

2.2.3 Maximum Average Degree

No approximations are currently available for $\text{MAD}_e(G)$ under either H_0 or H_A.

Figure 5 presents Monte Carlo simulation results for $\text{MAD}_e(G)$ for $H_0 : ER(n = 1000, p = 0.1)$ vs. $H_A : \kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$.

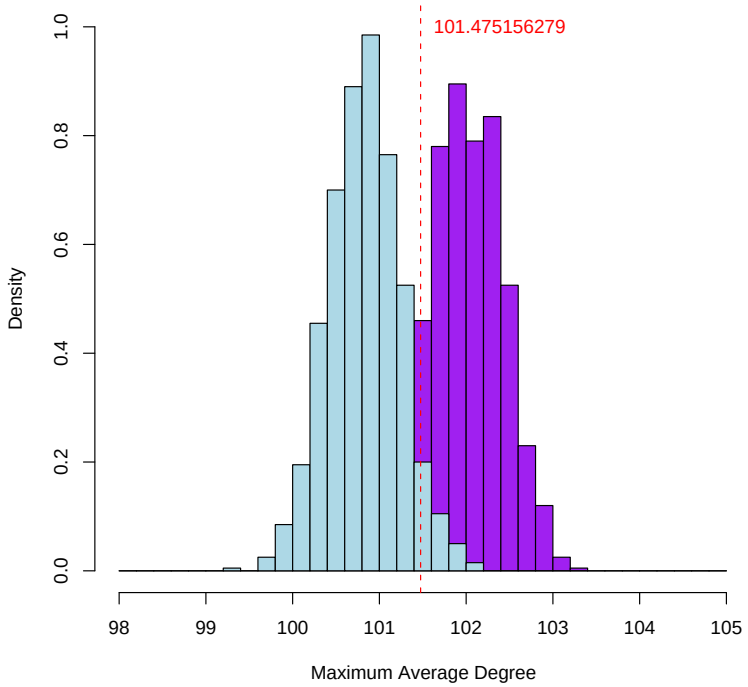

Figure 5: Monte Carlo simulation ($R = 1000$ replicates) for maximum average degree $\text{MAD}_e(G)$ for Erdős-Rényi ($n = 1000$, $p = 0.1$) in blue, and $\kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$ random graphs in purple. The critical value is denoted by the vertical dotted line ($\alpha = 0.05$). The Monte Carlo power estimate is $\hat{\beta} = 0.909$.

2.2.4 Scan Statistic

As with maximum degree, there is a limiting Gumbel approximation for our scan statistic for $H_0 : ER(n, p)$ (Rukhin 2008);

$$a = \frac{1}{2}p^3n^2 + p^2\sqrt{p(1-p)n^3}\sqrt{2\log n}\left(1 - \frac{\log\log n - \log(4\pi^2)}{4\log n}\right), \tag{19}$$

$$b = \frac{p^2\sqrt{p(1-p)n^3}}{\sqrt{2\log(n)}}, \tag{20}$$

$$f_S(s) \to \frac{1}{b}\exp\left[\frac{s-a}{b} - \exp\left(\frac{s-a}{b}\right)\right]. \tag{21}$$

A Gumbel approximation is also available under H_A (Rukhin 2008).

Figure 6 presents Monte Carlo simulation results for $S_1(G)$ for $H_0 : ER(n = 1000, p = 0.1)$ vs. $H_A : \kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$.

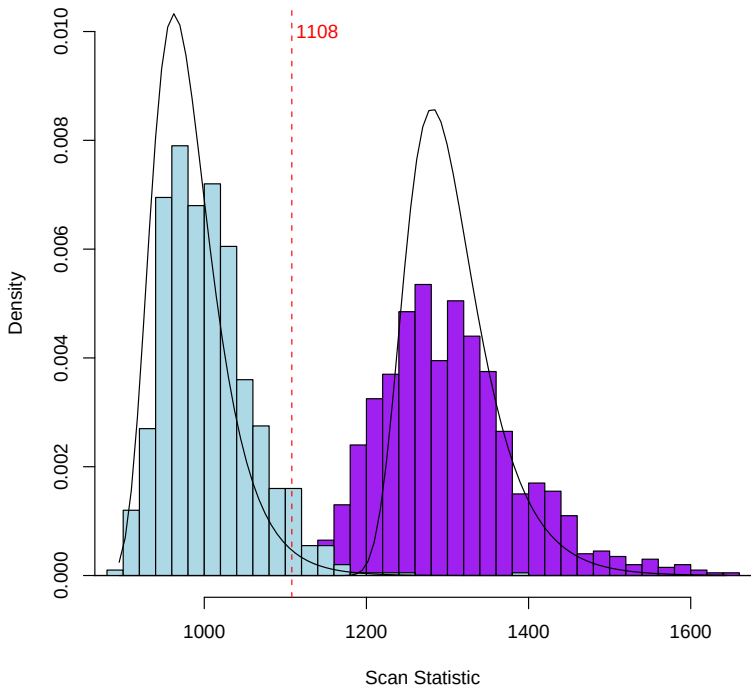

Figure 6: Monte Carlo simulation ($R = 1000$ replicates) for scan statistic $S_1(G)$ for Erdős-Rényi ($n = 1000$, $p = 0.1$) in blue and $\kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$ random graphs in purple, with the theoretical null Gumbel probability density function overlayed. The critical value is denoted by the vertical dotted line ($\alpha = 0.05$). The Monte Carlo power estimate is $\hat{\beta} = 0.999$. Exact calculations show that the true power for this case is $\beta = 1.0$.

2.2.5 Number of Triangles

Under both H_0 (Nowicki and Wierman 1988) and H_A (Rukhin 2008) a U-statistic approach demonstrates that $\tau(G)$ (properly normalized) is asymptotically normal.

Figure 7 presents Monte Carlo simulation results for $\tau(G)$ for $H_0 : ER(n = 1000, p = 0.1)$ vs. $H_A : \kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$.

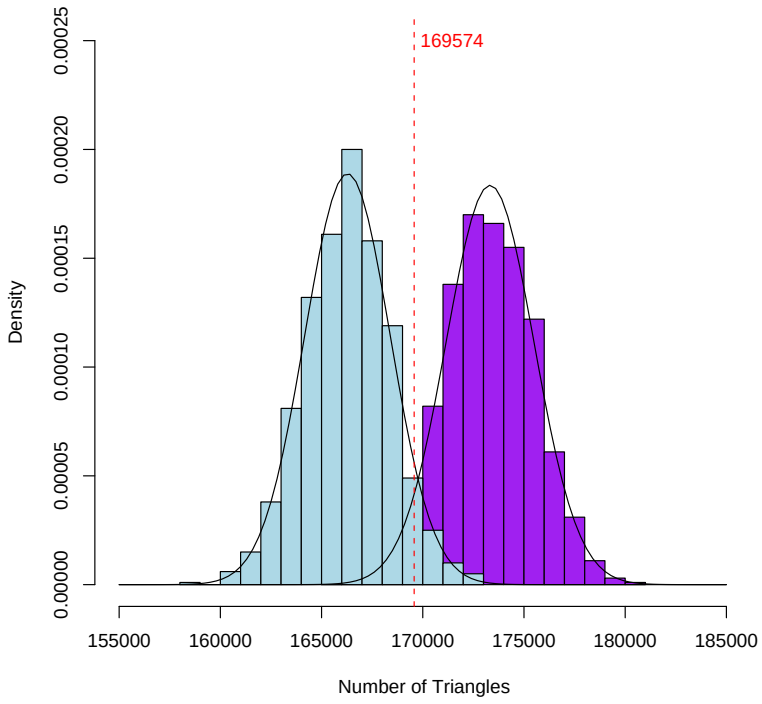

Figure 7: Monte Carlo simulation ($R = 1000$ replicates) for number of triangles $\tau(G)$ for Erdős-Rényi ($n = 1000$, $p = 0.1$) in blue and $\kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$ random graphs in purple with theoretical null and alternate normal probability density functions overlayed. The critical value is denoted by the vertical dotted line ($\alpha = 0.05$). The Monte Carlo power estimate is $\hat{\beta} = 0.962$. Exact calculations show that the true power for this case is $\beta = 0.958$.

2.2.6 Clustering Coefficient

For the clustering coefficient, $E[\mathrm{CC}(G)] = p$ under $H_0 : ER(n, p)$, since edges are independent; $P[\mathrm{size}(H) = 3 | \mathrm{size}(H) \geq 2, \mathrm{order}(H) = 3] = p$. Indeed, the ratio $\tau(G)/(\tau(G) + \mathrm{angles}(G))$ is asymptotically normal under H_0. Under H_A, one obtains a convolution of normals by considering induced subgraphs H of G with $\mathrm{order}(H) = 3$ and $\mathrm{size}(H) \geq 2$ conditionally, based on zero, one, two, or three of the vertices being among the m anomalous vertices V_A.

Figure 8 presents Monte Carlo simulation results for CC(G) for $H_0 : ER(n = 1000, p = 0.1)$ vs. $H_A : \kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$.

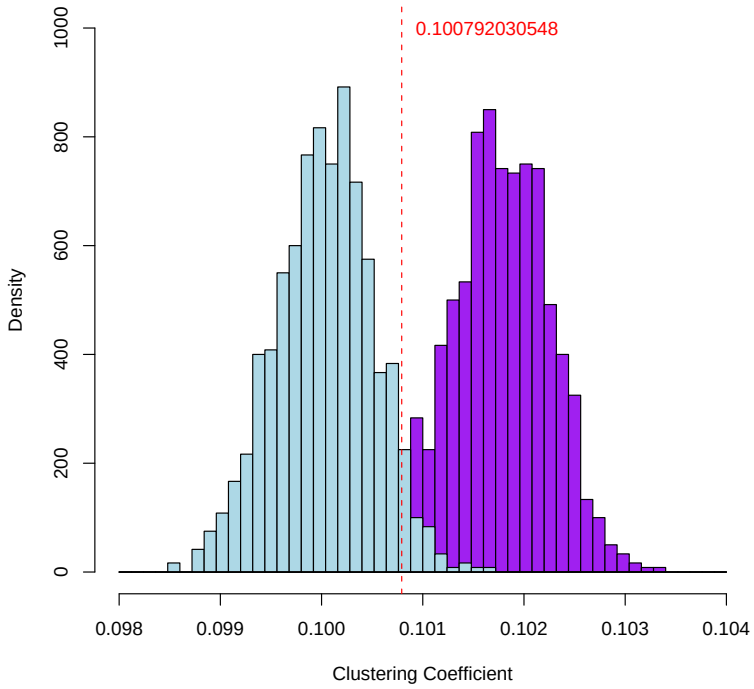

Figure 8: Monte Carlo simulation ($R = 1000$ replicates) for clustering coefficient CC(G) for Erdős-Rényi ($n = 1000$, $p = 0.1$) in blue and $\kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$ random graphs in purple. The critical value is denoted by the vertical dotted line ($\alpha = 0.05$). The Monte Carlo power estimate is $\hat{\beta} = 0.986$.

2.2.7 Average Path Length

As $n \to \infty$, the probability that an $ER(n, p)$ graph is connected goes to unity and asymptotic distributions for APL are available via consideration of sums of dependent random variables. However, for $n = 1000$, for example, the non-trivial probability that the graph is not connected implies that the altered definition for distance $l(u, v)$ when no path exists comes into play, complicating matters. We make the conjecture that APL, properly normalized, is approximately normal even for moderate n.

Figure 9 presents Monte Carlo simulation results for APL(G) for $H_0 : ER(n = 1000, p = 0.1)$ vs. $H_A : \kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$.

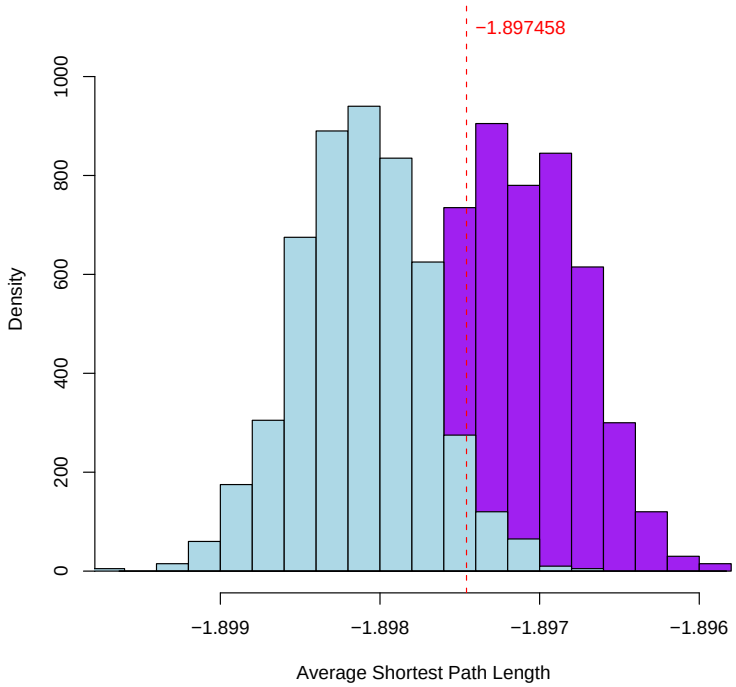

Figure 9: Monte Carlo simulation ($R = 1000$ replicates) for average path length APL(G) for Erdős-Rényi ($n = 1000$, $p = 0.1$) in blue and $\kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$ random graphs in purple. The critical value is denoted by the vertical dotted line ($\alpha = 0.05$). The Monte Carlo power estimate is $\hat{\beta} = 0.770$.

3 Experimental Design

Figures 3 – 9 present Monte Carlo power results for our seven invariants for $H_0 : ER(n = 1000, p = 0.1)$ vs. $H_A : \kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$. We proceed now to design and execute a Monte Carlo experiment to generate comparative power results over Θ_A.

3.1 Why Monte Carlo?

Distributions for random graph invariants are notoriously difficult to obtain – even for the (conceptually) simple invariants considered herein. Exact finite-sample distributions are unavailable for most invariants under H_A (and even under H_0) for all but extremely small $n = \text{order}(G)$. (And this is just for simple, mutually independent edge case we consider herein; these difficulties are compounded for generalizations to more elaborate random graph models.) In addition, it has been demonstrated (Rukhin and Priebe 2009) that power comparisons based on limiting distributions can be misleading. For these reasons, Monte Carlo is one of the few tools available for comparative power analysis. While it is true that our Monte

Carlo investigations provide only snapshots into test behavior, these snapshots provide new and valuable understanding for statistical inference on random graphs.

3.2 Monte Carlo Design

Our Monte Carlo experiment is performed with 1000 vertices ($|V| = n = 1000$), and the null probability of an edge between any pair of vertices is $p \in [0, 1]$. We fix $p = 0.1$ for the experiments presented here; that is, $H_0 : ER(1000, 0.1)$. A representative collection of $(m, q) \in \Theta_A$ is considered: $m \in \{5, 10, 15, \cdots, 100\}$ and $q \in \{0.10, 0.15, 0.20, \cdots, 0.90\}$. (For $q = p = 0.1$, H_0 holds.) For each specified (m, q) we perform $R = 1000$ Monte Carlo replicates, yielding statistical power estimates for each invariant. The result is comparative power function estimates across Θ_A, as shown in Figures 10 and 13.

3.2.1 Type I Error

To gauge the utility of the various graph invariants under consideration, we estimate the statistical power β – the probability of detecting an increase in local activity at a given test size α – for each invariant. The power of a test is the probability of rejecting the null hypothesis when it is false, which is easy to estimate using Monte Carlo methods. If $T(G)$ is the graph invariant of interest calculated from observed graph G, we first generate R independent, identically distributed (i.i.d.) graphs $G_1, G_2, \ldots, G_R$ under H_0. We calculate $T_r = T(G_r), r = 1, \ldots, R$, and consider the order statistics $T_{(1)} \leq \cdots \leq T_{(R)}$. H_0 is rejected when $T(G) > T_{(R(1-\alpha))}$, yielding a test that is approximately size α for R large (Bickel and Doksum 2001).

3.2.2 Power

If $T(\kappa_r)$ are the statistics generated based on graphs generated under the alternative hypothesis, $r = 1, \ldots, R$, then the estimated power $\hat{\beta}$ is given by

$$\hat{\beta} = \frac{1}{R} \sum_{r=1}^{R} I\{T(\kappa_r) > T_{(R(1-\alpha))}\}. \tag{22}$$

We use test size $\alpha = 0.05$ for all experiments. When $q = p = 0.1$, G is homogenous, as under H_0, and the power is $\beta = \alpha$ for all invariants.

3.2.3 Randomization

Since our statistics (graph invariants) are discrete random variables, we compensate for ties in the Monte Carlo tests via randomization (Bickel and Doksum 2001). To account for the case $T(\kappa_r) = T_{(R(1-\alpha))}$, a percentage of these are rejected in calculating $\hat{\beta}$.

Specifically, the null hypothesis is rejected not only when $T(\kappa_r) > T_{(R(1-\alpha))}$ but also, probabilistically, when $T(\kappa_r) = T_{(R(1-\alpha))}$. The quantity

$$\frac{\alpha - \hat{\alpha}_d}{\frac{1}{R}\sum_r I\{T(G_r) = T_{(R(1-\alpha))}\}} \tag{23}$$

is the randomization probability. The nominal size without randomization is denoted here by

$$\hat{\alpha}_d = \frac{1}{R}\sum_r I\{T(G_r) > T_{(R(1-\alpha))}\}.$$

4 Experimental Results

Notice that Figures 3 – 9 together demonstrate that for $H_0 : ER(n = 1000, p = 0.1)$ vs. $H_A : \kappa(n = 1000, p = 0.1, m = 50, q = 0.5)$ the invariant $S_1(G)$ is the most powerful statistic among those under consideration: Monte Carlo power estimates yield $\hat{\beta}_{S_1(G)} = 0.999 > \max\{\hat{\beta}_{\text{size}(G)} = 0.775, \hat{\beta}_{\delta(G)} = 0.793, \hat{\beta}_{\text{MAD}_e(G)} = 0.909, \hat{\beta}_{\tau(G)} = 0.962, \hat{\beta}_{\text{CC}(G)} = 0.986, \hat{\beta}_{\text{APL}(G)} = 0.770\}$. (The scan statistic's superiority is statistically significant, with $R = 1000$; paired analysis provides much stronger significance.) In this section we generalize this point investigation to a comparative power analysis over Θ_A.

4.1 Power Surface Plots

In order to determine the values of $(m, q) \in \Theta_A$ for which the invariants are effective for detecting increased local activity, we test over the previously specified range of values for m and q ($m \in \{5, 10, 15, \ldots, 100\}$, $q \in \{0.10, 0.15, 0.20, \ldots, 0.90\}$) with $\alpha = 0.05$, using $R = 1000$ Monte Carlo replicates each. The collection of Monte Carlo runs provide the data used to generate the statistical power surface plots shown in Figure 10.

The power surface plots for the invariants are superficially similar to one another, which makes determining the relative effectiveness of each invariant difficult. Powers for all invariants range from $\beta \approx \alpha$ for small m or q to $\beta \approx 1$ for large m and q. Substantial differences exist, but may not be apparent, between the various invariants for moderate m, q.

4.2 Power Difference Plots

In order to better analyze the comparative statistical power of our invariants, we examine the difference between pairs of statistical power surfaces. These plots allow comparative power analyses across Θ_A.

For many pairs of invariants, there exists a range of m and q for which each of the two invariants has a higher power; neither invariant has greater power over all of Θ_A. However, Figure 11 demonstrates that $S_1(G)$ dominates $\delta(G)$, rendering the latter "inadmissible". (This inadmissibility claim is supported by the

Monte Carlo results of Figure 11 for $H_0 : ER(n = 1000, p = 0.1)$ vs. $H_A : \kappa(n = 1000, p = 0.1, m, q)$ only. However, the scan statistic is specifically designed to improve upon maximum degree for "chatter" alternatives of the type represented by our κ random graph model, and we conjecture that this domination holds more generally. An asymptotic version of this result is available (Rukhin and Priebe 2009).) Figure 12 demonstrates, again for $H_0 : ER(n = 1000, p = 0.1)$ vs. $H_A : \kappa(n = 1000, p = 0.1, m, q)$, that size$(G)$ and APL(G) are indistinguishable in terms of power over all of Θ_A.

The comparative power surface plots shown in Figure 13 provide pairwise comparison of the remaining four invariants, excluding the inadmissible $\delta(G)$ and including size(G) but not the then-superfluous APL(G). Since powers are approximately α for small m or q and approximately 1 for large m and q for all invariants, power differences are approximately 0 for these extremes. Substantial differences are readily apparent between the various invariants for moderate m and q in these comparative power surfaces. No invariant dominates any other over all of Θ_A. Figure 14 presents in more detail the interesting case of $S_1(G)$ vs. $\tau(G)$. When m is large and q is small $\hat{\beta}_{\tau(G)} > \hat{\beta}_{S_1(G)}$, while when m is small and q is large $\hat{\beta}_{S_1(G)} > \hat{\beta}_{\tau(G)}$, and the power differences are large in both cases. Neither invariant dominates the other throughout Θ_A.

4.3 Most Powerful Statistic

When the power of all the invariants are examined together, the range of values for m and q for which each has the greatest power is shown in Figure 15a. Only $S_1(G)$, CC(G), and $\tau(G)$ show best power at least somewhere in Θ_A. The scan statistic $S_1(G)$ is best for moderate m, q; CC(G) dominates when m is small and q is large; and $\tau(G)$ dominates for large m and small q.

We have presented detailed results for $(m, q) \in \Theta_A$, but for just one choice of n, p – $n = 1000,\ p = 0.1$. Figures 15b and 15c present "most powerful statistic over Θ_A", analogous to Figure 15a, for other choices of n, p. In the case $n = 100, p = 0.1$ we see $S_1(G)$ as dominate for the large q small m region and MAD$_e(G)$ as dominate for the small q large m region. (Notice that we consider m much larger as a percentage of n in Figures 15b and 15c.) For the case $n = 100, p = 0.4$ the clustering coefficient CC(G) dominates. Some of the smattering effects in these figures is due to artifacts of the Monte Carlo; the basic structure of the plots is real. The fundamental result is that there does not necessarily exist a single *uniformly* most powerful statistic, across all of Θ_A.

We have performed extensive Monte Carlo Analysis generalizing Figure 15 for numerous n, p cases. The suggestive results seen in Figure 15 seem to hold generally: the scan statistic and the clustering coefficient are often most powerful, and $\tau(G)$ and MAD$_e(G)$ occasionally come into play; rarely if ever are the other invariants recommended.

5 Conclusions

Analytics are preferable to Monte Carlo. However, finite-sample comparative power analytics for random graphs are challenging even in this relatively simple setting, and asymptotic analytics can be at odds with finite sample truths even for extraordinarily large n (Rukhin and Priebe 2009). The snapshots into comparative test behavior available via Monte Carlo analysis provide new and valuable understanding for statistical inference on random graphs, and will form the foundation for comparative power investigations for larger graphs and more complex models.

From the statistical power surface plots, we observe that all the graph invariants we examined have power $\beta \approx 1$ for large m and q and power $\beta \approx \alpha$ for small m or q, as expected. For moderate m and q, the comparative behavior of the various invariants is quite complicated; different invariants dominate in different regions. In particular, there is in general a ridge/trough phenomenon running (nonlinearly) from large q small m to small q large m which seems to differentiate invariants – some invariants are recommended in the small q large m region and others are recommended in the large q small m region. There does not exist a uniformly most powerful statistic across all of Θ_A. That is, the specific alternative – how many anomalous vertices (m) and by how much are they anomalous (q) – determines the most powerful statistic. If a recommendation is required, without knowledge of the specific alternative m, q, we suggest (based on Figure 13 and related results) using scan statistic and clustering coefficient together, since the best of those two is rarely out-performed by much but there exists regions of Θ_A where each out-performs the other substantially. This requires two tests, and multiple-testing correction, but if no information is available regarding m and q then this seems a good course of action.

The "statistical inference on random graphs" considered herein is hypothesis testing. Of course, once one rejects in favor of $\kappa(n, p, m, q)$, the question of estimating m and q, as well as identifying the anomalous vertex set V_A, naturally arises. The more general inferential tasks are of substantial interest, but involve complicating issues best addressed after gaining solid understanding from our simpler comparative power analysis. For instance: we have treated p as known throughout this manuscript. Treating p as unknown both is more realistic and presents a confounding issue. Consideration of estimating p under H_A begins with assuming the anomalous m vertices are known – that is, we know the set V_A. Then

$$\widehat{p} = \text{size}(\Omega(V \setminus V_A))/\binom{n-m}{2}.$$

Thus the problem is to find V_A. One approach is to let $\widehat{V}_A = N[v^*]$ – that is, the vertex of maximum degree or maximum locality statistic $\text{size}(\Omega(N_1[v]))$ together with all its neighbors. This is a reasonable first estimator for p (and $|N[v^*]|$ and $\text{size}(\Omega(N_1[v^*]))/\binom{|N[v^*]|}{2}$ provide reasonable estimators for m and q). There are, however, many potential improvements available regarding the estimation of p, involving resampling or

iteration or bias correction based on asymptotic alternative distribution moments. In any event, we see that estimating p requires identifying $\widehat{V}_A$, which is clearly harder than the testing problem considered herein. Thus, while it is true that power analyses are affected by unknown p, we feel that full-scale consideration of this issue at this time would obscure the simpler, basic comparative power issues which can be elucidated by considering known p.

In summation, no one invariant is uniformly most powerful at detecting increases in local "chatter"; even in this relatively simple setting, our investigation suggests that finite sample statistical inference on random graphs poses significant complexities. Our Monte Carlo investigation provides useful insight into the comparative behavior of various invariants.

6 Acknowledgements

The authors thank Andrey Rukhin for assistance regarding invariant distributions, and three referees for insightful comments which were instrumental in improving this manuscript.

7 Supplemental Materials

Code for reproducing all the simulation results presented in this paper is available at `http://www.glencoppersmith.com/code/`.

References

Bickel, P. and Doksum, K. (2001), *Mathematical statistics*, Prentice Hall Upper Saddle River, NJ.

Bollobás, B. (2001), *Random Graphs*, Cambridge University Press.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004), "The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation," In *Proceedings LREC*, pages 837–840.

Grothendieck, J., Gorin, A., and Borges, N. (2008), "Social Correlates of Turn-taking Behavior," in preparation.

Horn, R. and Johnson, C. (1985), *Matrix Analysis*, Cambridge University Press.

Leenders, R. (1995), Structure and influence: Statistical models for the dynamics of actor attributes, network structure, and their interdependence, Thesis Publishers.

Newman, M.E.J. (2003), "The Structure and Function of Complex Networks," *SIAM Review*, 45(2):167–256.

Newman, M.E.J., Barabosi, A.-L., and Watts, D.J. (2006), *The Structure and Dynamics of Networks*, Princeton: Princeton University Press.

Nowicki, K. and Wierman, J. (1988), "Subgraph Counts in Random Graphs Using Incomplete UStatistics Methods," *Discrete Mathematics*, 72:299–310.

Priebe, C., Conroy, J., Marchette, D., and Park, Y. (2005), "Scan Statistics on Enron Graphs," *Computational & Mathematical Organization Theory*, 11(3):229–247.

Rukhin, A. (2008), Asymptotic Analysis of Various Statistics for Random Graph Inference, PhD thesis, Department of Applied Math and Statistics Dissertation, Johns Hopkins University.

Rukhin, A. and Priebe, C.E. (2009a), "On the Limiting Distribution of a Graph Scan Statistic," submitted for publication.

Rukhin, A. and Priebe, C.E. (2009b), "A Comparative Power Analysis of the Maximum Degree and Size Invariants for Random Graph Inference," submitted for publication.

Sen, S., Spatschek, O., and Wang, D. (2004), "Accurate, Scalable In-network Identification of P2P Traffic Using Application Signatures," In *Proceedings WWW'04*, pages 512–521.

Ullman, D. and Scheinerman, E. (1997), *Fractional Graph Theory*, Wiley.

West, D. (2001), *Introduction to Graph Theory*, Prentice Hall Upper Saddle River, NJ, second edition.

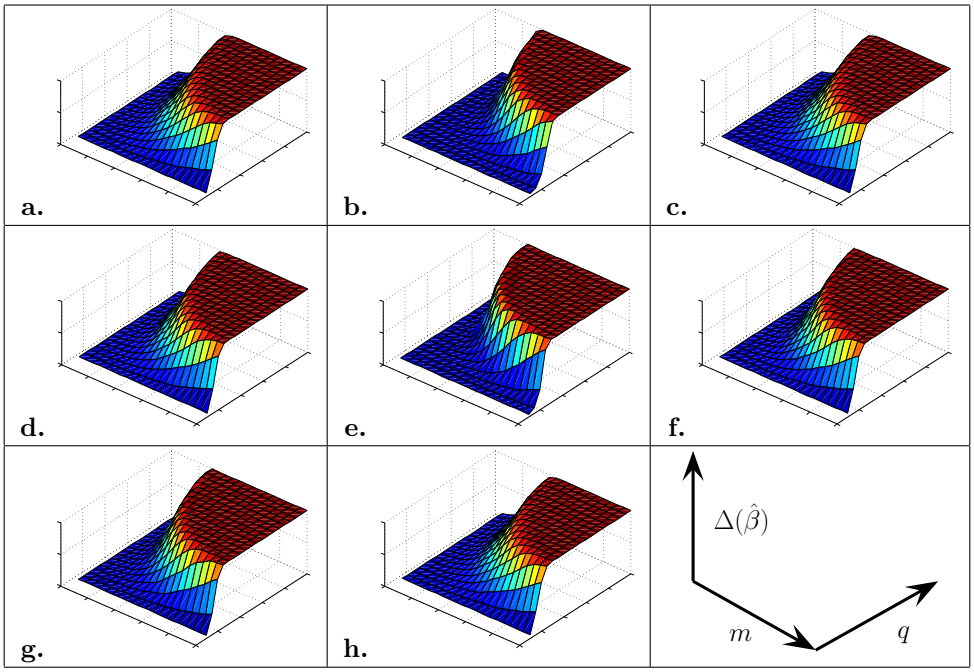

Figure 10: Power surface plots for the various graph invariants, as obtained from the Monte Carlo simulations for $n = 1000$, $p = 0.1$, $m \in \{5, 10, 15, \ldots, 100\}$, $q \in \{0.10, 0.15, 0.20, \ldots, 0.90\}$, $\alpha = 0.05$, and $R = 1000$. **a.** Number of edges, size(G). **b.** Maximum degree, $\delta(G)$. **c.** Greedy maximum average degree approximation, $\text{MAD}_g(G)$. **d.** Eigenvalue maximum average degree approximation, $\text{MAD}_e(G)$. **e.** Scan statistic, $S_1(G)$. **f.** Number of Triangles, $\tau(G)$. **g.** Global clustering coefficient, CC(G). **h.** Average Path Length, APL(G). Powers range from approximately α for small m or q to approximately 1 for large m and q for all invariants. Substantial differences exist, but may not be apparent, between the various invariants for moderate m, q; these differences are readily apparent in the pairwise comparisons (Figure 13).

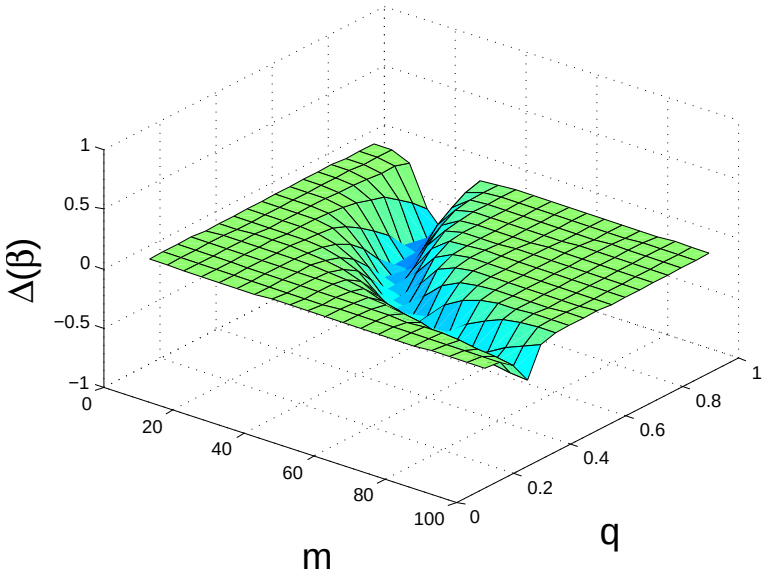

Figure 11: Comparative power surface for $\hat{\beta}_{\delta(G)} - \hat{\beta}_{S_1(G)}$ for $H_0 : ER(n = 1000, p = 0.1)$ vs. $H_A : \kappa(n = 1000, p = 0.1, m, q)$ for the range of m, q investigated. $S_1(G)$ has equal or superior power to $\delta(G)$ for all of Θ_A.

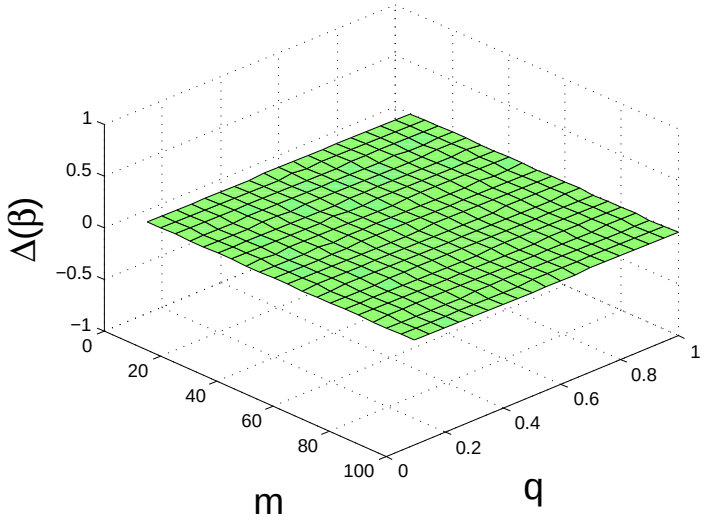

Figure 12: Comparative power surface for $\hat{\beta}_{\text{size}(G)} - \hat{\beta}_{\text{APL}(G)}$ for $H_0 : ER(n = 1000, p = 0.1)$ vs. $H_A : \kappa(n = 1000, p = 0.1, m, q)$ for the range of m, q investigated. The statistics size(G) and APL(G) have nearly identical power for all of Θ_A.

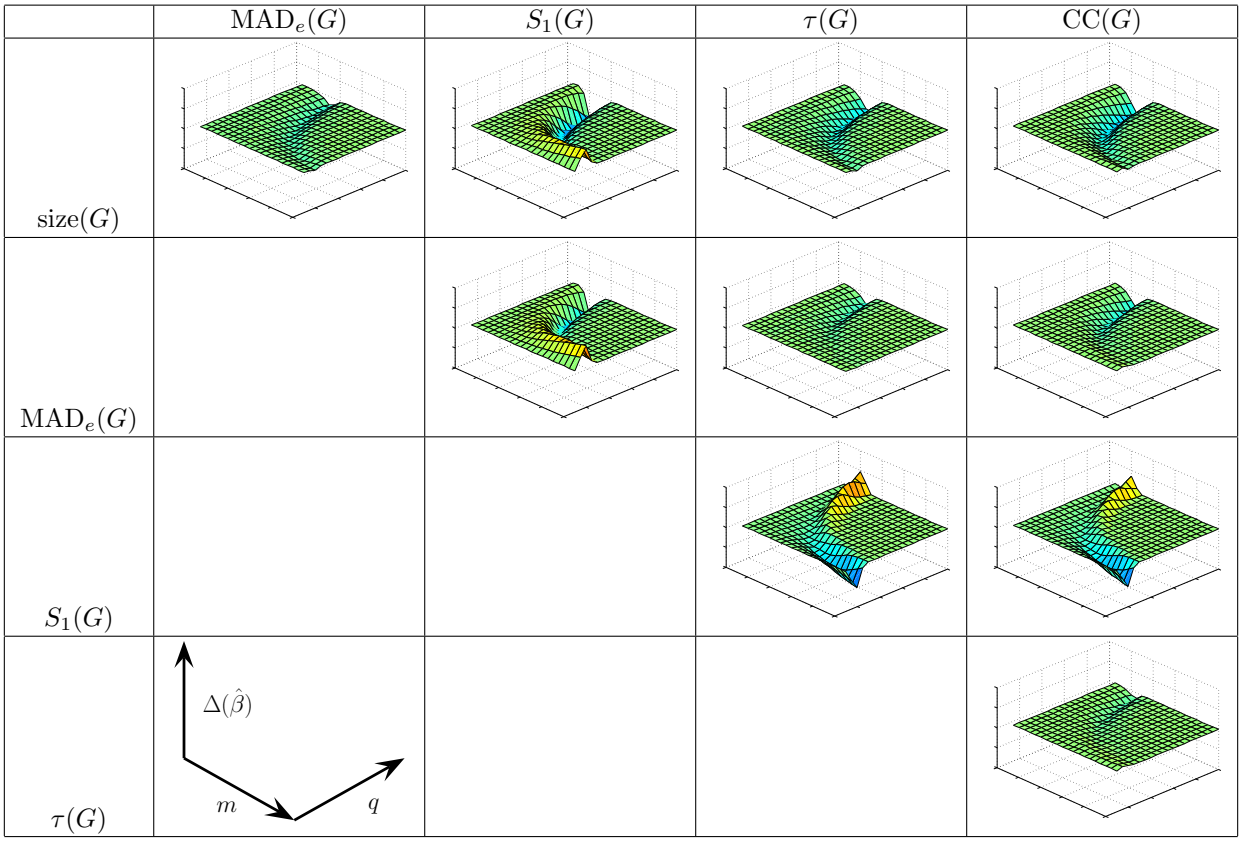

Figure 13: Comparative power surfaces for the various graph invariants, as obtained from Monte Carlo simulations for $n = 1000$, $p = 0.1$, $m \in \{5, 10, 15, \ldots, 100\}$, $q \in \{0.10, 0.15, 0.20, \ldots, 0.90\}$, $\alpha = 0.05$, and $R = 1000$. Each surface plot is representative of the power of the row invariant minus the power of the column invariant (e.g. the upper left corner depicts $\hat{\beta}_{\text{size}(G)} - \hat{\beta}_{\text{MAD}_e(G)}$) from Figure 10. Since powers are approximately α for small m or q and approximately 1 for large m and q for all invariants, power differences are approximately 0 in these regions. Substantial differences are readily apparent between the various invariants for moderate m, q in these comparative power surfaces. This phenomenon can be seen in more detail in Figure 14.

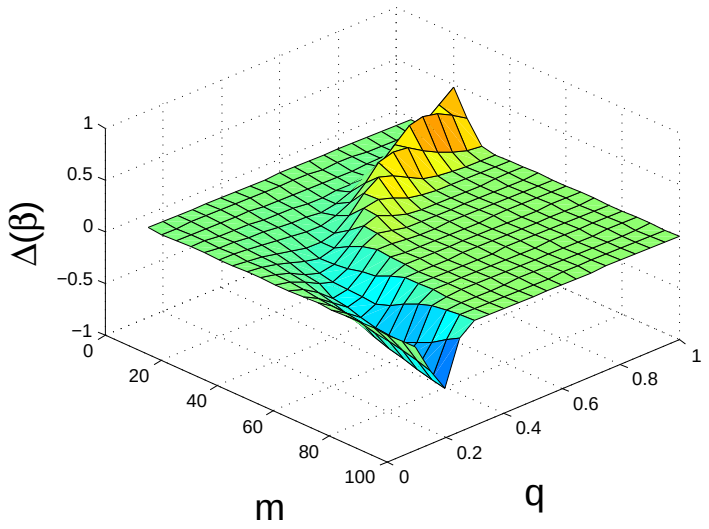

Figure 14: Comparative power surface $\hat{\beta}_{S_1(G)} - \hat{\beta}_{\tau(G)}$ for $H_0 : ER(1000, 0.1)$ vs. $H_A : \kappa(1000, 0.1, m, q)$ with $\alpha = 0.05$ and $R = 1000$, from Figure 13. Since powers are approximately α for small m or q and approximately 1 for large m and q for both invariants, power differences are approximately 0 in these regions. Substantial differences are readily apparent for moderate m, q: when m is large and q is small $\hat{\beta}_{\tau(G)} > \hat{\beta}_{S_1(G)}$; when m is small and q is large $\hat{\beta}_{S_1(G)} > \hat{\beta}_{\tau(G)}$. Neither invariant dominates the other throughout Θ_A.

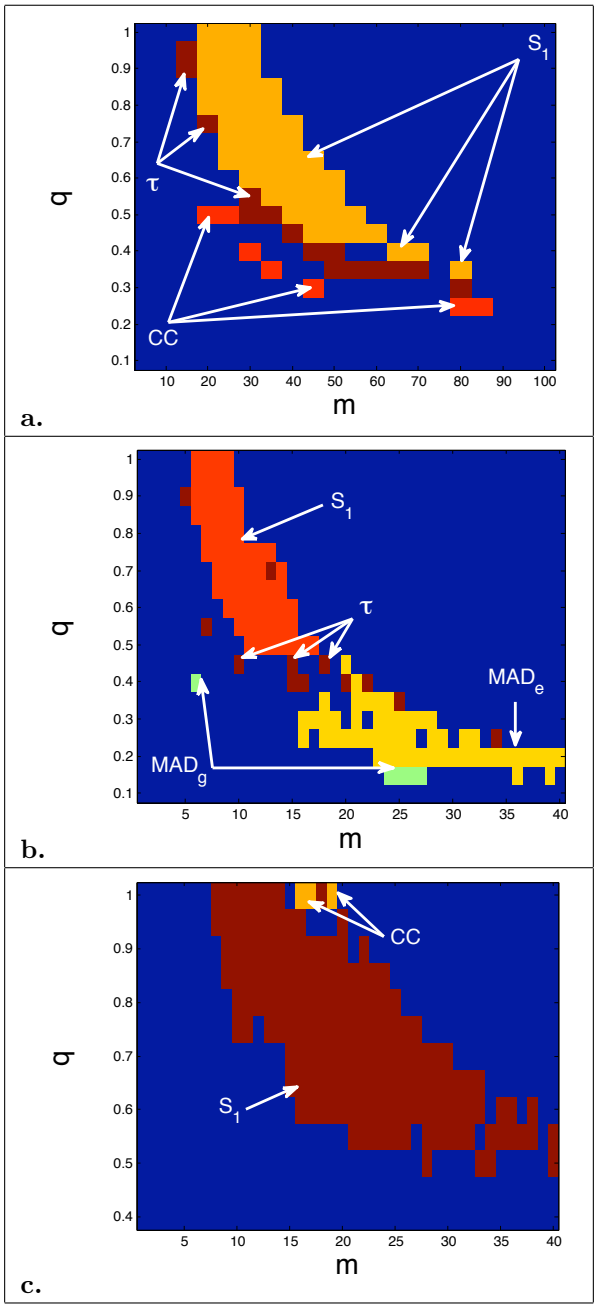

Figure 15: Most powerful statistic over Θ_A for selected n and p for $H_0 : ER(n,p)$ vs. $H_A : \kappa(n,p,m,q)$ with $\alpha = 0.05$ and $R = 1000$. The dark blue region is where no test is statistically significantly superior. For large m and q (relative to n and p), this is because all tests have $\hat{\beta} \approx 1$; for small m or small q, this is because all tests have $\hat{\beta} \approx \alpha$. **a.** $n = 1000$, $p = 0.1$, $m \in \{5, 10, 15, \ldots, 100\}$, $q \in \{0.10, 0.15, 0.20, \ldots, 1.0\}$ as in Figures 10 and 13. **b.** $n = 100$, $p = 0.1$, $m \in \{2, 4, 6, \ldots, 40\}$, $q \in \{0.10, 0.15, 0.20, \ldots, 1.0\}$. **c.** $n = 100$, $p = 0.4$, $m \in \{2, 4, 6, \ldots, 40\}$, $q \in \{0.40, 0.45, 0.50, \ldots, 1.0\}$.