Research Report

Department of Statistics

No. 2003:8

Random Graph Distributions of Degree variance

Jan Hagberg

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

Random Graph Distributions of Degree Variance

Jan Hagberg

Abstract

Exact and asymptotic distributions of the degree variance are investigated for Bernoulli graphs and uniform random graphs. For graphs of large order we show that the degree variance is approximately gamma distributed with parameters obtained from the first two moments of the degree variance. The usefulness of the results is illustrated by a graph centrality test with a critical value obtained from the gamma distribution.

Key words: Uniform Random Graphs, Bernoulli Graphs, Degree Variance, Gamma Approximation, Centrality Testing.

1 Introduction

Consider a random graph on n vertices with r edges, i.e. a random graph of order n and size r, and let X_i be the degree of vertex i, i.e. the number of edges incident to vertex i. The degree variance is defined as

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

where $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_i$. The distribution of S^2 depends on the random graph model that generates the edges. Section 2 gives a brief description of the two models of this paper: the $Bernoulli\,(n,p)$-model and the $Uniform\,(n,r)$-model. For further details on these models, see Hagberg (2003a). Some

properties of the degree variance and related enumeration problems are discussed in Section 3. Exact distributions of the degree variance for graphs of small order are given in Section 4 and approximate distributions for graphs of large order are given in Section 5. In Section 6, the accuracy of the approximations are investigated by comparing exact and simulated distributions with the approximate distributions. It is shown that we need to adjust the approximation for the dependence between the vertex degrees. Finally, in Section 7, the approximate distribution of S^2 is applied to graph centrality (or vertex heterogeneity) testing.

2 Random graph models

We consider a random graph on n labeled vertices. The vertices are labeled by integers $1, ..., n$. Let X_i be the number of edges incident to vertex i.

Model 1: $Bernoulli\,(n, p)$

With probability p, each pair of distinct vertices i and j is connected by an edge. These connections are made independently of each other. Let X_{ij} be an edge indicator that is 1 or 0 according to whether or not there is an edge between i and j. For $i = j$ it is convenient to define $X_{ii} = 0$ so that $X_i = \sum_{j=1}^{n} X_{ij}$. Since a vertex can have no more than $n-1$ edges it follows that

$$\begin{aligned} X_{ij} &\sim Bernoulli(p)\ , i \neq j, \\ X_i &\sim Bin(n-1, p)\ ,\ i = 1, 2, ..., n \text{ and} \\ R &\sim Bin\,(N, p)\,, \end{aligned}$$

where R is the number of edges and $N = \binom{n}{2}$.

Model 2: $Uniform\,(n, r)$

Let the number of edges, r, be fixed, $0 \leq r \leq N$. The r pairs of distinct vertices i and j connected by an edge, are chosen uniformly at random without replacement among the N pairs. It follows that X_{ij} is Bernoulli distributed and X_i is hypergeometrically distributed, i.e.

$$\begin{aligned} X_{ij} &\sim Bernoulli(\frac{r}{N}), \\ X_i &\sim Hypg.\,(r; n-1, N-n+1)\,. \end{aligned}$$

That is

$$P(X_i = x) = \frac{\binom{n-1}{x}\binom{N-n+1}{r-x}}{\binom{N}{r}} ,$$

$$\max\left\{0, r - \binom{n-1}{2}\right\} \leq x \leq \min\{n-1, r\} .$$

3 Degree variance

To avoid fractions, it is convenient to use the integer valued random variable

$$\begin{aligned} Z &= n^2 S^2 = n \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \\ &= \sum_{i<j} \left(X_i - X_j\right)^2 . \end{aligned} \tag{3.1}$$

We see that Z takes the value zero if and only if the graph is regular, i.e. all degrees are equal.

The distribution of the degree variance is related to the distribution of the ordered degree sequence, but is even more complicated to determine. Isomorphic graphs have the same ordered degree sequence and the same degree variance. Complementary graphs have the same degree variance even if their ordered degree sequences are not equal. Non-isomorphic graphs can have the same degree variance even if their ordered degree sequences are not equal.

Consider any given ordered degree sequence $(X_1, ..., X_n)$ where $X_1 \geq , ..., \geq X_n$ and its ordered complement, $(n-1-X_n, ..., n-1-X_1)$. The two ordered sequences correspond to a graph and its complement, and they have the same degree variance, but that value is not necessarily unique for these two graphs.

The two ordered degree sequences $(1,1,1,1,1,1)$ and $(3,3,3,3,3,3)$ both yield $Z = 0$ and the two ordered degree sequences $(4,1,1,1,1,0)$ and $(4,2,2,1,1,0)$ both yield $Z = 56$. The mean is not the same for the last two sequences, but the sequences $(3,1,1,1,0,0)$ and $(2,2,2,0,0,0)$ have the same mean and both yield $Z = 36$. So the same degree variance doesn't imply that the graphs are isomorphic, complementary, regular or have the same mean. The graphs corresponding to the six degree sequences above are shown in Figure 1.

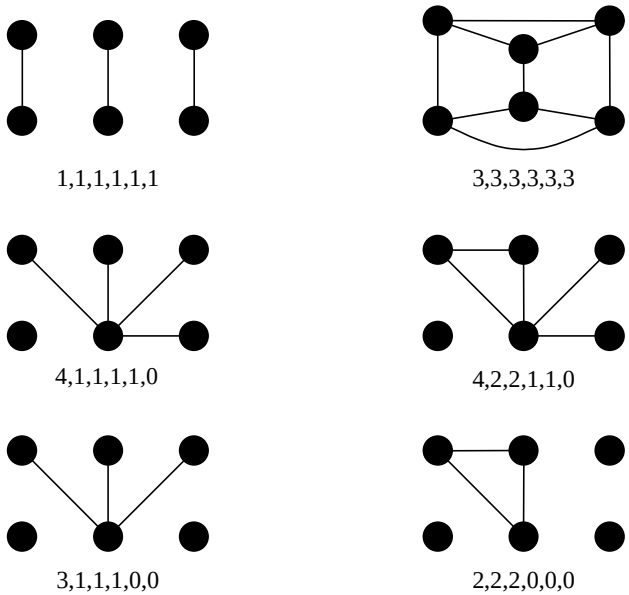

Figure 1. Six graphs and their ordered degree sequences. Graphs in the same row have the same degree variance.

An algorithm that generates all ordered degree sequences of a given length is developed by Cohen et al. (1994) and a formula for the number of $(0,1)$-matrices for any ordered degree sequence is given by Wang and Zhang (1998). For graphs of small order, the latter can also be computed by the program ZO 2.3, developed by Snijders (2002).

The number of ordered degree sequences increases rapidly with n, as indicated by Table 1. The number of unlabeled graphs increase even more rapidly. For methods of graphical enumeration see, for example, Harary (1969) , Deo (1973) or Sloane & Plouffe (1995).

n	Number of unlabeled graphs	Number of ordered degree sequences	Number of distinct s^2-values
1	1	1	1
2	2	2	1
3	4	4	2
4	11	11	4
5	34	31	11
6	156	102	14
7	1 044	342	43
8	12 346	1 213	34
9	274 668	4 361	102
10	12 005 168	16 016	111
11	1 018 997 864	59 348	296
12	165 091 172 592	222 117	262

Table 1. The number of unlabeled graphs, ordered degree sequences, and distinct s^2- values for $n = 1, ..., 12$.

Notice that the number of distinct s^2-values is not monotonically increasing with n.

4 Exact distributions

As indicated in the previous section, it is very time consuming to find the exact distribution of S^2 if $n > 7$. Even the task of determining the possible values of S^2 is cumbersome for modest values of n. From Table 2 below, it is possible to obtain simple probability functions of Z under the $Bernoulli\,(n, p)$ model of order 3 and 4:

$$n = 3 : Z \sim 2Be\,(3pq) \quad \text{and } n = 4 : Z \sim 4Bin\,(3, 2pq)\,.$$

In fact, for $n = 3$

$$Z = 2I\,(1 \leq R \leq 2)\,,$$

and for $n = 4$ some algebra reveals that

$$Z = 4\left[(X_{12} - X_{34})^2 + (X_{13} - X_{24})^2 + (X_{14} - X_{23})^2\right].$$

Due to the irregularities in the tails of the distributions and the rapidly increasing number of degree sequences, it is much harder, or in practice impossible, to derive the corresponding functions for the $Bernoulli\,(n,p)$ model of large order. Thus, there is a need for approximate methods.

The numbers needed for the exact distributions of the degree variance for uniform random graphs and Bernoulli graphs of order n for $n = 3, ..., 7$ are given below. Here $L(n,r,z)$ is the number of labeled graphs of order n, size r and degree variance z/n^2. For the $Uniform\,(n,r)$-model

$$P\,(Z=z) = \frac{L\,(n,r,z)}{2^N},$$

and for the $Bernoulli\,(n,p)$-model

$$P\,(Z=z) = \sum_r L\,(n,r,z)\,p^r q^{N-r}.$$

Notice that $\sum_z L\,(n,r,z) = \binom{N}{r}$ and $\sum_r \sum_z L\,(n,r,z) = 2^N$.

Table 2. The numbers $L\,(n,r,z)$ of labeled graphs of order n distributed according to r and z.

$z \setminus r$	0, 3	1, 2
0	1	
2		3
Sum	1	3

$n = 3$. $2^{\binom{3}{2}} = 8$ labeled graphs distributed according to z and r.

$z \setminus r$	0, 6	1, 5	2, 4	3
0	1		3	
4		6		12
8			12	
12				8
Sum	1	6	15	20

$n = 4$. $2^{\binom{4}{2}} = 64$ labeled graphs distributed according to z and r.

$z \setminus r$	0, 10	1, 9	2, 8	3, 7	4, 6	5
0	1					12
4			15	30		
6		10			70	
10						120
14			30	60		
16					75	
20						60
24				30		
26					60	
30						60
36					5	
Sum	1	10	45	120	210	252

$n = 5$. $2^{\binom{5}{2}} = 1024$ labeled graphs distributed according to z and r.

$z \setminus r$	0, 15	1, 14	2, 13	3, 12	4, 11	5, 10	6, 9	7, 8
0	1			15			70	
8		15	45		270	465		810
12				180			1080	
20			60		480	972		1800
24				180			1530	
32					405	810		
36				80			1080	
44					180	480		1080
48							810	
56					30	270		630
60							360	
68								360
72							75	
80						6		
Sum	1	15	105	455	1365	3003	5005	6435

$n = 6$. $2^{\binom{6}{2}} = 32768$ labeled graphs distributed according to z and r.

z \ r	0, 21	1, 20	2, 19	3, 18	4, 17	5, 16	6, 15	7, 14	8, 13	9, 12	10, 11
0	1							465			
6				105	315						5670
10		21					3507		9660		
12			105			2625				19355	
14								10500			
20				630	1890						48300
24							12810		35910		
26			105			4935				45360	
28								27300			
34				420	1890						62790
38							11970		41895		
40						6552				71190	
42								22365			
48				175	1365						79170
52							13440		46620		
54						3255				48055	
56								26880			
62					420						55230
66							6510		25620		
68						2100				51030	
70								13545			
76					105						50715
80							4095		25305		
82						840				27090	
84								9240			
90											21735
94							1470		10500		
96										18585	
98								3990			
104											18270
108							455		5355		
110						42				8190	
112								1890			
118											7455
122									1575		
124										4200	
132											3150
136									420		
138										875	
140								105			
146											210
150							7				
160											21
Sum	1	21	210	1330	5985	20349	54264	116280	203490	293930	352716

$n = 7$. $2^{\binom{7}{2}} = 2097152$ labeled graphs distributed according to z and r.

5 Gamma approximations

A random variable Y has a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$,denoted by $Y \sim Gamma(\alpha, \beta)$, if its density function is given by

$$g_{\alpha,\beta}(y) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} e^{-y/\beta} \ , \ 0 \leqslant y < \infty.$$

For $Y \sim Gamma(\alpha, \beta)$ it holds that

$$E(Y^k) = \frac{\beta^k}{\Gamma(\alpha)} \Gamma(\alpha + k) = \beta^k \prod_{j=0}^{k-1} (\alpha + j) \ , \ k \geqslant 0. \tag{5.1}$$

The first two central moments are thus

$$E(Y) = \alpha\beta \ \text{ and } Var(Y) = \alpha\beta^2. \tag{5.2}$$

Let $U_i, i = 1, 2, ..., n$ be a sequence of n *independent identically distributed* normal random variables with mean μ and variance σ^2, that is, $U_1, ..., U_n$ are *iid* $N(\mu, \sigma^2)$. Let $W = \frac{1}{n}\sum_{i=1}^{n} \left(U_i - \overline{U}\right)^2$where $\overline{U} = \frac{1}{n}\sum_{i=1}^{n} U_i$. Then, according to known results (Johnson & Kotz 1970), W is gamma distributed i.e.

$$W \sim Gamma\left(\frac{n-1}{2}, \frac{2\sigma^2}{n}\right) \tag{5.3}$$

and according to (5.1)

$$E(W) = \frac{n-1}{n}\sigma^2 \ \text{ and } Var(W) = 2\left(\frac{n-1}{n^2}\sigma^4\right). \tag{5.4}$$

The degrees of the vertices under the $Bernoulli(n, p)$ model are binomially distributed with $\mu = (n-1)p$ and $\sigma^2 = (n-1)pq$. Binomially distributed random variables are approximately normally distributed if their variances are sufficiently large. Thus, neglecting the weak pair wise dependence between the vertex degrees (See Hagberg (2000, 2003a).), we can argue that S^2 should be approximately

$$Gamma\left(\frac{n-1}{2}, \frac{2(n-1)pq}{n}\right). \tag{5.5}$$

However, due to the dependence between the vertex degrees, this gamma distribution does not have the correct mean and variance. A gamma distribution with the correct mean and variance can be obtained by choosing the gamma distribution parameters α and β so that $\alpha\beta = ES^2$ and $\alpha\beta^2 = VarS^2$, where ES^2 and $Var\left(S^2\right)$ are given by Hagberg (2000, 2003a). This leads to

$$\alpha = \frac{n(n-1)}{2\left[1+(n-6)pq\right]}pq \quad \text{and } \beta = \frac{2(n-2)}{n^2}\left[1+(n-6)pq\right]. \tag{5.6}$$

Table 3 shows the first three moments of S^2 derived under independence assumptions (unadjusted gamma) and adjusted for the dependence (adjusted gamma).

Moment	Unadjusted gamma	Adjusted gamma
1	$\frac{(n-1)^2}{n}pq$	$\frac{(n-1)(n-2)}{n}pq$
2	$\frac{(n+1)(n-1)^3}{n^2}p^2q^2$	$\frac{(n-1)(n-2)^2pq((n+4)(n-3)pq+2)}{n^3}$
3	$\frac{(n+3)(n+1)(n-1)^4}{n^3}p^3q^3$	$\frac{8(n-1)(n-2)^3}{n^5}pq$
		$+\frac{2(n-1)(n-2)^3\left(3n^2+5n-48\right)}{n^5}p^2q^2$
		$+\frac{(n-1)(n-2)^3(n-3)(n+4)\left(n^2+3(n-8)\right)}{n^5}p^3q^3$

Table 3. The first three approximate moments of S^2.

The exact first two moments of S^2 equal the adjusted gamma moments, and the exact third moment is equal to

$$\begin{aligned} E\left(S^2\right)^3 &= \frac{4\left(n-1\right)\left(n-2\right)^3}{n^5}pq \\ &+\frac{2\left(n-1\right)^{(3)}\left(3n-4\right)\left[\left(n-2\right)\left(n+6\right)-8\right]}{n^5}p^2q^2 \\ &+\frac{\left(n-1\right)^{(3)}\left[n^4\left(n+3\right)-4\left(3n-4\right)\left[3\left(n-2\right)\left(n+6\right)-n-4\right]\right]}{n^5}p^3q^3, \end{aligned} \tag{5.7}$$

where the exponent in parenthesis denotes the falling factorial.

Let D_n denote the difference between the exact third moment of S^2 and the third adjusted gamma moment and notice that the coefficients of

n^8, n^7,and n^6 are equal in the numerators of p^3q^3/n^5. We have that

$$\begin{aligned} D_n &= -\frac{4\left(n-1\right)\left(n-2\right)^3}{n^5}pq + \frac{4\left(n-1\right)\left(n-2\right)\left(3n^3-22n^2+48n-24\right)}{n^5}p^2q^2 \\ &\quad -\frac{64\left(n-1\right)^2\left(n-2\right)\left(n-3\right)\left(n-4\right)}{n^5}p^3q^3. \end{aligned}$$

For fixed p, $|D_n|$ increases with n and

$$\lim_{n\to\infty} D_n = D = 4p^2q^2\left(3-16pq\right). \tag{5.8}$$

The maximum value of D is $\frac{1}{16}$ and is obtained for $p = \frac{1}{2} \pm \sqrt{\frac{1}{8}}$, and the minimum value $-\frac{1}{4}$ is obtained for $p = \frac{1}{2}$ as shown in Figure 2.

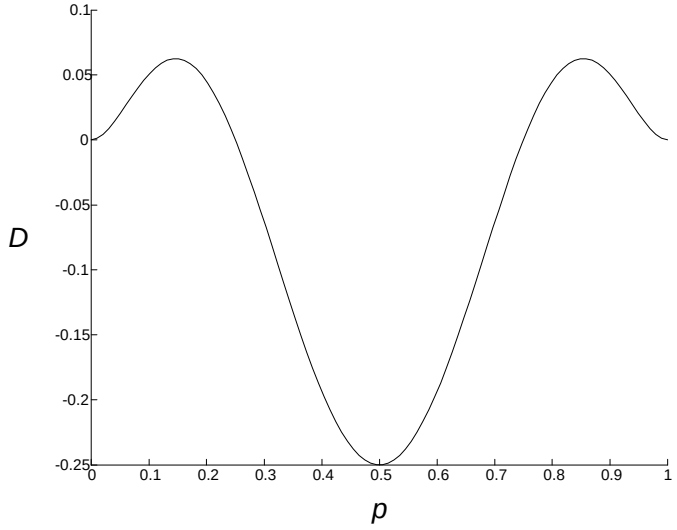

Figure 2. D plotted against p.

The corresponding differences between the exact moments and the moments of the unadjusted gamma distributed variable W tend to the following limits:

$$\begin{aligned} \left(ES^2 - EW\right) &\to -pq\ , \\ \frac{E\left(S^2\right)^2 - EW^2}{n} &\to -2p^2q^2 \text{ and} \\ \frac{E\left(S^2\right)^3 - EW^3}{n^2} &\to -3p^3q^3. \end{aligned} \tag{5.9}$$

Thus, the unadjusted gamma approximation gives a bias to the mean and increasing biases to higher moments. The adjusted gamma approximation with correct first two moments has a bias for the third moment which is bounded by $-1/4$ and $1/16$.

Since the distribution of S^2 is discrete and the gamma distribution is continuous, we can improve the approximation by the use of a *continuity correction.* The improvement is verified by simulations and exact calculations. Let $z_1 < z_2 < \cdots < z_m$ denote the ordered possible values of $Z = n^2S^2$. and let $s_j^2 = z_j/n^2$ for $j = 1, ..., m$ and let $G_{\alpha\beta}$ denote the distribution function of a $Gamma\,(\alpha, \beta)$ variable. We have

$$\begin{aligned} P(S^2 &= s_j^2) = P\left(z_j - \frac{z_j - z_{j-1}}{2} < Z < z_j + \frac{z_{j+1} - z_j}{2}\right) \\ &= P\left(z_j - a < Z < z_j + b\right) = P\left(s_j^2 - \frac{a}{n^2} < S^2 < s_j^2 + \frac{b}{n^2}\right) \\ &\approx G_{\alpha,\beta}\left(s_j^2 + \frac{b}{n^2}\right) - G_{\alpha,\beta}\left(s_j^2 - \frac{a}{n^2}\right) \\ &\approx \frac{a+b}{n^2} g_{\alpha,\beta}\left(s_j^2\right) = \frac{s_{j+1}^2 - s_{j-1}^2}{2} g_{\alpha,\beta}\left(s_j^2\right), \end{aligned} \tag{5.10}$$

$$P(S^2 \leqslant s_j^2) = P(S^2 \leqslant s_j^2 + \frac{b}{n^2}) \approx G_{\alpha,\beta}\left(s_j^2 + \frac{b}{n^2}\right) \tag{5.11}$$

and

$$P(S^2 \geqslant s_j^2) = P(S^2 \geqslant s_j^2 - \frac{a}{n^2}) \approx 1 - G_{\alpha,\beta}\left(s_j^2 - \frac{a}{n^2}\right). \tag{5.12}$$

where $a = \frac{z_j - z_{j-1}}{2}$ and $b = \frac{z_{j+1} - z_j}{2}$.

The possible z-values are not all known for large n. Table 2 lists the z-values for $n \leq 7$ and Table 7 contains the values for $n = 8$. All possible values of z for graphs of order $n = 9, 10, 11$ and 12 are listed in Table 8 at the end of the paper, and a method for deriving the values in graphs of larger order is treated in Hagberg (2003c). Below follows a method for *simplified continuity correction.*

Theorem 1 *For $n > 2$ and any two consecutive values, z_j and z_{j+1}, it holds that*

$$\begin{aligned} \min_{1 \leqslant j < m} (z_{j+1} - z_j) &= 2 \text{ if } n \text{ is odd,} \\ \min_{1 \leqslant j < m} (z_{j+1} - z_j) &= 4 \text{ if } n \text{ is even.} \end{aligned}$$

Proof. By writing

$$z = 2n\left[r + \sum_{i=1}^{n}\binom{x_i}{2}\right] - 4r^2, \tag{5.13}$$

where r is the number of edges, we see that z is even and z is also divisible by 4 if n is even. Thus, in order to prove the theorem we need to show that for some graph G of order n there is another graph G^* of order n such that

$$z\left(G^*\right) - z\left(G\right) = \begin{cases} 2 & \text{if} \quad n \text{ odd} \\ 4 & \text{if} \quad n \text{ even.} \end{cases}$$

Consider a graph G of order n and size $r \leq \binom{n}{2} - 1$ having $X_{12} = 0$. Let G^* be G extended with edge $(1,2)$. Then

$$\begin{aligned} z &= z\left(G\right) = n\sum_{i=1}^{n} x_i^2 - 4r^2, \\ z^* &= z\left(G^*\right) = n\left(\left(x_1+1\right)^2 + \left(x_2+1\right)^2 + \sum_{i=1}^{n} x_i^2\right) - 4\left(r+1\right)^2, \end{aligned}$$

and it follows that

$$z^* - z = 2n\left(1 + x_1 + x_2\right) - 4\left(2r+1\right).$$

Taking $x_1 = x_2 = 0$ and $r \leq \binom{n-2}{2}$ we obtain

$$z^* - z = 2n - 8r - 4 = 2 \quad \text{if } n = 4r+3$$

which proves the theorem for $n = 3, 7, 11, ...$

Taking $x_1 = x_2 = 1$ and $r \leq 2 + \binom{n-2}{2}$ we obtain

$$z^* - z = 6n - 8r - 4 = 2 \quad \text{if } n = 4k+1 \text{ and } r = 3k.$$

for some positive integer k, proving the theorem for $n = 5, 9, 13, ...$

Finally, taking $x_1 = 0, x_2 = 1$ and $r \leq 1 + \binom{n-2}{2}$ we obtain

$$z^* - z = 4n - 8r - 4 = 4 \quad \text{if } n = 2r+2\ ,$$

proving the theorem for $n = 2, 4, 6, ...$ ■

According to this theorem, if we observe a value s_j^2 and don't know the values of s_{j-1}^2 and s_{j+1}^2 we can use

$$a = b = \begin{cases} 1 & \text{if } n \text{ is odd} \\ 2 & \text{if } n \text{ is even} \end{cases} \tag{5.14}$$

in (5.10) - (5.12).

Under the $Uniform\,(n,r)$-model the degrees of the vertices are hypergeometrically distributed with mean $\mu = 2r/n$ and variance $\sigma^2 = 2r\,(n^2 - n - 2r)\,/(n^2\,(n+1))$. According to Johnson, Kotz & Kemp (1992) hypergeometric random variables can be approximated by the normal distribution. By using the same arguments as used for the $Bernoulli\,(n,p)$-model, it follows that gamma approximation is valid for S^2 in the $Uniform\,(n,r)$-model. In accordance with the $Bernoulli\,(n,p)$-model, we choose the gamma distribution parameters α and β so that $\alpha\beta = ES^2$ and $\alpha\beta^2 = VarS^2$, where ES^2 and $VarS^2$ are given by Hagberg (2000, 2003a). Thus, S^2 in the $Uniform(n,r)$ model is approximately $Gamma\,(\alpha\ ,\ \beta)$ where the parameters adjusted for dependence are

$$\alpha = \frac{r\,(n+2)\,[n\,(n-1)-4]\,[n\,(n-1)-2r]}{2n^2\,(r-1)\,[n\,(n-1)-2\,(r+1)]} \tag{5.15}$$

and

$$\beta = \frac{4\,(r-1)\,[n\,(n-1)-2\,(r+1)]}{(n+2)\,(n+1)\,[n\,(n-1)-4]}. \tag{5.16}$$

We can also improve the approximation by the use of a *continuity correction.* For the degree variance S^2 in $Uniform(n,r)$-graphs we use

$$a = b = n. \tag{5.17}$$

in (5.10) - (5.12). It is shown in Hagberg (2003b) that the difference between any two consecutive values of Z is $2n$, except for the right tail of the distribution. In the right tail the difference between any two consecutive values is at least $2n$.

6 Simulation results

In this section $F(s^2)$ means the exact or a simulated distribution function of S^2, $G(s^2)$ means the gamma distribution function in the adjusted approximation with continuity correction, and $G^*(s^2)$ means the gamma distribution function in the adjusted approximation without continuity correction. For the $Bernoulli(n,p)$-model, the function F is based on the exact distribution of S^2 for $n = 6$ and 7, it is based on the approximate distribution of S^2 obtained from 10^7 simulated graphs for $n = 8, 9, ..., 15, 20, 30$ and 10^6 simulated graphs for $n = 100$. Due to the increasing computer time needed to simulate graphs of higher order, fewer graphs are simulated for $n = 100$. Furthermore, to investigate the accuracy of the adjusted gamma approximation to the distribution function of S^2 in $Uniform(n,r)$-graphs, 10^6 graphs were simulated for $n = 8, ..., 12, 15$ and various values of r. The exact distribution is used for $n = 7$.

$n = 6$, $p = 0.1$						
z	$P(S^2 = \frac{z}{36})$	Adj. diff.	Unadj. diff.	$F(\frac{z}{36})$	Adj. diff.	Unadj. diff.
0	.210155	-.036779	.125352	.210155	-.036779	.125352
8	.467669	.133990	.145285	.677824	.097211	.270637
12	.051256	-.147356	-.228063	.729080	-.050145	.042573
20	.171050	.063438	.005814	.900130	.013293	.048388
24	.051431	-.004643	-.031431	.951561	.008651	.016957
32	.015618	-.012975	-.022267	.967179	-.004324	-.005311
36	.023013	.008633	.006686	.990192	.004308	.001376
44	.007374	.000210	.000620	.997566	.004519	.001996
48	.000314	-.003230	-.002396	.997880	.001289	-.000400
56	.001913	.000170	.000850	.999793	.001459	.000450
60	.000140	-.000715	-.000270	.999933	.000744	.000180
68	.000017	-.000400	-.000138	.999950	.000344	.000042
72	.000029	-.000174	-.000029	.999980	.000170	.000013
80	.000021	-.000078	-.000001	1.000000	.000093	.000013

$n = 6, \quad p = 0.5$						
z	$P(S^2 = \frac{z}{36})$	Adj. diff.	Unadj. diff.	$F(\frac{z}{36})$	Adj. diff.	Unadj. diff.
0	.005249	.002218	-.003899	.005249	.002218	-.003899
8	.097961	.047829	.038574	.103211	.050047	.034675
12	.076904	-.048359	-.024155	.180115	.001688	.010520
20	.202148	.033779	.082007	.382263	.035467	.092526
24	.104370	-.065564	-.017482	.486633	-.030098	.075045
32	.181274	.035667	.068299	.667908	.005570	.143344
36	.070801	-.041438	-.028068	.738709	-.035868	.115276
44	.106201	.025849	.023094	.844910	-.010020	.138370
48	.049439	-.005033	-.018376	.894348	-.015052	.119994
56	.056763	.021352	.002674	.951111	.006300	.122668
60	.021973	-.000295	-.020396	.973084	.006005	.102273
68	.021973	.008342	-.010732	.995056	.014348	.091541
72	.004578	-.003582	-.020363	.999634	.010766	.071178
80	.000366	-.004427	-.018460	1.000000	.006338	.052718

Table 4. The differences between the.exact distribution of S^2 and the adjusted and unadjusted gamma approximations respectively for $n = 6$, $p = 0.1$ and $p = 0.5$.

In Table 4 we have used the continuity correction (5.10) for the probability function, and the continuity correction (5.11) for the distribution function. Table 5 below shows the Kolmogorov distance, i.e. the greatest absolute deviation between the exact or simulated distribution function of S^2 and the adjusted and the unadjusted gamma approximation respectively for various values of n and p. Formally, the Kolmogorov distance between $F(s^2)$ and $G(s^2)$ is given by $\max_{s^2} |F(s^2) - G(s^2)|$. The corresponding distance in the upper decile of G is given within parenthesis. The accuracy of the approximation in the upper decile is important for hypothesis testing as shown in Section 7. For $n \leq 12$, the continuity correction (5.11) is used and for $n > 12$ the continuity correction (5.14) is used. Table 6 shows the corresponding values for the adjusted gamma approximation to S^2 in $Uniform(n, r)$-graphs where the continuity correction is given by (5.17).

n		$p = 0.1$	$p = 0.2$	$p = 0.4$	$p = 0.5$
7	Adj.	.1131 (.0252)	.0610 (.0112)	.0453 (.0076)	.0474 (.0088)
	Unadj.	.2854 (.0547)	.1586 (.0629)	.1319 (.1234)	.1334 (.1064)
8	Adj.	.0350 (.0063)	.0125 (.0015)	.0201 (.0047)	.0203 (.0048)
	Unadj.	.1557 (.0104)	.1088 (.0661)	.1118 (.0973)	.1100 (.0946)
9	Adj.	.0873 (.0162)	.0193 (.0053)	.0241 (.0042)	.0236 (.0051)
	Unadj.	.1617 (.0332)	.1157 (.0573)	.1053 (.0867)	.1076 (.0936)
10	Adj.	.0370 (.0019)	.0057 (.0014)	.0114 (.0028)	.0121 (.0033)
	Unadj.	.1387 (.0113)	.0981 (.0552)	.0939 (.0781)	.0944 (.0801)
12	Adj.	.0245 (.0054)	.0081 (.0012)	.0101 (.0025)	.0106 (.0028)
	Unadj.	.1366 (.0164)	.0949 (.0508)	.0883 (.0698)	.0883 (.0734)
15	Adj.	.0257 (.0087)	.0135 (.0034)	.0108(.0028)	.0100 (.0027)
	Unadj.	.1205 (.0213)	.0888 (.0466)	.0822 (.0595)	.0816 (.0614)
20	Adj.	.0136 (.0029)	.0056 (.0013)	.0046 (.0014)	.0049 (.0015)
	Unadj.	.0417 (.0727)	.0715 (.0352)	.0668 (.0478)	.0668 (.0486)
30	Adj.	.0111 (.0018)	.0045 (.0013)	.0028 (.0012)	.0039 (.0017)
	Unadj.	.0716 (.0110)	.0548 (.0281)	.0546 (.0357)	.0544 (.0367)
100	Adj.	.0020 (.0011)	.0033 (.0015)	.0011 (.0006)	.0009 (.0004)
	Unadj.	.0301 (.0075)	.0287 (.0131)	.0287 (.0154)	.0289 (.0162)

Table 5. The Kolmogorov distances for S^2. The distances in the upper deciles are given within parenthesis.

We see from Table 4 and 5 that the adjusted gamma approximation to the distribution function of S^2 in Bernoulli graphs works well, especially when $P\left(S^2 \geqslant s^2\right) \leqslant 0.10$. For $n = 8, ..., 12$ the adjusted approximation is very good when p is close to 0.2. For graphs of higher order the approximation is better when the variance is higher i.e. when p tends to 0.5. When p is fixed the accuracy of the approximation for graphs of order n is better than the accuracy for graphs of order $n + 1$ if n is even. The latter is due to the relative smoothness of the distribution when n is even and can be seen from Figure 3 and 4, that show the distribution of S^2 for $p = 0.5$ and $n = 7$ and 10 respectively. For further details on the smoothness of the distribution, see Hagberg (2003b). From Table 5 it can be seen that the unadjusted gamma approximation is bad, which agree with (5.9).

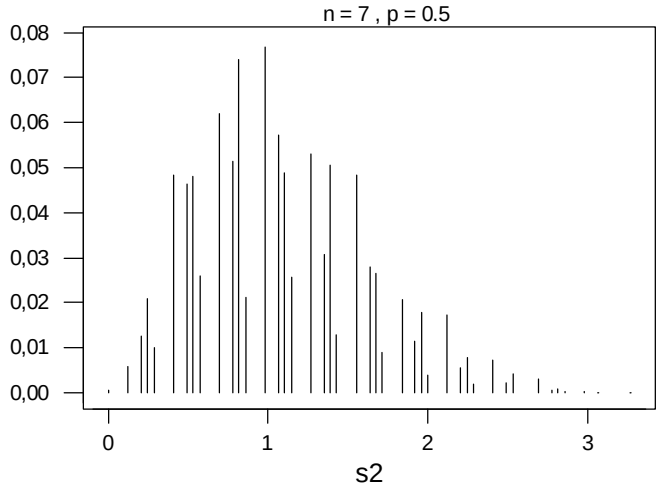

Figure 3. The exact distribution of S^2 for $n = 7$ and $p = 0.5$.

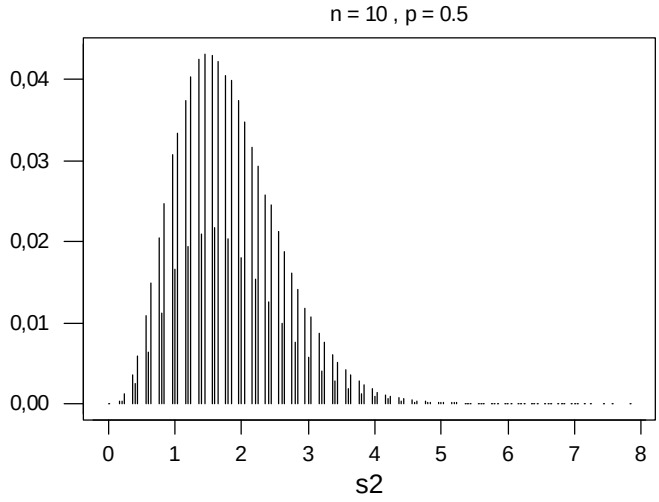

Figure 4. The simulated distribution of S^2 for $n = 10$ and $p = 0.5$.

Figure 5 below shows the differences $F\left(s^2\right) - G(s^2)$,under the $Bernoulli\,(n, p)$-

model, plotted against $F(s^2)$ for $n = 15$. The dependence structure of the differences varies for different values of n and p and can hardly be modeled without knowledge of the true distribution.

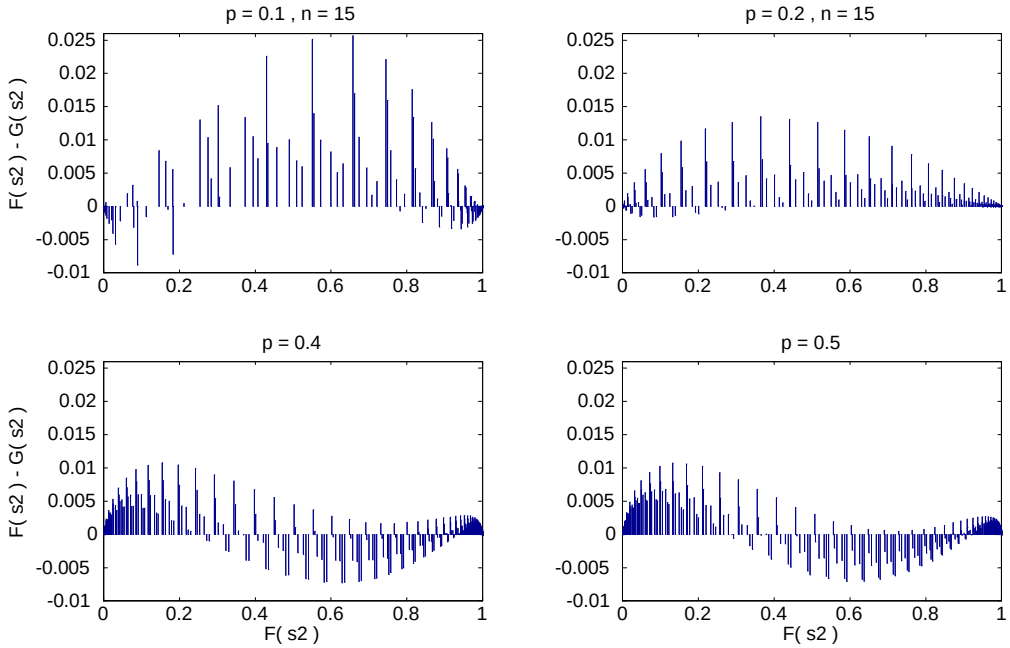

Figure 5. $F(s^2) - G(s^2)$ plotted against $F(s^2)$ for $n = 15$.

In Figure 6, $F(s^2) - G(s^2)$ is plotted against s^2 for $n = 15$ and $p = 0.5$. From the figure we get an inkling of the structure of the differences. The minimum value of $F(s^2) - G(s^2)$ is -0.007 and is obtained for $s^2 \approx 3 \approx E(S^2)$. For $n = 15$ and $p = 0.5$ we have that $P(S^2 \leq E(S^2)) \approx 0.54$. A figure similar to Figure 6 can be obtained if a binomial random variable X is approximated by a normal distributed random variable with mean np and variance npq and $P(X \leq E(X)) \approx 0.54$, using continuity correction. This is shown in Figure 7 where $X \sim Bin(100, 0.6)$ and $N(x)$ is the normal approximated distribution function of X. The minimum value in Figure 7 is -0.003 and is obtained for $x = 60 = E(X)$.

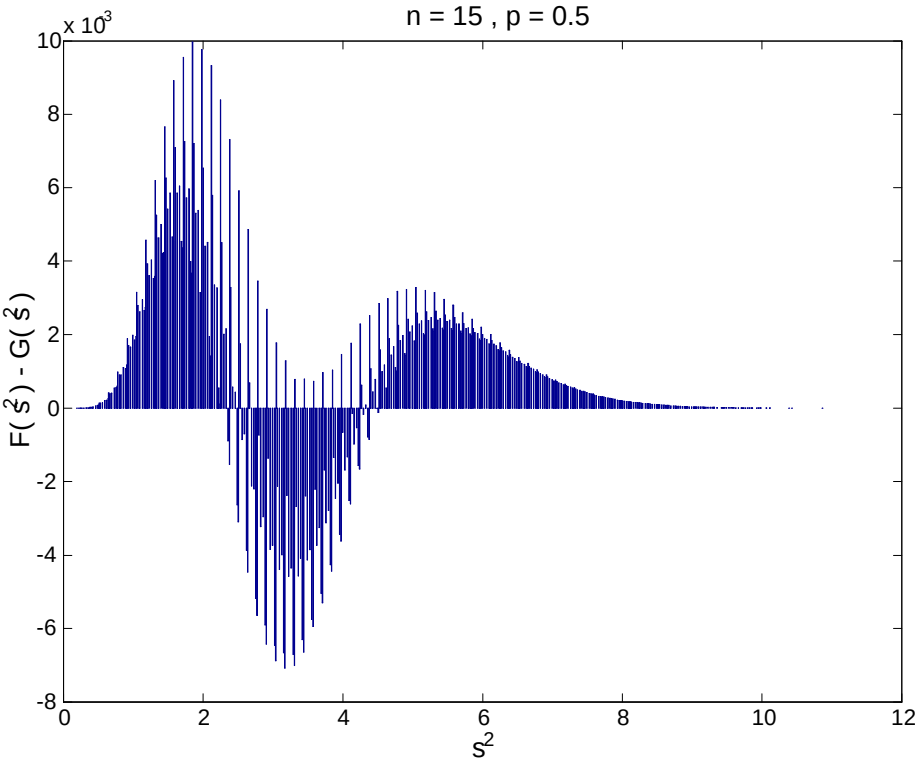

Figure 6. The differences between the simulated distribution function of S^2 and the adjusted gamma approximation plotted against s^2 for $n = 15$ and $p = 0.5$.

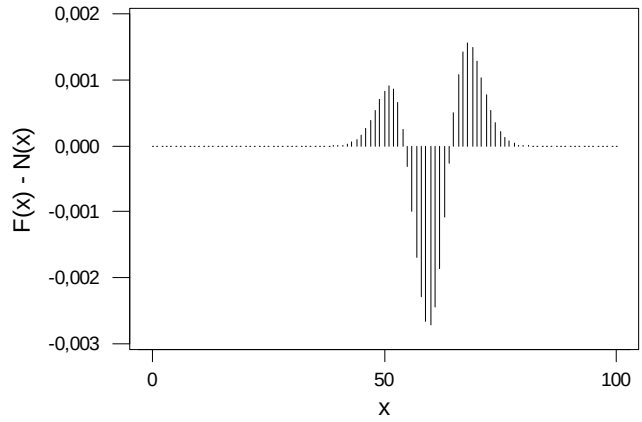

Figure 7. The differences between the exact and the normal approximated distribution function of X, $X \sim Bin(100,0.6)$.

n					
7		$r = 5$	$r = 7$	$r = 9$	$r = 10$
		.0319 (.0051)	.0421 (.0069)	.0361 (.0085)	.0267 (.0087)
10		$r = 5$	$r = 10$	$r = 17$	$r = 22$
		.0103 (.0103)	.0125 (.0030)	.0116 (.0036)	.0117 (.0035)
12		$r = 5$	$r = 10$	$r = 15$	$r = 30$
		.0118 (.0064)	.0052 (.0038)	.0077 (.0028)	.0111 (.0034)
15		$r = 5$	$r = 10$	$r = 20$	$r = 50$
		.0201 (.0075)	.0041 (.0029)	.0034 (.0028)	.0062 (.0024)

Table 6. The Kolmogorov distances for S^2 in the uniform(n, r) model.. The distances in the upper deciles of $G\,(s^2)$ are given within parenthesis.

The results in Table 6 show that the gamma approximation to the distribution function of S^2 in uniform random graphs is better than the corresponding approximation in Bernoulli graphs. This is explained by the smoothness of the distribution in uniform random graphs, and can be seen from Figure 8.

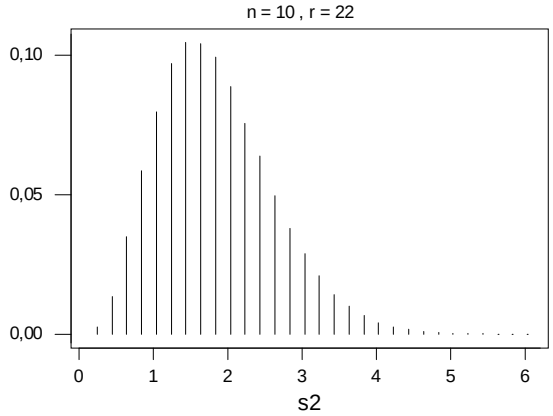

Figure 8. The simulated distribution of S^2 for $n = 10$ and $r = 22$.

	$p = 0.1$				$p = 0.5$		
z	F	$F - G^*$	$F - G$		F	$F - G^*$	$F - G$
0	0.0532	0.0532	0.0003		0.0002	0.0002	0.0001
12	0.2510	0.0743	0.0263		0.0067	0.0054	0.0043
16	0.4095	0.1357	-0.0088		0.0197	0.0156	0.0058
28	0.5515	**0.0036**	**-0.0350**		0.0594	0.0266	0.0180
32	0.7273	0.1047	0.0107		0.1050	0.0537	0.0166
44	0.8136	0.0233	0.0027		0.1656	0.0298	0.0119
48	0.8690	0.0394	-0.0072		0.2377	0.0653	0.0047
60	0.9193	0.0084	-0.0009		0.3291	0.0309	0.0085
64	0.9556	0.0269	0.0063		0.4107	0.0675	-0.0003
76	0.9666	0.0025	-0.0014		0.4802	**0.0027**	**-0.0189**
80	0.9803	0.0088	0.0003		0.5612	0.0409	-0.0203
92	0.9882	0.0022	0.0006		0.6432	**0.0051**	**-0.0127**
96	0.9936	0.0045	0.0012		0.7101	0.0370	-0.0110
108	0.9962	0.0015	0.0009		0.7584	**-0.0053**	**-0.0184**
112	0.9979	0.0020	0.0008		0.8091	0.0198	-0.0144
124	0.9985	0.0004	0.0002		0.8523	**-0.0007**	**-0.0096**
128	0.9992	0.0007	0.0002		0.8895	0.0192	-0.0035
140	0.9996	0.0003	0.0002		0.9139	**0.0018**	**-0.0039**
144	0.9998	0.0004	0.0002		0.9372	0.0141	-0.0002
156	0.9999	0.0002	0.0001		0.9555	0.0063	0.0029
160	0.9999	0.0001	0.0001		0.9692	0.0133	0.0047
172	1.0000	0.0001	0.0000		0.9765	0.0050	0.0030
176	1.0000	0.0000	0.0000		0.9846	0.0092	0.0042
188	1.0000	0.0000	0.0000		0.9902	0.0059	0.0047
192	1.0000	0.0000	0.0000		0.9941	0.0076	0.0048
204	1.0000	0.0000	0.0000		0.9962	0.0046	0.0040
208	1.0000	0.0000	0.0000		0.9979	0.0051	0.0036
220	1.0000	0.0000	0.0000		0.9988	0.0032	0.0029
224	1.0000	0.0000	0.0000		0.9995	0.0033	0.0024
236	1.0000	0.0000	0.0000		0.9997	0.0020	0.0019
240	1.0000	0.0000	0.0000		0.9999	0.0019	0.0015
252	1.0000	0.0000	0.0000		1.0000	0.0012	0.0008
272	1.0000	0.0000	0.0000		1.0000	0.0005	0.0003
300	1.0000	0.0000	0.0000		1.0000	0.0001	0.0001

Table 7. Gamma approximation G with and G^* without continuity correction in the $Bernoulli\,(n, p)$-model of order $n = 8$.

Finally, in Table 7 we have compared $F(s^2) - G^*(s^2)$ with $F(s^2) - G(s^2)$ for every possible value of S^2 in the $Bernoulli(n,p)$-model of order $n = 8$. The points of the distributions where $|F(s^2) - G(s^2)| > |F(s^2) - G^*(s^2)|$ are given in bold. For $p = 0.1$, we see that $G^*(s^2)$ is better than $G(s^2)$ for $s^2 = 28/64$ only. For $p = 0.5$, $G^*(s^2)$ is better than $G(s^2)$ in 5 of the 34 possible values of S^2. Other values of p for $n = 8$ give similar results. The conclusion is that we in general improve the approximation by using the continuity correction for $n = 8$. Further simulations show that the importance of the continuity correction decreases when n increases. It seems that the fraction of points where $|F(s^2) - G(s^2)| > |F(s^2) - G^*(s^2)|$ increases with n. However, in the calculations and simulations preformed for $n \leq 15$, $\max_{s^2} |F(s^2) - G(s^2)| < \max_{s^2} |F(s^2) - G^*(s^2)|$ for both the models.

7 Application to Graph Centrality Testing

Several measures of graph centrality have been developed over the years, see for example Freeman (1978), Snijders (1981a) and Wasserman&Faust (1994). One of them is based on the degree variance, and this centrality of a graph can be interpreted as a measure of vertex heterogeneity in the graph. By using the approximate probability distribution of S^2, we can assess a critical value to test the hypothesis of no centrality against alternatives with centrality. The null hypothesis will be rejected if the observed value s^2 is large enough, that is, if the probability value $P(S^2 \geqslant s^2 | H_0)$ is small enough.

Note that we have to reject the $Bernoulli(n,p)$-model and the $Uniform(n,r)$-model if the null hypothesis is rejected. In fact the null hypothesis of no centrality is modeled by the Bernoulli graph and the uniform random graph. But, departures from these graphs does not imply centrality. Centrality here means, that the degrees vary more than they do in Bernoulli or uniform graphs. If the degrees vary less, as they do for instance in regular graphs, then this is not considered to be a violation of H_0. Below we apply the outlined method to a well known network.

One part of the network data compiled by Padgett, (Padgett & Ansell (1993)) consist of marriage relations among 16 families in the 15th century Florence, Italy. In Figure 9 we have drawn an edge between a pair of vertices i.e. a pair of families if a member of one family marries a member of the other. A more detailed description of the network can be found in Wasserman & Faust (1994).

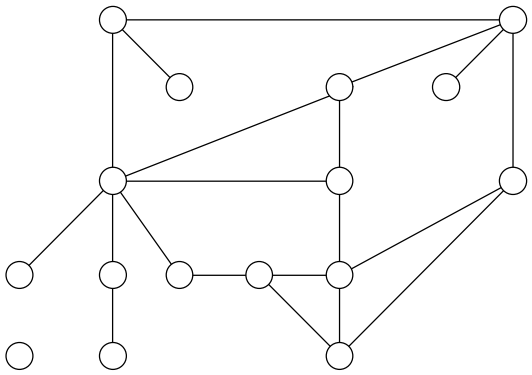

Figure 9. Marital relations between Padgett's Florentine families.

The statistics of the network are:

$$n = 16 \ , \ r = 20 \ , \ \widehat{p} = \frac{1}{6} \text{ and } s^2 = 2.125 \ .$$

To estimate the α and β parameters, we use (5.6) for the Bernoulli $(n, \widehat{p})$ model and (5.15), (5.16) for the uniform(n, r) model. We also apply the continuity corrections given by (5.14) and (5.17).

If we assume that the edges are generated according to a Bernoulli$(n, \widehat{p})$ model, S^2 is approximately $Gamma(6.9767 \ , \ 0.2613)$ and $P\left(S^2 \geqslant 2.125 - \frac{2}{256}\right) = 0.30$. This is no strong evidence against the hypothesis of no graph centrality. However, since the true value of p might differ from the estimate $\widehat{p}$, the probability value is uncertain.

To see how the uncertainty about p might affect our conclusions, we calculate a approximate 95% confidence interval for p according to

$$\widehat{p} \pm 2\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{N}}$$

and obtain the interval (0.10 , 0.23).

From the lower endpoint of the interval we have

$$\widehat{\alpha} = 5.6474 \ , \ \widehat{\beta} = 0.2066 \ , \ P\left(S^2 \geqslant 2.125 - \frac{2}{256}\right) = 0.04$$

and from the upper endpoint we have

$$\widehat{\alpha} = 7.7085 \ , \ \widehat{\beta} = 0.3058 \ , \ P\left(S^2 \geqslant 2.125 - \frac{2}{256}\right) = 0.57.$$

Hence, the two endpoints yield two different conclusions about graph centrality and indicate that the $Bernoulli(n,p)$-model might be inappropriate for graphs of low order.

If we assume a $Uniform(n,r)$-model, S^2 is approximately $Gamma(8.8218$, $0.2084)$ and $P\left(S^2 \geqslant \frac{544-16}{256}\right) = 0.32$. Under this model, there is no strong evidence against the hypothesis that there is no graph centrality. Wasserman and Faust (1994) come to the same conclusion in some of their investigations of these data. They also applied other models and got other findings about the structure of this network.

.

References

[1] Cohen, B., Eades, P., Ruskey, F. & Scott, A. (1994). *Alley CATs in Search of Good Homes.* Paper presented at 25th S.E. Conference on Combinatorics, Graph Theory, and Computing, Congressus Numerantium, 102: 97-110. Available at http://www.csr.uvic.ca/~fruskey/Publications/AlleyCat.html

[2] Deo, N. (1974). *Graph Theory With Applications to Engineering and Computer Science.* New Jersey: Prentice-Hall.

[3] Freeman, L. (1978). *Centrality in Social Networks: Conceptual Clarification.* Social Networks, 1: 215-239.

[4] Hagberg, J (2000). *Centrality Testing and the Distribution of the Degree Variance in Bernoulli Graphs.* Stockholm University, Department of Statistics.

[5] Hagberg, J (2003a). *General Moments of Degrees in Random Graphs.* Stockholm University, Department of Statistics.

[6] Hagberg, J (2003b). *extreme Values and Other Attained Values of the Degree Variance in Graphs.* Stockholm University, Department of Statistics.

[7] Harary, F. (1969). *Graph Theory*. Reading: Perseus Books.

[8] Johnson, N. & Kotz, S. (1970). *Continuous Univariate Distributions-1*. Boston: Houghton Mifflin.

[9] Johnson, N., Kotz, S. & Kemp, A. (1992). *Univariate Discrete Distributions, 2nd ed*. New York: Wiley.

[10] Padgett, J.F. & Ansell, C.K. (1993). *Robust action and the rise of the Medici, 1400-1434*. American Journal of Sociology. 98, 1259-1319.

[11] Sloane, N. & Plouffe, S. (1995). *The Encyclopedia of Integer Sequences*. San Diego: Academic Press.

[12] Snijders, T. (1981a). *The Degree Variance: An Index of Graph Heterogeneity*. Social Networks, 3: 163-174

[13] Snijders, T. (1981b). *Maximum Value and Null Moments of the Degree Variance*. TW-report 229. Department of Mathematics, University of Groningen.

[14] Snijders, T. (2002). ZO 2.3. Available at http://stat.gamma.rug.nl/stocnet/.

[15] Wang, B.Y. & Zhang, F. (1998). *On the Precise Number of (0,1)- matrices in $\mathfrak{A}$(R,S)*. Discrete Mathematics, 187, 211-220.

[16] Wasserman, S. & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. New York Cambridge University Press.

Table 8. The possible values of the degree variance times n^2 for $n = 9, 10, 11, 12$.

$n = 9$:
0 8 14 18 20 26 32 36 38 44 50 54 56 62 68 72 74 80 86 90 92 98 104 108 110 116 122 126 128 134 140 144 146 152 158 162 164 170 176 180 182 188 194 198 200 206 210 216 218 224 230 234 236 242 248 252 254 260 266 270 272 278 284 288 290 296 302 306 308 314 320 324 326 332 338 342 344 350 356 360 362 368 374 378 380 386 392 396 398 404 410 414 416 422 428 432 434 446 450 458 470 504

$n = 10$:
0 16 20 24 36 40 44 56 60 64 76 80 84 96 100 104 116 120 124 136 140 144 156 160 164 176 180 184 196 200 204 216 220 224 236 240 244 256 260 264 276 280 284 296 300 304 316 320 324 336 340 344 356 360 364 376 380 384 396 400 404 416 420 424 436 440 444 456 460 464 476 480 484 496 500 504 516 520 524 536 540 544 556 560 564 576 580 584 596 600 604 616 620 624 636 640 644 656 660 664 676 680 684 696 700 704 716 724 744 756 784

$n = 11$:
0 10 18 22 24 28 30 32 40 44 46 50 52 54 62 66 68 72 74 76 84 88 90 94 96 98 106 110 112 116 118 120 128 132 134 138 140 142 150 154 156 160 162 164 172 176 178 182 184 186 194 198 200 204 206 208 216 220 222 226 228 230 238 242 244 248 250 252 260 264 266 270 272 274 282 286 288 292 294 296 304 308 310 314 316 318 326 330 332 336 338 340 348 352 354 358 360 362 370 374 376 380 382 384 392 396 398 402 404 406 414 418 420 424 426 428 436 440 442 446 448 450 458 462 464 468 470 472 480 484 486 490 492 494 502 506 508 512 514 516 524 528 530 534 536 538 546 550 552 556 558 560 568 572 574 578 580 582 590 594 596 600 602 604 612 616 618 622 624 626 634 638 640 644 646 648 656 660 662 666 668 670 678 682 684 688 690 692 700 704 706 710 712 714 722 726 728 732 734 736 744 748 750 754 756 758 766 770 772 776 778 780 788 792 794 798 800 802 810 814 816 820 822 824 832 836 838 842 844 846 854 858 860 864 866 868 876 880 882 886 888 890 898 902 904 908 910 912 920 924 926 930 932 934 942 946 948 952 954 956 964 968 970 974 976 978 986 990 992 996 998 1000 1008 1012 1014 1018 1020 1022 1030 1034 1036 1040 1044 1052 1056 1058 1062 1064 1066 1074 1086 1106 1110 1124 1152 1176

$n = 12$:
0 20 24 32 36 44 48 56 60 68 72 80 84 92 96 104 108 116 120 128 132 140 144
152 156 164 168 176 180 188 192 200 204 212 216 224 228 236 240 248 252
260 264 272 276 284 288 296 300 308 312 320 324 332 336 344 348 356 360
368 372 380 384 392 396 404 408 416 420 428 432 440 444 452 456 464 468
476 480 488 492 500 504 512 516 524 528 536 540 548 552 560 564 572 576
584 588 596 600 608 612 620 624 632 636 644 648 656 660 668 672 680 684
692 696 704 708 716 720 728 732 740 744 752 756 764 768 776 780 788 792
800 804 812 816 824 828 836 840 848 852 860 864 872 876 884 888 896 900
908 912 920 924 932 936 944 948 956 960 968 972 980 984 992 996 1004 1008
1016 1020 1028 1032 1040 1044 1052 1056 1064 1068 1076 1080 1088 1092
1100 1104 1112 1116 1124 1128 1136 1140 1148 1152 1160 1164 1172 1176
1184 1188 1196 1200 1208 1212 1220 1224 1232 1236 1244 1248 1256 1260
1268 1272 1280 1284 1292 1296 1304 1308 1316 1320 1328 1332 1340 1344
1352 1356 1364 1368 1376 1380 1388 1392 1400 1404 1412 1416 1424 1428
1436 1440 1448 1452 1460 1464 1472 1476 1484 1488 1496 1500 1508 1512
1520 1524 1532 1548 1560 1568 1584 1592 1620 1652 1728