Research Report

Department of Statistics

No. 2000:5

Testing Centrality in Random Graphs

Christian Tallberg

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

Testing Centrality in Random Graphs

Christian Tallberg*

Abstract

When testing centrality in random graphs it is of importance to specify models that capture the irregularities in the structure due to centrality. In this paper we propose using the well-known block models in an attempt to capture such irregularities. The baseline model, revealing no centrality structure, used in this paper is the Bernoulli model. It is shown that the maximum likelihood estimators of the parameters in the block model are tedious to obtain, and that the distribution of the likelihood ratio is difficult to derive analytically. Therefore, various tests of centrality in random graphs are presented where the power functions of the test quantities are estimated by performing computer simulations. The tests are based on centrality indices that are evaluated at actor level. These indices are then aggregated across all actors in order to obtain a centrality index at group level. Two of the tests proposed are based on degree and eight of them are based on distance. None of the tests is uniformly most powerful. The tests where the group level index is defined as an average of the actor level indices show poor power compared to the tests that indicate the variability of the actor level indices. Among the tests based on variability of the actor level indices, the test quantities that include the maximum of the actor level indices generate a higher power than the tests based on the variance of the actor level indices.

Keywords: Random block models, Bernoulli graph, Degree centrality, Closeness centrality, Power of centrality tests

*Department of Statistics, Stockholm University. E-mail: Christian.Tallberg@stat.su.se. The author would like to thank Ove Frank and Mattias Villani for helpful comments. Partial financial support from the Swedish Council of Research in Humanities and Social Sciences (HSFR), grant No. F0750/96, is gratefully acknowledged.

1 Introduction

1.1 Social network modeling

In an association of members it may be desireable to possess the knowledge if any of them can be considered as more popular, important or influencial in some aspect. If we can ascertain that this is the case, we may then wish to inform ourselves of some properties or characteristics of these members for different reasons. One reason could be just of descriptive nature; we want to examine the background characteristics of these central members. If we have several associations of similar kind the next step would be to compare the background characteristics of the central members in the associations in order to detect similarities and differences.

In this paper block models are presented in an attempt to capture diversions among members in a social network due to centrality. When block modeling is mentioned in social network literature the approach is usually slightly different. There are two main approaches to block modeling. In the first one, which is primarily descriptive, the focus is to capture some of the general features of a network's structure by partitioning members with the same or similar adjacency structure into the same block. In the second approach, stochastic block modeling, actors in the same block must not necessarily have the same or similar realized structure, but the probability distribution over the set of possible structures is the same. These models are described by Holland, Laskey and Leinhardt (1983), Wang and Wong (1987) and Wasserman and Anderson (1987). Frank, Hallinan and Nowicki (1985), Frank, Komanska and Widaman (1985) used the block modeling technique as a data analytic method to reduce the number of parameters in log-linear models. However, the purpose of this paper is to show that block modeling can be a useful technique in testing centrality in social networks.

In Section 2 the stochastic model assumed under non-centrality is described. Moreover, the idea behind the concept block modeling within the context of centrality is treated and some desired properties induced by centrality are discussed. In Section 3 maximum likelihood estimators are derived, both without any restrictions on the parameters and with restrictions that one would expect under centrality structure. In Section 4 ten test quantities for testing centrality structure are described, and in Section 5 the power functions of the tests are estimated by simulation methods.

1.2 Notation

The members and their relation structure are represented by graphs, where the members are referred to as vertices and the relations are referred to as edges.

Let G be an undirected graph without loops where n, the order of the graph, is known. Let $V = \{1, ..., n\}$ be the vertex set of G and E the edge set. Further, we let A be the corresponding $n \times n$ adjacency matrix of G where the elements are given by

$$a_{ij} = a_{ji} = \begin{cases} 1 & \text{if there exists an edge between vertex } i \text{ and vertex } j,\ i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

In the context of graph theory the number of edges in G is often referred to as the size of G and it is given by

$$|E| = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij}.$$

The degree of any vertex i in G is defined as the number of edges incident to vertex i and is denoted by a_i, thus

$$a_i = \sum_{j=1}^{n} a_{ij} \ = \sum_{j=1}^{n} a_{ji}.$$

The maximum degree in G is denoted $\max_i a_i$ and the mean degree of G is denoted $\bar{a}$ and given by

$$\bar{a} = \frac{1}{n} \sum_{i=1}^{n} a_i = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}.$$

Finally the variance of the degrees is denoted s_a^2 and given by

$$s_a^2 = \frac{1}{n} \sum_{i=1}^{n} (a_i - \bar{a})^2 .$$

A *walk* in a graph is an alternating sequence of vertices and edges, starting and ending with vertices, in which each edge is incident with the vertices following and preceding it in the sequence. Vertices and edges may be repeated. A *path* is a walk in which all vertices and edges are distinct. The *length* of a path is the number of edges used. The *geodesic* is the shortest

path between two vertices. If any two vertices are connected by a walk the graph is said to be connected. The maximal connected subgraphs of a disconnected graph are called the connected components of the graph. Let D be an $n \times n$ distance matrix of G where the element d_{ij} is defined as the length of the geodesic between vertex i and vertex j. If there is no geodesic between vertex i and vertex j, i.e. the vertices are in two different components, $d_{ij} = \infty$. A consequence is that mathematical functions of the distance such as the average distance can only be evaluated in connected components. In a connected graph, the average distance is given by

$$\bar{d} = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}.$$

If the graph is disconnected the average distance can be calculated in each connected component. For a more detailed description of concepts in graph theory see Palmer (1985), or the extensive work on social networks by Wasserman and Faust (1994).

2 Centrality and block models

2.1 Centrality concepts

Three centrality concepts are usually mentioned in the literature, see for example Freeman (1979), Wasserman and Faust (1994), namely degree centrality, closeness centrality and betweenness centrality. All three are in different ways trying to capture the popularity, influence or importance of members or actors in a social network. The choice of an appropriate centrality concept and its associated measure depends on the context of the application. According to Freeman (1979) degree-based measures are indexes of an actors communication activity, betweenness-based measures are useful if we are interested in an actors control of communication, and closeness-based measures indicates the level of an actors independence or efficiency.

Centrality indices are evaluated for all n actors indicating the status of each actor according to popularity, influence etc.. To obtain a centrality index at group level the centrality indices are aggregated across all actors. There are different measures at group level such as the maximum, the average, or the variance of the actor level indices. By using average as measure we define group level centrality as the compactness of the network. According to Freeman (1979) the group level index should be high if one single actor is more central than all the other actors. Therefore an appropriate group level

index should capture the variability of the actor centrality indices. In this paper a social network is considered as central if the actors can be partitioned into blocks where the probability of an edge between two actors is distinct in different blocks. If this is the case, we should obtain a larger variability of the centrality indices at actor level.

Centrality based on degree focuses on adjacency; an actor adjacent to a large number of actors in the network is considered as central. Hence, degree is the actor level centrality measure. Group level centrality measures are for instance the difference between the maximum degree and the average degree, which is suggested in this paper, and the variance of the degrees recommended by Snijders (1981a, 1981b).

Centrality based on geodesic distances is more dependent on indirect ties compared to centrality based on degree which only involves direct ties. The point of using distance as a measure of centrality is that an actor will be considered as central if he can interact quickly with all others. A desireable property of a distance-based centrality index is, that it decreases for actor i when the distance to the other actors increases. Such an index, based on the inverse of the sum of the distances from actor i to all the other actors, was proposed by Sabidussi (1966). The disadvantage with Sabidussi's index is that it can only be evaluated for connected graphs, since the distance between two disconnected vertices is set to infinity. In this paper similar measures are introduced with weights allowing us to handle distance in disconnected graphs.

The third centrality concept, also based on geodesics, is betweenness centrality. If actor j and actor k has to pass actor i in order to reach each other on their shortest path, then actor i is defined as being between actor j and actor k. Freeman (1977) introduced a centrality index based on betweenness where all geodesics between actor j and actor k are equally likely to be used. The number of geodesics between actor j and actor k is denoted g_{jk} and the probability of using any one is g_{jk}^{-1}. The betweenness index for actor i is then evaluated as the sum of the probabilities of choosing the geodesics he is contained in,

$$\sum_{j<k} \frac{g_{jk}(i)}{g_{jk}}$$

for i distinct from j and k, where $g_{jk}(i)$ is the number of geodesics from j to k containing actor i. A group index can be obtained as an average or as a variability measure in analogy with the definitions for the other centrality concepts.

2.2 Models

2.2.1 The Bernoulli model

As our stochastic nullhypothesis model showing no centrality structure we will use the *Bernoulli* (p) *model*. That is, the edges of G are generated independently with an unknown common edge probability $p = P(a_{ij} = 1)$, $i \neq j$. The probability of G is given by

$$P(G) = \begin{cases} p^r (1-p)^{\binom{n}{2}-r}, & \text{if } G \text{ is of order } n \text{ with } r \text{ edges} \\ 0, & \text{otherwise.} \end{cases}$$

This model, commonly used to study random graph properties, see for instance Palmer (1985), is in this paper considered as a baseline model for testing centrality.

2.2.2 Block models with centrality

As an alternative random graph model that captures centrality against which we can test the *Bernoulli* (p) model, we use a block model. Block modeling means that we have a graph of fixed order n where V is partitioned into K mutually exclusive non-empty vertex blocks, $V_1, ..., V_K$. Each pair of vertices in the graph is independently given an edge with a probability that depends on which blocks the vertices belong to.

Let p_{kl} be the edge probability between any vertex i in block k and any vertex j in block l, for all k, l. The block with the largest edge probability, $\max_k p_{kk}$, is defined as the central block. Moreover, we let E_{kl} denote the set of edges between any vertex i in block k and any vertex j in block l.

Since there are K blocks in G, there are $K + \binom{K}{2}$ edge probabilities (K within blocks and $\binom{K}{2}$ between blocks) of which some may be equal. Note that a minimal requirement for the model to be a block model, is that not all the p_{kk} is equal. In order to estimate these probabilities we also need to identify the vertex blocks. If K is large this will be a tedious task even though it can be handled with more computing time. For illustrative purposes when deriving analytical results and performing simulation studies, the number of blocks is limited to be two.

In the case $K = 2$, centrality block models with block 1 as the central block is defined by

$$\begin{cases} p_{11} > p_{22} \\ p_{11} \geq p_{12} \end{cases} \tag{1}$$

or

$$\begin{cases} p_{11} > p_{22} \\ p_{12} \geq p_{22} \end{cases} . \tag{2}$$

In (1) we don't allow the edge probability between the blocks to be larger than the edge probability within the central block, whereas in (2) the edge probability between the blocks is not allowed to be smaller than the edge probability within the non-central block. By applying restrictions on the probabilities, we will obtain some desired properties in a graph that is considered as central.

Lemma 1 *The expected degree of any central vertex is larger than the expected degree of any non-central vertex.*

The proof is given only when the restrictions under the first formulation hold. A proof when the restrictions under the second formulation hold can be obtained in a similar way. Some further notation and remarks are in order before the proof.

Let $V_1 = \{1, ..., \ n_1\}$ be the vertex set in G consisting of central vertices with a common edge probability p_{11}. Let $V_2 = \{n_1 + 1, ..., \ n_1 + n_2\}$, $n_1 + n_2 = n$, be the vertex set in G consisting of non-central vertices with a common edge probability p_{22}, where $V = V_1 \cup V_2$ and $V_1 \cap V_2 = \phi$. Let p_{12} be the common edge probability between any vertex i in V_1 and any vertex j in V_2.

According to the $Bernoulli(p)$ model the degree of vertex i, $a_i, i = 1, ..., n,$ is a random variable

$$a_i = \sum_{j \in V} a_{ij} = a_{iV}$$

and its distribution is $Bin(n-1, p)$. Note that a_i are not independent since $\sum_{i=1}^{n} a_i = 2r$.

An arbitrarily chosen vertex i in V can be adjacent to vertices in both V_1 and V_2 and its degree can be obtained as the sum of the following two numbers:

- a_{iV_1} = *the number of edges from vertex i to vertices in V_1*,

$$a_{iV_1} = \sum_{j \in V_1} a_{ij}, \quad i \in V$$

where

$$\begin{cases} a_{iV_1} \sim Bin\left(n_1 - 1, p_{11}\right), \ i \in V_1 \\ a_{iV_1} \sim Bin\left(n_1, p_{12}\right), \ i \in V_2 \end{cases} .$$

- a_{iV_2} = *the number of edges from vertex* i *to vertices in* V_2,

$$a_{iV_2} = \sum_{j \in V_2} a_{ij}, \quad i \in V$$

where

$$\begin{cases} a_{iV_2} \sim Bin\left(n_2, p_{12}\right), \ i \in V_1 \\ a_{iV_2} \sim Bin\left(n_2 - 1, p_{22}\right), \ i \in V_2 \end{cases}.$$

Proof. We shall prove that

$$\begin{aligned} E(a_{iV_1}) + E(a_{iV_2}) &> E(a_{jV_2}) + E(a_{jV_1}) \text{ for } i \in V_1 \textit{ and } j \in V_2, \text{ that is} \\ (n_1 - 1)\, p_{11} + n_2 p_{12} &> (n_2 - 1)\, p_{22} + n_1 p_{12}, \text{ or equivalently} \\ (n_1 - 1)\, p_{11} &> (n_2 - 1)\, p_{22} + (n_1 - n_2)\, p_{12}. \end{aligned} \tag{3}$$

Replacing the left hand side of this inequality with

$$(n_1 - n_2 + n_2 - 1)\, p_{11} = (n_2 - 1)\, p_{11} + (n_1 - n_2)\, p_{11}$$

and applying (1) we see that the inequality is true if either $n_1 > 1$ or $n_2 > 1$. ■

Note that for $n_1 = 1$ and $n_2 > 1$ we obtain from (3) that $p_{12} > p_{22}$, and for $n_1 > 1$ and $n_2 = 1$ that $p_{11} > p_{12}$.

If V is partitioned into two blocks, V_1 and V_2, there are three probabilities that we need to estimate, namely $P = (p_{11}, p_{22}, p_{12})$. If the blocks are unknown we also need to estimate V_1 and V_2.

3 Maximum likelihood estimators of centrality structure

3.1 Maximum likelihood estimators if the blocks are known

Let G be a random graph of a fixed and known order n consisting of K blocks. Note that an important assumption in this paper is that the order of the blocks n_k, where $\sum_{k=1}^{K} n_k = n$, is also considered as fixed and known. The number of edges between block k and block l is denoted r_{kl}, and $r = \sum_{k=1}^{K} \sum_{l=k}^{K} r_{kl}$ is the number of edges in G.

If the blocks are known then it is possible to observe r_{kl} and obtain the maximum likelihood estimators of the edge probabilities P. Since independent dyads are assumed, i.e. the edges between different pairs of individuals are assumed stochastically independent, the likelihood for a realization of G is

$$L(P) = \left[\prod_{k=1}^{K} \prod_{l=k}^{K} p_{kl}^{r_{kl}} (1 - p_{kl})^{n_{kl} - r_{kl}} \right],$$

where

$$n_{kl} = \begin{cases} \binom{n_k}{2} \text{ for } k = l \\ n_k n_l \text{ for } k \neq l \end{cases}$$

and $P = (p_{kl})$ and $R = (r_{kl})$.

It can easily be proved that in this case the maximum likelihood estimator of each parameter p_{kl} is

$$\hat{p}_{kl} = \frac{r_{kl}}{n_{kl}} \text{ for all } k, l. \tag{4}$$

3.2 Maximum likelihood estimators if the blocks are unknown

3.2.1 Without restrictions on the parameters

When $K = 2$ and the vertex sets are unknown we have to estimate the three probabilities $P = (p_{11}, p_{22}, p_{12})$ and one partition into two blocks, V_1 and V_2. The likelihood is now a function of P and V_1 (since $V_2 = V \cap \bar{V}_1$), where V_1 is latent, and the likelihood is dependent on V_1 only through r_{11}. Therefore, r_{kl} is now seen as a function of V_1, but for notational simplicity we will drop V_1, and denote $r_{kl}(V_1)$ as r_{kl} in most of the paper. Note that in a realization of a random graph the elements of A, a_{ij}, are observable and therefore $r = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij}$ is observable, but since V_1 and V_2 not are known each edge frequency r_{11}, r_{12} and r_{22} is unobservable. The likelihood function for $K = 2$ with unknown V_1 is given by

$$\begin{aligned} L(P, V_1) &= p_{11}^{r_{11}} (1 - p_{11})^{n_{11} - r_{11}} p_{22}^{r_{22}} (1 - p_{22})^{n_{22} - r_{22}} \\ &\quad \times p_{12}^{r - r_{11} - r_{22}} (1 - p_{12})^{n_{12} - r + r_{11} + r_{22}}. \end{aligned}$$

We want to obtain the maximum likelihood estimators of P and V_1 simultaneously. Similar problems are dealt with by Bock (1996) and Jansson

(1997) where the maximum of the likelihood function is determined iteratively. Bock mentioned the procedure as the "classification maximum likelihood approach" and maximized the likelihood over clusters and parameters.

Two approaches to obtain the maximum likelihood estimators are discussed here. In the first approach the likelihood function is considered for fixed P, where it depends on V_1 only through r_{11} and r_{22}. If the factors that are not depending on r_{11} and r_{22} are put into a constant, we can write the likelihood as

$$L(P, V_1) = \left(\frac{p_{11}}{1-p_{11}}\frac{1-p_{12}}{p_{12}}\right)^{r_{11}} \left(\frac{p_{22}}{1-p_{22}}\frac{1-p_{12}}{p_{12}}\right)^{r_{22}} \times \textit{constant.}$$

We see that the maximum of $L(P, V_1)$ is depending on P. Without the restrictions on P given in (1), there are numerous ways to choose V_1 as any set of n_1 vertices that maximizes $L(P, V_1)$. Even with restrictions on P the maximum likelihood estimator of V_1 is not unique. Examples of graphs where there exist no unique maximum likelihood estimator of V_1 are given in Section 3.2.2 and illustrated in Figure 2.

A more feasible approach is to consider V_1 as fixed and the likelihood as a function of P without restrictions. $L(P, V_1)$ is maximized over P by $\hat{P}$ given by (4), where $L\left(\hat{P}, V_1\right)$ depends on V_1 (and indirectly on V_2) only through $R = (r_{11}, r_{12}, r_{22})$. The likelihood function is given by

$$\begin{aligned} L\left(\hat{P}, V_1\right) &= \left(\frac{r_{11}}{n_{11}}\right)^{r_{11}} \left(1-\frac{r_{11}}{n_{11}}\right)^{n_{11}-r_{11}} \left(\frac{r_{22}}{n_{22}}\right)^{r_{22}} \left(1-\frac{r_{22}}{n_{22}}\right)^{n_{22}-r_{22}} \\ &\quad \times \left(\frac{r-r_{11}-r_{22}}{n_{12}}\right)^{r-r_{11}-r_2} \left(1-\frac{r-r_{11}-r_{22}}{n_{12}}\right)^{n_{12}-r+r_{11}+r_{22}}. \end{aligned}$$

The behaviour of the likelihood function is investigated through the differentials of the log likelihood function. For notational simplicity we denote $L\left(\hat{P}, V_1\right)$ as $\hat{L}$ in the rest of this section. In order to obtain the differentials we use the entropy function

$$h(p) = -(p \ln p + q \ln q), \; p + q = 1.$$

Then the likelihood function can be expressed as

$$\ln \hat{L} = -\left(n_{11} h\left(\frac{r_{11}}{n_{11}}\right) + n_{22} h\left(\frac{r_{22}}{n_{22}}\right) + n_{12} h\left(\frac{r-r_{11}-r_{22}}{n_{12}}\right)\right)$$

The first differentials are

$$\begin{aligned}
\frac{\partial \ln \hat{L}}{\partial r_{11}} &= -h'\left(\frac{r_{11}}{n_{11}}\right) + h'\left(\frac{r - r_{11} - r_{22}}{n_{12}}\right) = \ln \frac{\frac{r_{11}}{n_{11}}}{1 - \frac{r_{11}}{n_{11}}} + \ln \frac{1 - \frac{r - r_{11} - r_{22}}{n_{12}}}{\frac{r - r_{11} - r_{22}}{n_{12}}} \\
\frac{\partial \ln \hat{L}}{\partial r_{22}} &= -h'\left(\frac{r_{22}}{n_{22}}\right) + h'\left(\frac{r - r_{11} - r_{22}}{n_{12}}\right) = \ln \frac{\frac{r_{22}}{n_{22}}}{1 - \frac{r_{22}}{n_{22}}} + \ln \frac{1 - \frac{r - r_{11} - r_{22}}{n_{12}}}{\frac{r - r_{11} - r_{22}}{n_{12}}}
\end{aligned}$$

and the second differentials are

$$\begin{aligned}
\frac{\partial^2 \ln \hat{L}}{\partial r_{11}^2} &= -\frac{1}{n_{11}} h''\left(\frac{r_{11}}{n_{11}}\right) - \frac{1}{n_{12}} h''\left(\frac{r - r_{11} - r_{22}}{n_{12}}\right) \\
&= \frac{1}{r_{11}\left(1 - \frac{r_{11}}{n_{11}}\right)} + \frac{1}{(r - r_{11} - r_{22})\left(1 - \frac{r - r_{11} - r_{22}}{n_{12}}\right)} \\
\frac{\partial^2 \ln \hat{L}}{\partial r_{22}^2} &= -\frac{1}{n_{22}} h''\left(\frac{r_{22}}{n_{22}}\right) - \frac{1}{n_{12}} h''\left(\frac{r - r_{11} - r_{22}}{n_{12}}\right) \\
&= \frac{1}{r_{22}\left(1 - \frac{r_{22}}{n_{22}}\right)} + \frac{1}{(r - r_{11} - r_{22})\left(1 - \frac{r - r_{11} - r_{22}}{n_{12}}\right)} \\
\frac{\partial^2 \ln \hat{L}}{\partial r_{11} r_{22}} &= -\frac{1}{n_{12}} h''\left(\frac{r - r_{11} - r_{22}}{n_{12}}\right) = \frac{1}{(r - r_{11} - r_{22})\left(1 - \frac{r - r_{11} - r_{22}}{n_{12}}\right)}.
\end{aligned}$$

Due to the fact that all the second differentials are positive and from the following inequalities

$$\begin{cases} \frac{\partial^2 \ln \hat{L}}{\partial r_{11}^2} > \frac{\partial^2 \ln \hat{L}}{\partial r_{11} r_{22}} \\ \frac{\partial^2 \ln \hat{L}}{\partial r_{22}^2} > \frac{\partial^2 \ln \hat{L}}{\partial r_{11} r_{22}} \end{cases},$$

the matrix of second differentials must be positive definit. Hence, there exists a local minimum. The optimal values giving the minimum of the likelihood is obtained by solving the equation system

$$\begin{aligned}
\frac{\partial \ln \hat{L}}{\partial r_{11}} &= 0 \\
\frac{\partial \ln \hat{L}}{\partial r_{22}} &= 0.
\end{aligned}$$

In an arbitrarily chosen graph the minimum is obtained when the proportions of edges in the central and the non-central blocks both equal the proportion of edges in the graph, $\frac{r_{11}}{n_{11}} = \frac{r_{22}}{n_{22}} = \frac{r}{\binom{n}{2}}$. Since the matrix of second differentials

is positive definite, V_1 and V_2 that maximizes $\hat{L}$ must be chosen such that (r_{11}, r_{22}) is a boundary point. The complexity of determining V_1 and V_2 is illustrated in Figure 1. The subplots to the left show different realizations of assumed block model graphs of order $n = 7$ and with a central block of order $n_1 = 3$. The subplots to their right are the corresponding level curves of the natural logarithm of the likelihood function. In Figure 1a), $\hat{V}_1 = \{2, 3, 7\}$ where $\hat{r}_{11} = 3$ and $\hat{r}_{22} = 5$. That is $\hat{L}$ is maximized when the number of edges within the central block and within the non-central block is maximized in its range in G. In Figure 1b) $\hat{V}_1 = \{2, 3, 6\}$, $\hat{r}_{11} = 3$ and $\hat{r}_{22} = 1$. Thus, $\hat{L}$ obtains its maximum if the number of edges within the central block is maximized in its range and the number of edges within the non-central block is minimized in its range. In Figure 1c) $\hat{V}_1 = \{1, 3, 5\}$, $\hat{r}_{11} = 0$ and $\hat{r}_{22} = 0$. Thus, to obtain the maximum of $\hat{L}$ the number of edges within the central block and the number of edges within the non-central block is minimized in its range. The last case illustrated in Figure 1d) shows that $\hat{L}$ obtains its maximum for two sets of V_1. That is there exists no unique maximum likelihood estimator of V_1. $\hat{L}$ is maximized if the number of edges within the central block is maximized in its range, which is the case for $\hat{V}_1 = \{1, 2, 4\}$ or $\hat{V}_1 = \{3, 5, 6\}$ where $\hat{r}_{11} = 3$ and $\hat{r}_{22} = 4$.

We see that in one realization of G both $\hat{r}_{11}$ and $\hat{r}_{22}$ should be maximized to obtain $\hat{V}_1$ that maximizes the likelihood function. In a second realization of G both $\hat{r}_{11}$ and $\hat{r}_{22}$ should be minimized and in a third realization one of them should be minimized and the other should be maximized to obtain $\hat{V}_1$ that maximizes $\hat{L}$. These examples illustrate that there is no simple rule how to choose $\hat{V}_1$ to maximize the likelihood function. Note that r_{11} and r_{22} are descrete variables and that r_{11} doesn't take all the values in the interval $0 \leqslant r_{11} \leqslant \min(n_{11}, r)$ and r_{22} doesn't take all the values in the interval $0 \leqslant r_{22} \leqslant \min(n_{22}, r)$. In Figure 1 admissible values of (r_{11}, r_{22}) are represented with dots.

3.2.2 With restrictions on the parameters

We now proceed to derive the maximum likelihood estimators with restrictions on the probabilities P. Deriving the maximum likelihood estimators under the restrictions proposed in (1) involves complicated analytical methods. For simplicity, suppose that the edge probability within the central block is the same as the edge probability between the central block and the non-central block. Then the restrictions on the edge probabilities are given

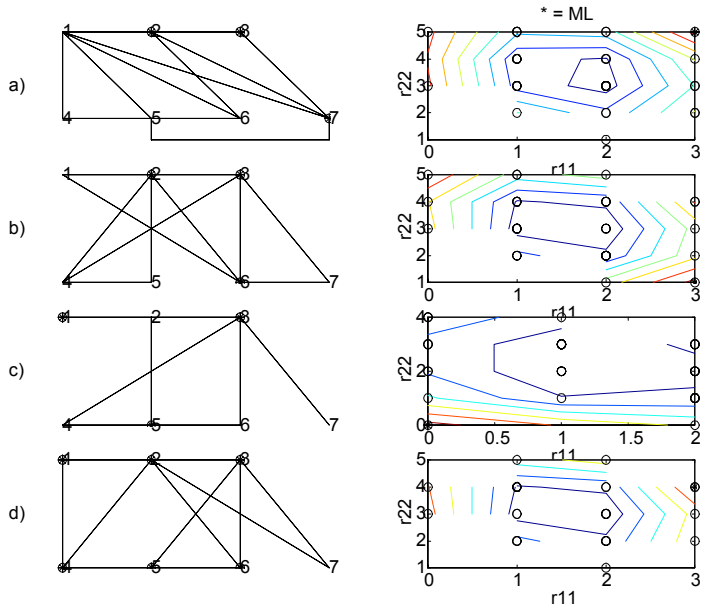

Figure 1: The left subplots show realizations of assumed block model graphs of order $n = 7$ and with a central block of order $n_1 = 3$. The right subplots are the corresponding level curves of the logarithm of the likelihood function.

by

$$\begin{aligned} p_{11} - p_{22} &> 0 \\ p_{11} &= p_{12} \end{aligned} \tag{5}$$

which enables us to derive the maximum likelihood estimators of P, V_1 and V_2. When the number of blocks is restricted to be two it is convenient to reparametrize. If we let $p_{22} = p$ and introduce a new parameter

$$\Delta = p_{11} - p > 0,$$

the block model is reformulated with the two unknown parameters, p and Δ. p is now the edge probability within the non-central block, and $p + \Delta$ is the edge probability within the central block and between the two blocks.

As in Section 3.2.1 there are two approaches to determine the maximum likelihood estimators iteratively. First we consider the likelihood function for fixed p and Δ.

$$L(p, \Delta, V_2) = (p+\Delta)^{r-r_{22}} (1-p-\Delta)^{\binom{n}{2}-r-n_{22}+r_{22}} p^{r_{22}} (1-p)^{n_{22}-r_{22}}.$$

If the factors not depending on r_{22} are put into a constant we can rewrite the likelihood function as

$$\begin{aligned} L(p, \Delta, V_2) &= \left(\frac{p(1-p-\Delta)}{(1-p)(p+\Delta)} \right)^{r_{22}} \times \textit{constant} \\ &= \left(\frac{q-\Delta}{q} \frac{p}{p+\Delta} \right)^{r_{22}} \times \textit{constant}, \ \Delta > 0, \ p+q=1. \end{aligned}$$

Since $\frac{q-\Delta}{q} < 1$ and $\frac{p}{p+\Delta} < 1$, $L(p, \Delta, V_2)$ is a decreasing function of r_{22}. Thus, $\hat{V}_2$ is any set of n_2 vertices with minimal number of edges $\hat{r}_{22} = r_{22}\left(\hat{V}_2\right)$. The objective then is to find p and Δ maximizing $L(p, \Delta, \hat{V}_2)$.

Alternatively, we consider $L(p, \Delta, V_2)$ for fixed V_2 as a function of p and Δ. $L(p, \Delta, V_2)$ is maximized over p and Δ by $\hat{p} = \frac{r_{22}}{n_{22}}$ and $\hat{\Delta} = \frac{r-r_{22}}{\binom{n}{2}-n_{22}} - \frac{r_{22}}{n_{22}}$, where $L\left(\hat{p}, \hat{\Delta}, V_2\right)$ depends on V_2 only through $r_{22}(V_2)$. The likelihood function is then given by

$$\begin{aligned} L\left(\hat{p}, \hat{\Delta}, V_2\right) &= \left(\frac{r-r_{22}}{\binom{n}{2}-n_{22}} \right)^{r-r_{22}} \left(1 - \frac{r-r_{22}}{\binom{n}{2}-n_{22}} \right)^{\binom{n}{2}-r-n_{22}+r_{22}} \\ &\quad \times \left(\frac{r_{22}}{n_{22}} \right)^{r_{22}} \left(1 - \frac{r_{22}}{n_{22}} \right)^{n_{22}-r_{22}}. \end{aligned} \tag{6}$$

From investigating the first and second differentials of $L\left(\hat{p}, \hat{\Delta}, V_2\right)$ we see that it is a convex function with a minimum at $r_{22} = r\frac{n_{22}}{\binom{n}{2}}$. Therefore, $L\left(\hat{p}, \hat{\Delta}, V_2\right)$ must obtain its maximum on the boundary. Due to the restriction $\Delta > 0$, $L\left(\hat{p}, \hat{\Delta}, V_2\right)$ is maximized if $\hat{V}_2$ is any set of n_2 vertices with $\hat{r}_{22}$ edges where $\hat{r}_{22}$ is the lower boundary value of r_{22}, $0 \leqslant \hat{r}_{22} \leqslant r\frac{n_{22}}{\binom{n}{2}}$.

The maximum likelihood estimator of V_2 is not unique. There is usually more than one solution $\hat{V}_2$ that will maximize $L\left(\hat{p}, \hat{\Delta}, V_2\right)$.

An alternative to find V_2 with a minimal r_{22}, is to find V_1 with a maximal $r_{11} + r_{12}$. This is due to the relationship $r_{11} + r_{12} = r - r_{22}$. If the likelihood is expressed as a function of V_1 instead of as a function of V_2, $L\left(\hat{p}, \hat{\Delta}, V_1\right)$ is a convex function with a minimum at $r_{11} + r_{12} = r\left(\frac{\binom{n}{2} - n_{22}}{\binom{n}{2}}\right)$. Then $\hat{V}_1$ maximizing $L\left(\hat{p}, \hat{\Delta}, V_1\right)$ must be any set of n_1 vertices with $\hat{r}_{11} + \hat{r}_{12}$ chosen as the upper boundary value of $r_{11} + r_{12}$ so that $r\left(\frac{\binom{n}{2} - n_{22}}{\binom{n}{2}}\right) \leqslant \hat{r}_{11} + \hat{r}_{12} \leqslant \min\left(\binom{n}{2} - n_{22}, r\right)$.

Note that finding $\hat{V}_1$ with a maximal $\hat{r}_{11} + \hat{r}_{12}$ is not necessarily equal to finding $\hat{V}_1$ with maximum sum of degrees. There is a distinction between the two cases even though they might seem similar. The sum of the degrees of V_1 is $\sum_{i \in V_1} a_i = 2r_{11} + r_{12}$ and hence, $r_{11} + r_{12} = \sum_{i \in V_1} a_i - r_{11}$. For $n_1 = 1$ it follows that $r_{11} = 0$ and the quantities are equal.

For $n_1 = 2$, the procedure of obtaining $\hat{V}_1$ is a bit more complex. Since

$$r_{11} + r_{12} = \begin{cases} \sum_{i \in V_1} a_{i.} - 1, & \text{if } r_{11} = 1 \\ \sum_{i \in V_1} a_{i.}, & \text{if } r_{11} = 0 \end{cases}, \tag{7}$$

the procedure of finding $\hat{V}_1$ can conveniently distinguish two cases:

1. If there are at least two non-adjacent vertices with maximum degree, then we must choose two of them arbitrarily to form $\hat{V}_1$.

2. If we there are at least two adjacent vertices with maximum degree that are reciprocally adjacent, $\hat{V}_1$ must contain one of them. The choice of the other one leaves us with two options; either we chose one with maximum degree adjacent to the first one, or we choose one with one

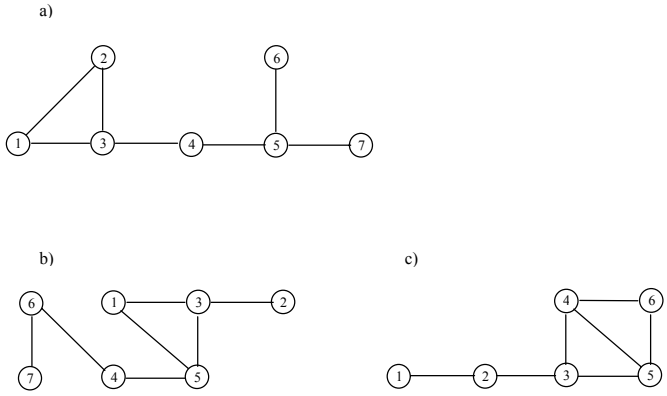

Figure 2:

degree less than maximum degree that is not adjacent to the first one. Both sets of $\hat{V}_1$ will yield the same $\hat{r}_{11} + \hat{r}_{12}$ and therefore maximize the likelihood.

Figure 2a) illustrates how $\hat{V}_1$ is obtained in case 1. Both vertex number 3 and vertex number 5 have maximum degree. If $n_1 = 1$, one of them is chosen arbitrarily for $\hat{V}_1$. That is, $\hat{V}_1 = \{3\}$ or $\hat{V}_1 = \{5\}$. If $n_1 = 2$, the only solution is $\hat{V}_1 = \{3, 5\}$. This is due to the fact that vertex number 3 and vertex number 5 are non-adjacent.

Figure 2b) illustrates how $\hat{V}_1$ is chosen in a graph that corresponds to case 2. Vertex number 3 and vertex number 5 still have maximum degree. Thus, $\hat{V}_1$ has the same solution as in figure 2a) for $n_1 = 1$. For $n_1 = 2$, $\hat{V}_1 = \{3, 5\}$ is again one solution but it is not unique. Since vertex number 3 and vertex number 5 are adjacent a second solution is obtained if $\hat{V}_1$ consists of two non-adjacent vertices, one vertex with maximum degree and one vertex with degree one less than maximum degree. $\hat{V}_1 = \{3, 4\}$, $\hat{V}_1 = \{3, 6\}$ and $\hat{V}_1 = \{5, 6\}$ are therefore also solutions.

Figure 2c) illustrates a graph that corresponds to case 2 for $n_1 = 3$. By choosing $\hat{V}_1$ to contain the three vertices with maximum degree that are reciprocally adjacent, $\hat{V}_1 = \{3, 4, 5\}$, we have that $2\hat{r}_{11} + \hat{r}_{12} = 9$ and $\hat{r}_{11} + \hat{r}_{12} = 6$. This is not the solution that maximizes the likelihood function. If we instead choose $\hat{V}_1$ to contain two of the adjacent vertices with maximum degree and one vertex with degree one less than maximum not adjacent to any of these two vertices, for example $\hat{V}_1 = \{2, 4, 5\}$, then we have that $2\hat{r}_{11} + \hat{r}_{12} = 8$ and $\hat{r}_{11} + \hat{r}_{12} = 7$. The sum of the degrees is less for the solution $\hat{V}_1 = \{2, 4, 5\}$, but $L\left(\hat{p}, \hat{\Delta}, V_1\right)$ depends on $\hat{V}_1$ only through the

sufficient statistics $\hat{r}_{11} + \hat{r}_{12}$. Since $\hat{r}_{11} + \hat{r}_{12}$ is larger for $\hat{V}_1 = \{2, 4, 5\}$ than for $\hat{V}_1 = \{3, 4, 5\}$, $L\left(\hat{p}, \hat{\Delta}, V_1\right)$ is maximized by $\hat{V}_1 = \{2, 4, 5\}$ and not by $\hat{V}_1 = \{3, 4, 5\}$.

As n_1 increases the number of ways to choose $\hat{V}_1$ increases leading to more time consuming work. If n_1 is large and n_2 is small it is more convenient trying to find $r_{22}\left(\hat{V}_2\right)$ that maximizes $L\left(\hat{p}, \hat{\Delta}, V_2\right)$.

If at least one of n_1 or n_2 is equal to 1 or 2 the degree of each vertex is enough as a sufficient statistic to estimate V_1 and V_2. If $n_1 = 3$ and $n_2 \geqslant 3$ (or $n_2 = 3$ and $n_1 \geqslant 3$), we see from the procedure described above of finding $\hat{V}_1$ $\left(\text{or } \hat{V}_2\right)$ that a sufficient statistic is every vertex degree and the column elements in the adjacency matrix, A, for the vertices with maximum degree. As n_1 and n_2 increase, we need more information of the structure in A for the statistic to be sufficient. If n_1 and n_2 is large enough the complete matrix A is needed in order to obtain a sufficient statistic.

Suppose we have a realization of a graph generated by a random process and we want to test if we have centrality. A standard test procedure is to perform the likelihood ratio test. Assuming a Bernoulli distribution with a common edge probability under the null hypothesis and a block model under the alternative hypothesis, the likelihood ratio is

$$LR = \frac{L\left(H_0\right)}{L\left(H_1\right)} = \frac{L_0\left(\hat{p}\right)}{L_1\left(\hat{P}, \hat{V}_2\right)} = \frac{L_0\left(\hat{p}\right)}{\max\limits_{\hat{r}_{22}} L_1\left(\hat{P}, V_2\right)},$$

where the numerator is

$$L_0\left(\hat{p}\right) = \left(\frac{r}{\binom{n}{2}}\right)^r \left(1 - \frac{r}{\binom{n}{2}}\right)^{\binom{n}{2}-r}$$

and the denominator is as defined in (6). Now we need to find a critical region

of size α to be able to carry out the test. The null hypothesis is rejected if the likelihood ratio, LR, is less than or equal to some constant c. That is, we need to obtain a critical region, $C = \{LR \leqslant c\}$, of size α for testing H_0 against H_1, where c is selected such that $P\left(LR \leqslant c; H_0\right) = \alpha$.

In order to find C we have to know the distribution of LR. It will be a difficult task to derive the distribution analytically. An alternative is to estimate the distribution by performing computer simulations. Since this

procedure also includes determining $r_{22}(V_2)$, which involves complicated and tedious work as already has been discussed, it is more convenient to focus on computer simulations from the start. That is, we decide on a proper test that will capture centrality and then estimate the distribution of the test by computer simulations. Tables of C are then obtained for the desired levels of size α. A benefit with this approach is that we don't have to limit ourselves to test quantities based on degree. Therefore, the group of tests are extended to include test quantities based on distance. These tests are compared with tests based on degree by investigating the power functions.

4 Tests of centrality

4.1 Introduction

Abandoning the idea of testing centrality by comparing the likelihood functions for different block models we concentrate instead of developing tests where the key concept is computer simulation. The distributions and the critical values of the test statistics are estimated under the null hypothesis. In order to decide which of the test statistics is most suitable for testing centrality, we want to know how likely the test is to reject the null hypothesis of no centrality if the graph has been generated by some model of central actors. Thus, the power functions are estimated and compared for the different test statistics. Performing simulation studies allows us to use tests based on other centrality measures than those based on degree. In this study ten tests are investigated, eight of them are based on distance and two of them are based on degree.

4.2 Tests based on degree.

The two tests in this study that are based on degree both measure the variability of the vertex centrality indices. The first test quantity is the difference of the maximum degree and the mean degree,

$$T_1 = \max_i \ a_i - \bar{a}.$$

The second test, the variance of the degrees

$$T_2 = s_a^2$$

is often recommended as an index of group level centrality; see for instance Snijders (1981a,1981b). There is a wide range of other test quantities based

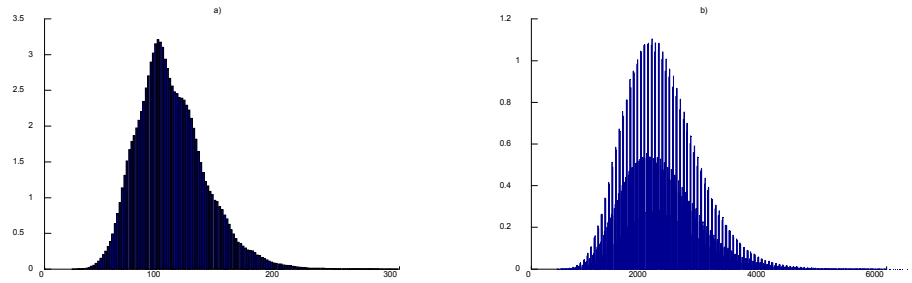

Figure 3: Simulated null distributions of two centrality test statistics where the edges are generated from a *Bernoulli* (p) graph of order $n = 30$ and edge probability $p = 0.1$. The number of simulations is 1000000; a) The simulated probability distribution of T_1; b) The simulated probability distribution of T_2.

on degree proposed in the literature; see, for instance, Kephart (1950), Blau (1977) and Freeman (1979).

In the literature, the distributions of the test statistics have only been studied for one of them, namely T_2. From Figure 3, showing the simulated null distributions of T_1 and T_2, we see that the distribution of T_2 can be approximated with a gamma distribution. For a more detailed reading on the degree variance in Bernoulli graphs, see Hagberg (2000).

4.3 Tests based on distance

Out of the eight tests based on distance, six measure the variability of the actor level indices and two are averages of the actor level indices. The eight tests are separated into two groups, each group including four tests.

In the first group of centrality indices we focus on connected components. Let n_i be the number of vertices in the component of vertex i. The average distance from vertex i to its other connected vertices is denoted by $\bar{d}_i$. Since the minimum value of $\bar{d}_i$ is 1 for $n_i > 1$, we define $\bar{d}_i = 1$ if vertex i is isolated. To obtain a centrality index that increases as the length of the geodesics decrease, we use the reciprocal of $\bar{d}_i$ rather than $\bar{d}_i$ as a centrality index on actor level. The range of $1/\bar{d}_i$ is between 0 and 1, and its maximum is obtained for any vertex i that is adjacent to all the other vertices in the same component. If a Bernoulli model generates graphs with a small common edge probability we will obtain a relatively large proportion of vertices in components of small order. These vertices have high centrality values and will therefore to a large extent determine the group level centrality index.

If the probability is small enough a vast number of vertices will be isolated and the consequence is that we will accept centrality. The interpretation is that we have a realization of a random graph consisting of isolated vertices that are central in their own component. Therefore the actor centrality index is multiplied with a weight of size n_i/n, whose purpose is to give more importance to vertices in larger components. The centrality index for vertex i is then given by $c_i = n_i/n\bar{d}_i$ for $n_i \geqslant 1$. Note that if G is connected then $c_i = 1/\bar{d}_i$.

In the second group of test statistics based on distance we define $\frac{1}{d_{ij}} = 0$ if $d_{ij} = \infty$. By applying this definition we don't have to focus on distance in connected components. A second centrality index for any vertex i is now given by

$$c_i' = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{d_{ij}}.$$

c_i' also possesses the desireable property of a centrality index of having a range between 0 and 1 where a maximum attained index value implies centrality.

The following eight tests of centrality based on c_i and c_i' are presented in this study :

- $T_3 = \max_i c_i$
- $T_4 = \bar{c} = \frac{1}{n} \sum_i^n c_i$
- $T_5 = \max_i c_i - \bar{c}$
- $T_6 = s_c^2 = \frac{1}{n} \sum_i (c_i - \bar{c})^2$
- $T_7 = \max_i c_i'$
- $T_8 = \bar{c}' = \frac{1}{n} \sum_i^n c_i'$
- $T_9 = \max_i c_i' - \bar{c}'$
- $T_{10} = s_{c'}^2 = \frac{1}{n} \sum_i (c_i' - \bar{c}')^2$

5 Power against centrality

To decide which of the tests that are most suitable for testing centrality, we want to know how likely the test is to reject the null hypothesis of no centrality if a realization of G is generated by a block model containing a central block. That is, we want to know the power of the tests. In order to investigate the power of the different test statistics, we have performed a simulation study. The model that generates the graphs is limited to be determined by the two parameters introduced in Section 3.2.2, p and Δ. As an illustration of the behaviour of the power function of the test statistics we assigned p the value 0.1 and Δ the values 0.0 (the null distribution model), 0.2, 0.4, 0.6, and 0.8. Any value of p could have been used but the tests are good for small values of p or equivalently, for large values of Δ.

1000000 random graphs with a fixed order of 20 were generated for each Δ. For $\Delta = 0$, the critical values, $c_{i;\alpha}$, were determined such that we could obtain significance levels of approximately the same size for all the test quantities in order to be able to compare the power functions. Since the test statistics are discrete we may, particularly in graphs of smaller order for which the possible values of the test statistics are small, find it difficult to obtain critical values at the desired significance level α. In this study a less commonly used significance level of approximately size 0.032 (for T_3 $\alpha = 0.035$) was conveniently chosen. In Table 1, showing the results of the estimated power functions for various values of Δ for $n_1 = 1$ and $n_1 = 2$, the estimated significance level of each test statistic is given for $\Delta = 0$.

From Table 1 and Figure 4 it is clear that none of the test quantities is uniformly most powerful. If we distinguish between the two block sizes we see that for $n_1 = 1$, T_1 is a uniformly most powerful test. For $n_1 = 2$, there are three test quantities that yield power functions that are similar, namely T_2, T_3 and T_7.

Of the two test statistics based on degree, T_1 and T_2, we see that for $n_1 = 1$, T_1 yields a slightly better power than T_2 and for $n_1 = 2$, T_2 yields a slightly better power than T_1.

A test quantity based on c_i corresponding to the same test quantity based on c_i', generates approximately the same power function. This is due to the fact that both indices are weighted means; c_i can be considered as the inverse of the arithmetic mean and c_i' is the inverse of the harmonic mean.

In general, the tests based on variability of the actor level centralities show higher power than the two tests where the group level index is an average of the actor level centralities. An exception is that T_4 and T_8 has higher power than T_5 and T_9 for $n_1 = 2$.

T_6 and T_{10}, the variance of c_i and c_i' respectively, besides showing a poor

				n_1	=	1				
Δ	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
0.0	3.2	3.2	3.5	3.2	3.2	3.3	3.2	3.2	3.2	3.2
0.2	32.2	29.1	29.1	18.6	20.8	15.1	29.7	16.1	23.1	11.3
0.4	89.6	84.6	83.3	58.7	73.6	37.8	84.8	48.3	78.2	23.3
0.6	99.9	99.7	99.5	93.6	98.3	49.0	99.6	85.7	99.0	20.6
0.8	100.0	100.0	100.0	100.0	100.0	24.9	100.0	99.7	100.0	3.9
				n_1	=	2				
Δ	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
0.0	3.2	3.2	3.5	3.2	3.2	3.3	3.2	3.2	3.2	3.2
0.2	44.6	53.5	53.5	44.4	22.1	21.0	54.3	39.0	22.0	14.6
0.4	97.1	98.7	98.0	94.2	68.8	33.8	98.3	89.1	69.4	16.0
0.6	100.0	100.0	100.0	100.0	96.9	23.4	100.0	99.8	97.5	6.8
0.8	100.0	100.0	100.0	100.0	100.0	37.6	100.0	100.0	100.0	2.8

Table 1: For $n_1 = 1$ and $n_1 = 2$ the simulated power functions of the test statistics are given for $n = 20$, $p = 0.1$ and various Δ based on 1000000 replications

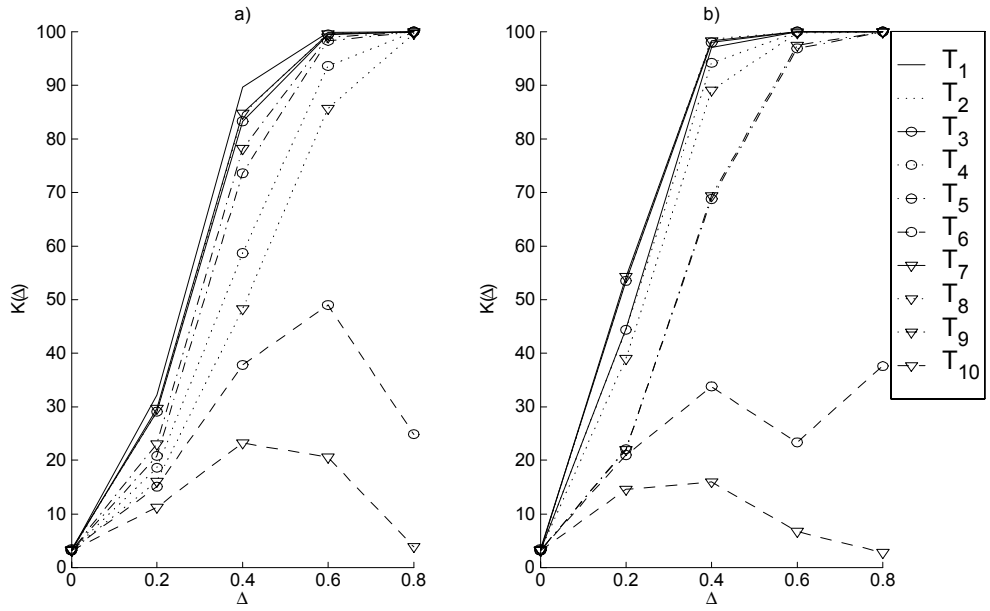

Figure 4: The simulated power functions of the tests in random graphs of order $n = 20$, $p = 0.1$ and various Δ based on 1000000 replications; a) $n_1 = 1$; b) $n_2 = 2$.

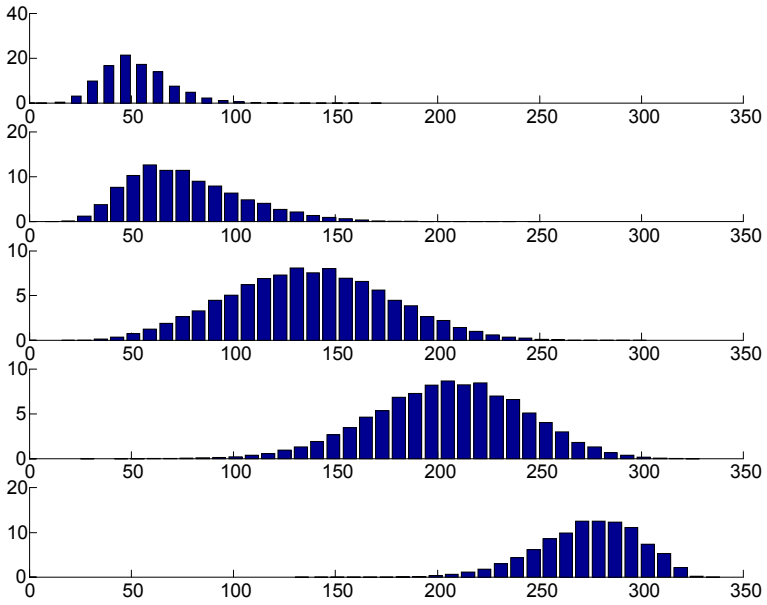

Figure 5: The simulated distributions of T_1 in random graphs of order $n = 20$, $p = 0.1$ and various Δ based on 1000000 replications.

power also has a power function that is approximately decreasing for $\Delta > 0.4$. The explanation to this behaviour of the power function is that if centrality exists, then the expected degree is higher for the central actors. As Δ increases the expected geodesic distance between this actor and any other actor decrease, resulting in that also the expected geodesic distance between any pairs of actors in G will decrease. Consequently we will obtain a larger homogeneity in the actor centrality indices which implies that also the variability will decrease. Therefore the variance of the actor centrality indices is not a consistent measure when centrality is based on distance. That is, if centrality means that some actors in the social network generate edges with a larger probability. In Figures 5 - 9 the distributions of T_1, T_2, T_3, T_5 and T_6 are illustrated for various Δ.

As Δ increases the simulated distributions of the tests T_1, T_2, T_3 and T_5 move to the right which implies that the power functions also increase. Figure 9 shows that the simulated distribution of T_6 doesn't move to the right which implies that T_6 has a non-increasing power function. Furthermore, the distribution appears to be uni-modal under the null hypothesis, but when

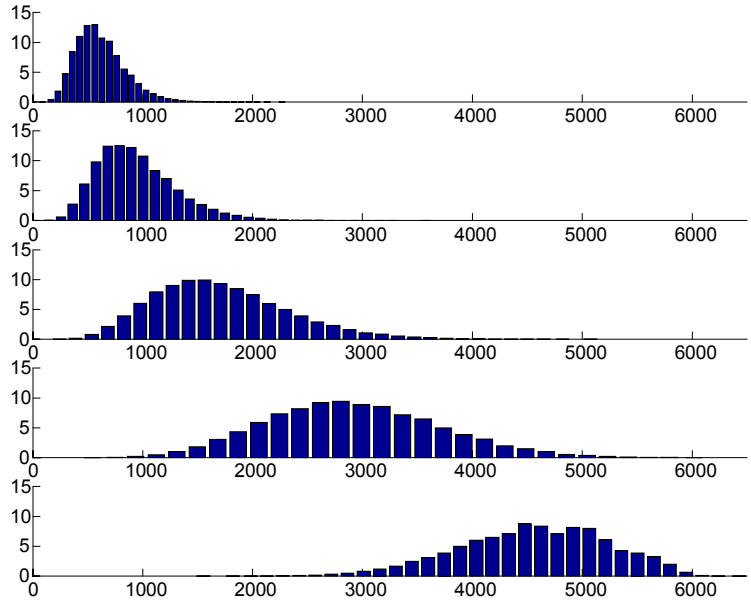

Figure 6: The simulated distributions of T_2 in random graphs of order $n = 20$, $p = 0.1$ and various Δ based on 1000000 replications.

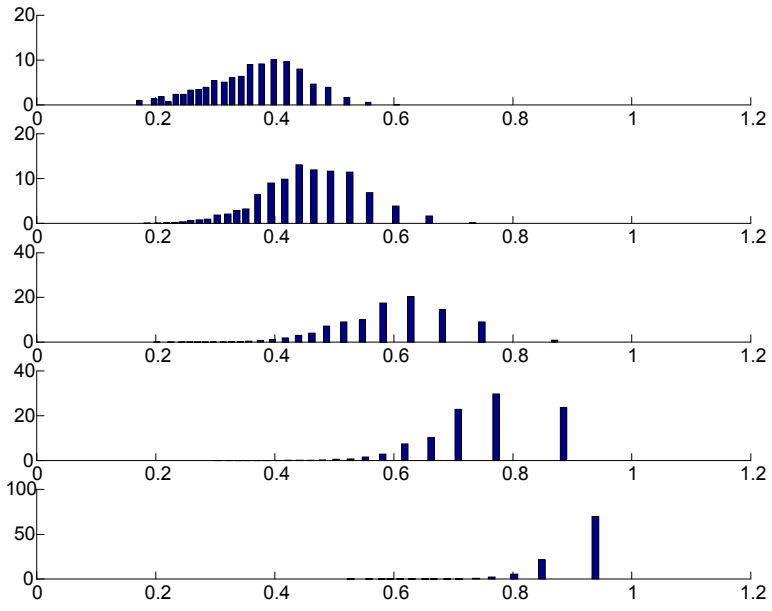

Figure 7: The simulated distributions of T_3 in random graphs of order $n = 20$, $p = 0.1$ and various Δ based on 1000000 replications.

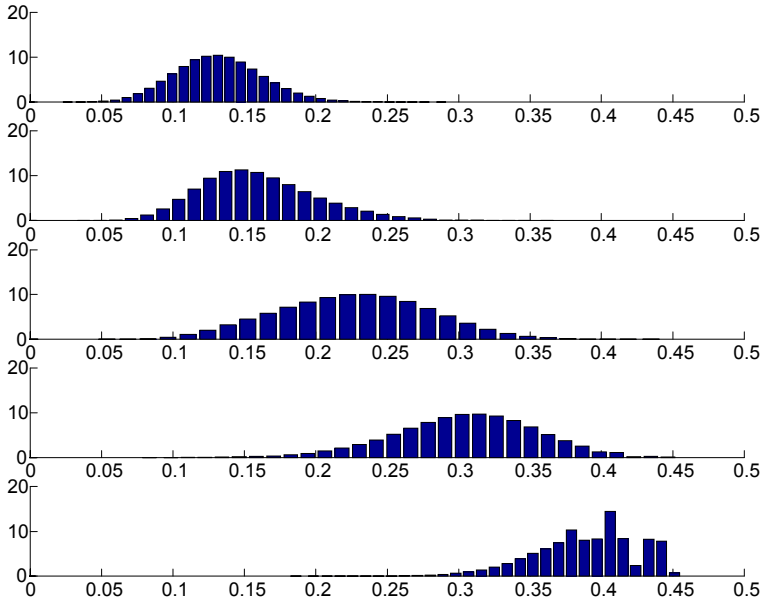

Figure 8: The simulated distributions of T_5 in random graphs of order $n = 20$, $p = 0.1$ and various Δ based on 1000000 replications.

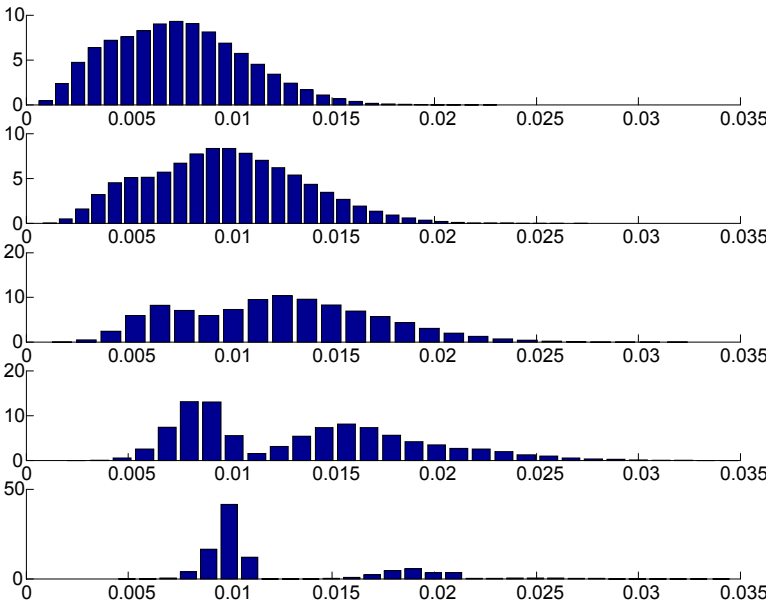

Figure 9: The simulated distributions of T_6 in random graphs of order $n = 20$, $p = 0.1$ and various Δ based on 1000000 replications.

Δ increases the distribution more and more attains a bi-modal shape. As is explained above, the reason for this behaviour could be that the geodesic between a pair of vertices mainly takes two values. As Δ increases the main body of the distribution is centered around the lower value since most of the vertices will be adjacent.

6 Concluding remarks

The $Bernoulli(p)$ model is considered as the stochastic model when no centrality exists. Under centrality, block models of unknown blocks of known orders are considered. Likelihood ratio testing is attempted but the procedure is rejected partly because of the complexity of obtaining the maximum likelihood estimators. The procedure of obtaining the maximum likelihood estimators includes identifying the blocks, which turns out to be a rather tedious and time consuming work when the blocks are of order larger than two. Since another difficulty is to derive the distribution of the likelihood ratios analytically, the idea of using the likelihood ratio test procedure is abandoned.

Instead the focus is on estimating the distributions of some test statistics that capture centrality by performing computer simulations. A large number of graphs of order 20 is generated and the critical values and power functions are estimated. There exists no test that is uniformly most powerful, but the test quantities that are based on the maximum of the centrality indices generate large power functions. Test T_2, the variance of the vertex degrees, performs well whereas the two tests T_6 and T_{10}, the variance of the actor level indices when centrality is based on distance, constantly exhibit a poor power. This is due to the paradox that increasing actor centrality induces less variability when centrality is based on distance.

The rather simple null hypothesis model, $Bernoulli(p)$, may not cover all the non-centrality irregularities and therefore it is always a risk that we are testing other deviations than the ones based on centrality. One may also be sceptical to what actually is rejected if the null hypothesis states that the edges are generated accordingly to a $Bernoulli$ model for which no vertices are central. Now, rejection of the null hypothesis might be due to inaccuracy in the modeling of the non-centrality structure rather than evidence of centrality. Another critical assumption is the validity of independent dyads. Intuitively one should expect a dyad dependency that is inversely proportional to the order of G. Modeling with dependent dyads should open possibilities to obtain more elaborate models; see Frank and Strauss (1986).

In this study the order of G has been fixed and the order of the central

blocks under the alternative hypothesis has been restricted to $n_1 = 1$ and $n_1 = 2$. Furthermore, the number of blocks under the alternative hypothesis are only considered for $K = 2$. An extension in future work would be to allow a larger variation of the edge probabilities and the number of blocks. Further we should consider not only to have randomized size, but also assume a model with a randomized order of G and a randomized order of the blocks. These generalizations induce complexity that might be required of appropriate models.

References

[1] Bock, H. (1996). Probability models and hypotheses testing in partitioning cluster analysis. In *Clustering and Classification*, editors P. Arabie, L.J. Hubert, G. De Soete. Singapore: World Scientific.

[2] Blau, P.M. (1977). *Inequality and Heterogeneity.* New York: Free Press.

[3] Frank, O., Hallinan, M. and Nowicki, K. (1985). Clustering of dyad distributions as a tool in network modeling. *Journal of Mathematical Sociology,* ***11,*** 47-64.

[4] Frank, O., Komanska, H. and Widaman, K.F. (1985). Cluster analysis of dyad distributions in networks. *Journal of Classification,* **2**, 219-238.

[5] Frank, O. and Strauss, D. (1986). Markov Graphs. *Journal of the American Statistical Association*, **81**, 832-842.

[6] Freeman, L.C. (1977). A set of measures of centrality based on betweeness. *Sociometry,* **40**, 35-41.

[7] Freeman, L.C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, **1**, 215-239.

[8] Hagberg, J. (2000). Centrality testing and the distribution of the degree variance in Bernoulli graphs. Unpublished manuscript. Department of Statistics, Stockholm University.

[9] Holland, P.W., Laskey, K.B. and Leinhardt, S. (1983). Stochastic blockmodels; Some first steps. *Social Networks,* ***5,*** 109-137.

[10] Jansson, I. (1997c). Popularity structure in friendship networks. Research Report 1997:8. Department of Statistics, Stockholm University.

[11] Karlberg, M. (1997). Testing transitivity in graphs. *Social Networks*, **19**, 325-343.

[12] Karlberg, M. (1999). Testing transitivity in digraphs. *Sociological Methodology*, **29**, 225-251.

[13] Kephart, W.M. (1950). A quantitative analysis of intragroup relationships. *American Journal of Sociology*, **55**, 544-549.

[14] Palmer, E. (1985). *Graphical Evolution*. New York: Wiley.

[15] Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, **31**, 581-603.

[16] Snijders, T.A.B. (1981a). The degree variance: An index of graph heterogeneity. *Social Networks*, **3**, 163-174.

[17] Snijders, T.A.B. (1981b). Maximum value and null moments of the degree variance. TW-report 229. Department of Mathematics, University of Groningen.

[18] Wang, Y.J. and Wong, G.Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, **82**, 8-19.

[19] Wasserman, S. and Anderson, C. (1987). Stochastic *a posteriori* blockmodels: Construction and assessment. *Social Networks*, **9**, 1-36.

[20] Wasserman, S. and Faust, K. (1994). *Social Network Analysis: methods and applications*. New York: Cambridge University Press.