Social Networks 59 (2019) 171–183

Contents lists available at ScienceDirect

Social Networks

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

Permutation and randomization tests for network analysis

Mark M. Fredrickson[a,*], Yuguo Chen[b]

[a] Department of Statistics, University of Michigan, 1085 South University Ave, Ann Arbor, MI 48109, United States
[b] Department of Statistics, University of Illinois at Urbana-Champaign, 725 South Wright Street, Champaign, IL 61820, United States

ARTICLE INFO

Keywords:
Randomization
Permutation
Hypothesis testing
Causal inference
Quadratic assignment procedure (QAP)
Edge counts
Mahalanobis distance
Coefficient of determination
Clustering
Centrality

ABSTRACT

Permutation tests have a long history in testing hypotheses of independence between nodal attributes and network structure, though they are often thought less informative than parametric modeling techniques. In this paper, we show that when the nodal attribute is random assignment to a treatment condition, permutation tests provide a valid test of the causal effect of treatment. We discuss existing test statistics used in network permutation tests and propose several new statistics. In simulations we find that these statistics perform well compared to parametric tests and that specific statistics can be selected to provide power against common network models. We illustrate the methods with gene-wide association study performed on randomized study participants and an observational study of gender membership on Scandinavian corporate boards.

1. Introduction

As a tool to relate nodal covariates to network features, permutation tests have a long history in the network analysis literature. Dating back to at least Mantel (1967), permutation approaches work by relabeling nodes uniformly at random, leading to a constrained permutation on the observed network's adjacency matrix. The permuted adjacency matrix is then compared to a matrix expressing a variable measured for each dyad through some type of correlation measure. If the observed relationship between the network and the variable is extreme compared to values observed under the permutations, the researcher can reject the hypothesis that the network and the variable are statistically independent.

Such permutation tests are attractive as they require very few assumptions, compared to parametric graph models.[1] This simplicity, however, is sometimes thought to be a limiting factor as well. As part of a broader survey of network analysis techniques, Snijders (2011) largely dismisses permutation approaches, saying "[permutation] approaches are useful, but they are not discussed further here because they regard network structure as nuisance rather than substance and do not attempt to model network dependencies" (p. 134). The goal of this paper is to show that permutation approaches have significant merit as a tool in network analysis. First, for causal questions when the nodal covariate of interest is the treatment assignment of a randomized controlled trial, we show that randomization tests of no treatment effect are precisely permutation tests of independence between the treatment and the network. This permits testing causal hypotheses without requiring strong assumptions on the network's data generating process. Second, through careful selection of a test statistic researchers can make tests sensitive to specific network structure, such as increased clustering or centrality associated with the nodal covariate or treatment assignment. Even if the larger goal is modeling, permutation and randomization tests can be used to refine the early model building process before committing to a particular set of assumptions for a network's data generating process.

Causal attributions are difficult in any context and particularly vexing in network analysis. In recent years, the Neyman–Rubin causal model, also known as the "potential outcomes framework", has become a valuable tool for framing causal inference in statistical terms (for discussions from a non-network perspective, see Imbens and Rubin (2015) and Hernán and Robins (2019)). In this model, observations are thought to have a set of fixed potential outcomes, each revealed by a particular treatment assignment. When treatment is assigned independently of the potential outcomes, statistical analysis reveals the causal effect of treatment. As randomized controlled trials exhibit precisely this form of treatment assignment, the potential outcomes

* Corresponding author.
E-mail addresses: mfredric@umich.edu (M.M. Fredrickson), yuguo@illinois.edu (Y. Chen).

[1] Extensive coverage of parametric approaches can be found in Kolaczyk (2009), Goldenberg et al. (2010), Snijders (2011), Fienberg (2012), Hunter et al. (2012), O'Malley (2013), and Amati et al. (2018).

https://doi.org/10.1016/j.socnet.2019.08.001

Available online 22 August 2019

framework justifies their claim to the "gold standard" of causal inference, and we place special emphasis on networks measured after nodes are randomly assigned to treatment conditions. See Matous and Wang (2019) for a recent example of this type of study design.

In most statistical applications of the Neyman–Rubin model, it is assumed that the assignment of one unit does not influence the potential outcomes of other units (Rubin, 1980). These strong independence assumptions seem out of place in a network analysis context (Fienberg, 2012).[2] To avoid this difficulty, we highlight the role of a specific null hypothesis, the sharp null hypothesis of no effect, which states that the observed network would have been identical under any possible treatment assignment. This null hypothesis naturally leads to a testing strategy in which treatment assignment is repeatedly shuffled in order to create a null distribution (Fisher, 1935; Rosenbaum, 2002, 2010), a procedure which we describe as a "randomization test." Randomization tests justify inference on the known randomization mechanism, rather than appealing to stronger parametric assumptions. Outside of the network context, an extensive literature has shown how regression modeling fails to take into account the true stochastic nature of randomized treatment and can lead to biased effect estimates (Berk, 2004; Freedman, 2008a,c,b). In our simulation results, we see that parametric network analysis methods can achieve excellent power to reject false null hypotheses when parametric assumptions are true. When assumptions fail to hold, however, Type I error can be remarkably poor.

When treatment assignment is performed using complete random assignment — the number of treated and control nodes is fixed, with all assignments equally probable — randomization tests are equivalent to permutation tests using a categorical nodal variable (Maritz, 1981). At the heart of randomization and permutation tests is a test statistic that maps the permuted node labels and observed network to a scalar value. Test statistics vary in their functional form and, consequently, are sensitive to different alternative hypotheses, deviations from the null hypothesis of independence between the network and the nodal attribute. Mantel (1967) proposed a linear statistic, equivalent to the correlation between the network and treatment indicators, and similar approaches were introduced in several other disciplines (Whaley, 1983; Good, 2005, chapter 10). Linear statistics also form the basis of "quadratic assignment procedure" (QAP) methods that use linear regression techniques to define test statistics (Baker and Hubert, 1981; Krackhardt, 1987, 1988; Dow and de Waal, 1989; Nyblom et al., 2003). Edge count statistics can also be non-linear. We propose using the probability mass function of the edge counts as a test statistic. A similar approach can also be found in Chen and Friedman (2017), who proposed a Mahalanobis distance based on edge counts in the context of high dimensional inference. Interestingly, this statistic is quadratic in the within treatment and control group edge counts and can be shown to be similar to the coefficient of determination (R^2) from a QAP regression. As these methods are functions of edge counts within subgraphs defined by the nodal attribute, they tend to be sensitive against alternatives that change the number of edges within groups, but may be less sensitive to alternatives that change higher order properties of the graph. We also introduce statistics that are sensitive to treatment effects that increase clustering or centrality for nodes with different levels of the treatment assignment. We show that these statistics can be selected to provide power to detect the presence of treatment effects in the network, including those corresponding to several well known network formation models.

The rest of the paper is arranged as follows. In Section 2.1 we review causal inference approaches to randomized trials that use the randomization procedure as the "reasoned basis" for inference (Fisher, 1935), in particular how this approach can be extended to networks. In Sections 2.2 and 2.3 we develop local and global tests that are sensitive to the various ways in which treatments can influence network features. In Section 3 we evaluate the statistical properties of the proposed methods in a variety of simulated networks. In Section 4 we apply the methods to a gene-wide association study after a randomized controlled trial and to a network of board members for a set of Norwegian companies. Section 5 concludes with a brief discussion.

2. Method

2.1. Causal inference for networks

Consider n units in a study with a random treatment $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)'$, where $Z_i \in \{0, 1\}$. As $\mathbf{Z}$ is controlled by the researcher, the distribution of $\mathbf{Z}$ is known. Typically, and throughout this document, $\mathbf{Z}$ is generated by selecting n_1 units for treatment, setting $Z_i = 1$ and selecting the remaining n_0 units to receive control, setting $Z_i = 0$. For simplicity, we take all $\mathbf{Z}$ to be equally probable. Extensions with constrained randomization, when units are first blocked by similar characteristics for example, are immediate. To simplify later notation, for any variable write $\mathbf{Z}^{(k)} = (I(Z_1 = k), I(Z_2 = k), \ldots, I(Z_n = k))'$, where I is the indicator function.

In the potential outcomes framework (Neyman, 1923), each unit's response is a fixed value indexed by the treatment assignment: $Y_i = y_i(\mathbf{Z})$, with $\mathbf{Y} = \mathbf{y}(\mathbf{Z}) = (y_1(\mathbf{Z}), y_2(\mathbf{Z}), \ldots, y_n(\mathbf{Z}))'$. Unit i would have response $y_i(\mathbf{z})$ if $\mathbf{Z} = \mathbf{z}$, but for some other treatment assignment $\mathbf{u}$, the response would be $y_i(\mathbf{u})$. We say that a treatment has an effect if $y_i(\mathbf{z}) \neq y_i(\mathbf{u})$ for at least one unit i. By the fundamental problem of causal inference (Holland, 1986), we cannot observe both $y(\mathbf{z})$ and $y(\mathbf{u})$, so we must perform inference to determine if treatment has an effect. In this paper we use a sharp null hypothesis of no effect that states $H_0 : \mathbf{y}(\mathbf{z}) = \mathbf{y}(\mathbf{u})$ for all $\mathbf{z}$ and $\mathbf{u}$. Under this hypothesis, for any treatment assignment $\mathbf{Z}$, we would have observed precisely the same outcome $\mathbf{y}$, and a null distribution for the hypothesis results from evaluating a test statistic $T(\mathbf{Z}, \mathbf{y})$ over the distribution of $\mathbf{Z}$ keeping $\mathbf{y}$ fixed.

While the null hypothesis is quite specific, there are many possible alternative hypotheses that would have $\mathbf{y}(\mathbf{z}) \neq \mathbf{y}(\mathbf{u})$ for at least some $\mathbf{z}$ and $\mathbf{u}$. The researcher can focus attention on one particular alternative hypothesis through the choice of the test statistic $T(\mathbf{Z}, \mathbf{y})$. For example, a difference of means statistic would be sensitive to treatments that made the two groups different on average, but insensitive to treatments that only operated on the variance of the outcome. Under the null hypothesis, the distribution of $T(\mathbf{Z}, \mathbf{y})$ is entirely determined by the distribution of $\mathbf{Z}$, and inference proceeds by enumerating, or sampling from, $\mathbf{Z}$ to compute the distribution of T (Fisher, 1935; Maritz, 1981; Rosenbaum, 2002). When the statistic T is selected such that large values of T indicate evidence against the null hypothesis in favor of the alternative, the p-value of the test is given by

$$p^+(\mathbf{z}, \mathbf{y}) = \Pr(T(\mathbf{Z}, \mathbf{y}) \geq T(\mathbf{z}, \mathbf{y})) = \sum_{\mathbf{Z} \in \Omega} \Pr(\mathbf{Z}) I(T(\mathbf{Z}, \mathbf{y}) \geq T(\mathbf{z}, \mathbf{y})),$$

where $\mathbf{z}$ and $\mathbf{y}$ are the observed treatment and outcome and Ω is the sample space of possible assignments. An analogous statistic, $p^-(\mathbf{z}, \mathbf{y})$, computes an appropriate tail probability when small values of T are evidence against the null. In general, two-sided tests can be constructed using $p(\mathbf{z}, \mathbf{y}) = 2\min(p^+(\mathbf{z}, \mathbf{y}), p^-(\mathbf{z}, \mathbf{y}))$ (Cox, 2006, chapter 3).

While the definition of $p(\mathbf{z}, \mathbf{y})$ encompasses any treatment assignment mechanism, in this paper we focus on complete random assignment for which $\Pr(\mathbf{Z} = \mathbf{z}) = (n_1!(n - n_1)!)/n!$. In this situation, randomization tests are mathematically equivalent to two-sample permutation tests (Maritz, 1981). Nevertheless, we emphasize that permutation tests do not necessarily imply a causal interpretation outside of random assignment. With this caveat in mind, throughout the

[2] Our interest in this paper is networks measured after treatment. There is also a developing literature that extends the Neyman–Rubin model inference for numeric outcomes in the presence of pre-treatment networks that may allow treatment to *spillover* from treated to control nodes (Rosenbaum, 2007; Bowers et al., 2013; Choi, 2017; Aronow and Samii, 2017; Athey et al., 2018).

rest of the paper, for simplicity we refer to "treatment" and "control" groups, even in situations in which the group labels were not randomly assigned.The randomization inference approach generalizes to networks in a natural way. Consider applying treatment to n_1 of the n units and then measuring a simple network for those for all n units (i.e., an undirected network with no self-loops). Rather than focus on the n subjects in the experiment, we shift focus to the $m = n(n-1)/2$ possible connections between them. As with other types of outcomes, we can posit the existence of *potential networks* composed of *potential edges*. Each dyad (i, j), $i < j$, may have one of four possible treatment assignments, depending on whether i, j, both, or neither is treated. If we assume that edge (i, j) would behave in the same fashion if either one of its endpoints were treated, we can state the treatment levels as $W_{ij} = Z_i + Z_j \in \{0, 1, 2\}$,writing $\mathbf{W} = (W_{12}, W_{13}, \ldots, W_{(n-1)n})'$. For each dyad (i, j), let $y_{ij}(\mathbf{W}) = 1$ if units i and j have a link following treatment $\mathbf{W}$ and $y_{ij}(\mathbf{W}) = 0$ otherwise.

If treatment had no effect, then we would observe the same network $\mathbf{y}$ under all treatment assignments. We can apply randomization inference to the network by selecting a test statistic $T(\mathbf{W}, \mathbf{y})$. In the next section, we consider statistics that operate on $\mathbf{y}$ directly. In Section 2.3 we consider test statistics $T(\mathbf{Z}, g(\mathbf{y}))$ that operate on $\mathbf{y}$ through a function g that summarizes each node's position within the network. Throughout both sections we emphasize alternatives for which the test statistics are expected to exhibit high power, the probability of rejecting a false null hypothesis. By selecting statistics that have high power against interesting alternatives, researchers can detect interesting treatment effects in the network.

2.2. Local approaches

In this section, we discuss test statistics that operate on local features of the graph, with locality defined by the treatment and control subgraphs. In general, these statistics tabulate some feature within each of the treatment and control subgraphs and compare the two groups on this feature. These statistics can be expressed through cross classifying edges by treatment assignment, and they bear a strong resemblance to graph models descended from log-linear models (Holland and Leinhardt, 1981; Fienberg and Wasserman, 1981b,a). Unlike approaches based on logistic regression or other parametric models, however, inference is non-parametric, using randomization or permutation to provide valid tests.

As in the previous section, let $Y_{ij}(\mathbf{W}) = 1$ when there is a link in the network between i and j. Define $\mathbf{W}^{(k)} = (I(W_{12} = k), I(W_{13} = k), \ldots, I(W_{(n-1)n} = k))'$. We notate the number of dyads with the treated group as $m_2 = n_1(n_1 - 1)/2$, the number of dyads in the control group as $m_0 = n_0(n_0 - 1)/2$, and the number of treated-control dyads as $m_1 = n_1 n_0$. We label the total number of edges in the network as $R = \mathbf{Y}'\mathbf{Y}$. Cross classifying the edges by treatment assignment leads to the edge count statistics $R_2 = \mathbf{W}^{(2)\prime}\mathbf{Y}$, $R_0 = \mathbf{W}^{(0)\prime}\mathbf{Y}$, and $R_1 = \mathbf{W}^{(1)\prime}\mathbf{Y}$, the edges within and across the treated and control groups, respectively. Under the sharp null hypothesis of no effect, or the equivalent permutation hypothesis, the total number of edges R is a fixed value, though R_2, R_1, and R_0 will vary under different treatment assignments. We discuss several test statistics that use these edge counts to capture the relationship of treatment with the network.

Mantel (1967) introduced a family of permutation tests based on linear test statistics of the form $T_{\text{Mantel}} = \mathbf{Y}'\mathbf{U}$. Statistics of this form have been popularized under the title of quadratic assignment procedure (QAP) methods in the behavioral sciences (Baker and Hubert, 1981), where "quadratic" refers to the number of nodes, rather than the number of edges. As a linear sta/tistic, Mantel's statistic can also be derived from a bivariate linear regression of $\mathbf{Y}$ on $\mathbf{U}$, leading to extensions that can include additional covariates (Krackhardt, 1987, 1988; Dekker et al., 2007). As a special case, when $\mathbf{U} = a\mathbf{W}^{(2)} + b\mathbf{W}^{(1)} + c\mathbf{W}^{(0)}$, Mantel's statistic is a linear combination of the edge counts: $T_{\text{Mantel}}(R_2, R_1, R_0) = aR_2 + bR_1 + cR_0$. Selecting particular values for a, b, and c makes the test sensitive to particular alternative hypotheses. Dow and de Waal (1989) used $(a = 1/m_2, b = 0, c = 0)$ to test how much more compact the treatment group was than the rest of the graph, $(a = 0, b = 1/m_1, c = 0)$ to test how close the treatment group would be to the rest of the network, and $(a = 1/m_2, b = -1/m_1, 0)$ to combine these tests. Nyblom et al. (2003) also considered the case when only a is non-zero and extended the method to consider other nodal covariates.

Methods based on edge counts should be expected to have good power against alternative hypotheses that have consistent effects within or across the treatment and control groups. If treatment is thought to induce effects only between treated units, but not between treated and control units or within the control group, a useful statistic can be constructed by setting $(a = 1/m_2, b = 0, c = 0)$, or equivalently $(a = 1/m_2, b = -1/(m_1 + m_0), c = -1/(m_1 + m_0))$. Similarly, if the treatment regime makes both treated and control units similarly insular, the statistic that sets $(a = 0, b = 1/m_1, c = 0)$ should be sensitive to deviations from the null hypothesis. For less well specified alternatives, using $(a = 1/m_1, b = 0, c = -1/m_0)$ may be a good general choice. We explore the power of the last statistic against several network formation models in Section 3.

Linear combinations of R_2, R_1, and R_0 are not the only method by which these edge counts can be applied. Motivated by using networks to reduce the size of high-dimensional two-sample testing, Chen and Friedman (2017) proposed a Mahalanobis distance computed from R_2 and R_0:

$$T_{\text{CF}}(R_2, R_0) = \begin{pmatrix} R_2 - \mu_2 \\ R_0 - \mu_0 \end{pmatrix}' \Sigma_{20}^{-1} \begin{pmatrix} R_2 - \mu_2 \\ R_0 - \mu_0 \end{pmatrix}. \tag{1}$$

The moments $\mu_2 = Rm_2/m$, $\mu_0 = Rm_0/m$, and Σ_{20} (given in Appendix A) can be computed from combinatorial analysis under the sharp null hypothesis (Frank, 1977, 1978; Chen and Friedman, 2017). As R_2 or R_0 vary from their respective expected values, T_{CF} will increase. In large samples, R_2 and R_0 are approximately normal (Chen and Friedman, 2017), suggesting T_{CF} will have good power against alternative hypotheses for which μ_2 and μ_0 are different and Σ_{20} remains largely unchanged.

In small or moderate networks, however, the distribution of (R_2, R_0) may be decidedly non-normal, so these power properties may not hold. To motivate the next statistic, observe that

$$\begin{aligned} \mathrm{p}^+ &= P(T_{CF}(R_2, R_0) \geq T_{CF}(r_2, r_0)) \\ &= P(\exp\{T_{\text{CF}}(R_2, R_0)\} \geq \exp\{T_{\text{CF}}(r_2, r_0)\}) \\ &= P(\mathrm{f}(R_2, R_0) \leq \mathrm{f}(r_2, r_0)) = \mathrm{P}(-\mathrm{f}(R_2, R_0) \geq -\mathrm{f}(r_2, r_0)), \end{aligned}$$

where f is the distribution function for a normal distribution with mean (μ_2, μ_0) and variance Σ_{20}. In small samples, where the distribution is not necessarily normal, the true probability mass function (PMF) of (R_2, R_0) can still be used as test statistic:

$$T_{\text{PMF}}(R_2, R_0) = -f(R_2, R_0), \tag{2}$$

where f is the probability mass function of (R_2, R_0). Under the null hypothesis of no effect, f depends on the precise structure of the network and must be found by either complete enumeration of all possible $\mathbf{Z}$ or by sampling from $\mathbf{Z}$ using a Monte Carlo approach.

The relationship between T_{CF} and T_{PMF} is similar to different test statistics used in applying Fisher's exact test for a binary outcome in the non-network setting. In this setting the test statistic has the well-known hypergeometric distribution, and Freeman and Halton (1951) suggested using the hypergeometric PMF as a test statistic in two-tailed tests. Similar to the T_{CF} statistic, Radlow and Alf (1975) suggested a χ^2 statistic instead. Gibbons and Pratt (1975) and Agresti (2013, Section 3.5.3) discussed the relative merits of these approaches, which may be informative in selecting between T_{CF} and T_{PMF}. In our experience, both methods perform similarly, particularly when the permutation distribution of T_{CF} is used rather than the normal approximation. Since the

true PMF of (R_2, R_0) is not known in advance, for even moderate networks it must be estimated, which does induce some additional error that tends to make the test conservative.

Another method of using edge counts in a non-linear way comes from the connection between T_{Mantel} and linear regression. In a bivariate ordinary least squares (OLS) regression of $\mathbf{Y}$ on $\mathbf{W}^{(k)}$, it is straightforward to show that the estimated coefficient for $\mathbf{W}^{(k)}$ is proportional to $R_k - \mu_k$, where μ_k is the expected value for R_k under the null hypothesis of no effect. Extending the regression to include all three indicators $\mathbf{W}^{(0)}$, $\mathbf{W}^{(1)}$, and $\mathbf{W}^{(2)}$, but necessarily no intercept, a statistic based on the coefficient of determination[3] is equivalent to

$$T_{\text{CoD}}(R_2, R_1, R_0) = \frac{R_0^2}{m_0} + \frac{R_1^2}{m_1} + \frac{R_2^2}{m_2}.$$

For details, see Appendix B. As this statistic is quadratic in the three R_k, it can be expected to perform similarly to T_{CF}, though it does not account for the covariance between the edge counts. As we will see in Section 3, T_{CF} and T_{CoD} exhibit nearly equivalent power in a wide variety of contexts, with one occasionally exhibiting slightly more power than the other. T_{CoD} therefore has the advantage of being slightly simpler to calculate than T_{CF}, while providing nearly identical performance in many settings.

Edge count statistics are not the only feature of a graph that can be used in a randomization or permutation test. Statistics that count triangles, paths of a certain length, or other local features could be used in place of the edge counts statistics. Alternatively, researchers might wish to use higher order features of the graph not easily described by any local feature. In the next section, we propose methods that incorporate global aspects of the graphs, such as clusters and graph topology in order to test the sharp null hypothesis of no effects.

2.3. Global approaches

In the previous section, we considered statistics that operated by splitting the network into treatment and control subgraphs and comparing features of the two subgraphs, such as edge counts. In this section, we introduce statistics that analyze the entire graph and reduce the graph to a vector of node level summaries. Critically, the first step in this process, analyzing the entire graph, is done without respect to the observed treatment assignment. Only after performing this analysis is the treatment assignment information used to construct a randomization test.

Our first set of test statistics are constructed by applying community detection algorithms to the graph.[4] In this paper, we focus on graph partitioning algorithms that generate a single label for each node: $\mathbf{C} \in \{0, \ldots, k-1\}^n$. We use these labels to construct a $2 \times k$ table counting treatment and control nodes in each cluster. Under the sharp null hypothesis that treatment had no effect, we would have seen exactly the same network, and therefore the clustering algorithm would have provided exactly the same labels for any treatment assignment. In the simplest case when $k = 2$, we have reduced the network to a binary variable measured for each node. With the total number assigned to treatment and control conditions fixed, and the number of units in each cluster fixed under the null hypothesis of no effect, then $\mathbf{Z}'\mathbf{C}$ follows a hypergeometric distribution (Fisher, 1935). The hypothesis can then be tested using Fisher's exact test. For $k > 2$ clusters, several extensions exist that generalize the 2×2 methods to $2 \times k$ tables (Agresti, 1992; Hirji and Johnson, 1996).

Fig. 2 provides a graphical representation of using clustering to create a hypothesis test. For a small simulated network and treatment assignment, the figure shows the initial network, clustering without treatment assignment labels, adding the labels back, and cross classifying the node-cluster counts. In this example, spectral clustering was performed to partition the graph into two blocks, though any other clustering procedure may be used.

Researchers have identified many other global properties of graphs that can be used to describe their topology. Those methods that assign numerical or ordinal scores to nodes can be used to construct tests as well. One key area of inquiry in social network analysis is ranking nodes on their "centrality" to the network. There are several different measures of centrality (Freeman, 1978), typically based on either graph theoretic quantities such as the number of paths in which a node is present (Borgatti, 2005; Borgatti and Everett, 2006) or spectral decomposition of the graph (Bonacich, 1972, 2007).

We use the spectral definition of centrality that defines centrality as the eigenvector of the largest eigenvalue λ of the adjacency matrix A: $A\mathbf{x} = \lambda\mathbf{x}$. For each node, x_i can be thought of as proportional to the sum of the centrality scores of i's neighbors, where λ is the constant of proportionality. While it may sound circular, this definition captures the fact that central nodes are those that are connected to other central nodes. When there are multiple disconnected components to the graph, there will be multiple eigenvectors for λ, with the ith entry being non-zero for only one vector for each node i. In that case, we take x_i to be the non-zero entry for any of the matching eigenvectors. To perform inference, the x_i values can be ranked to perform a Wilcoxon–Mann–Whitney (WMW) test of the hypothesis that treatment had no effect on the network (Lehmann, 1975; Maritz, 1981). Fig. 1 plots the network used in the previous example with node sizes proportional to the rank of the centrality of the node, as measured by eigenvector centrality.

3. Simulations

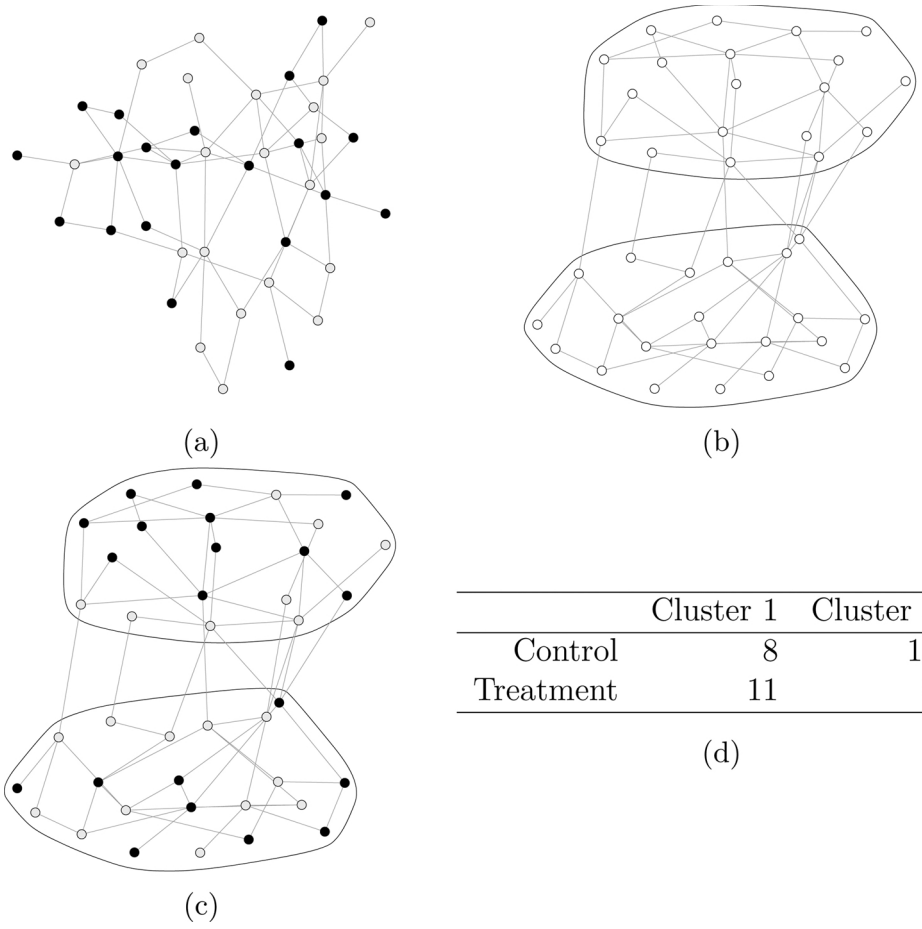
In the following simulations, we investigate statistical properties of the randomization and permutation tests discussed in this paper. In particular, we investigate the Type I and Type II error, the probability of rejecting true and false null hypotheses, respectively. To investigate Type I error, we generate networks and then assign treatment and control labels uniformly at random, independent of the network. To investigate the power (i.e., not making a Type II error), we first assign treatment to the nodes and then allow the treatment assignment of nodes to influence the formation of edges. Using several common network generation methods, we parameterize the model to include the treatment assignment in increasingly influential ways. We fix $\alpha = 0.05$ and find the simulations in which the null hypotheses are rejected.

All simulations contain 100 units with 50 of those assigned to the treatment condition. We allocate the first 50 nodes to the treatment condition and allocate the remaining nodes to the control condition. For each of $k = 500$ replications, a network is generated and the strict null hypothesis of no effect is tested at the $\alpha = 0.05$ level using several test statistics. We consider the following test statistics:

- Edge Diff: The difference of edge proportions within the treated and control groups, $T(R_2, R_0) = R_2/m_2 - R_0/m_0$, used in a two-tailed test.
- CF: The Mahalanobis statistic of Chen and Friedman used in a randomization test.
- CoD: The coefficient of determination from a QAP regression on indicators created for each level of $\mathbf{W} \in \{0, 1, 2\}^n$ used in a randomization test.

[3] The coefficient of determination is more commonly known as the "multiple R^2" and is used to simultaneously describe the strength of the linear relationship between the outcome and predictors while also explaining the percentage of variance in the outcome reduced by the fitted regression. Here we use the slightly longer title, "coefficient of determination," to disambiguate it from the squared total number of edges. We thank an anonymous reviewer for prompting us to consider this statistic.

[4] Reviews of community detection algorithms can be found in Schaeffer (2007), Fortunato (2010), Coscia et al. (2011), Nascimento and de Carvalho (2011), Fortunato and Castellano (2012), Harenberg et al. (2014), Amelio and Pizzuti (2014), and Bedi and Sharma (2016).



	Cluster 1	Cluster 2
Control	8	12
Treatment	11	9

(d)

Fig. 1. A graphical representation of using community detection to form a hypothesis test of the sharp null of no effects. Panel (a) shows an example network with treated (black) and control (gray) nodes. In panel (b), the treatment assignments are ignored and clustering is performed. In panel (c), treatment labels are returned and assignment-cluster totals are used to form panel (d).

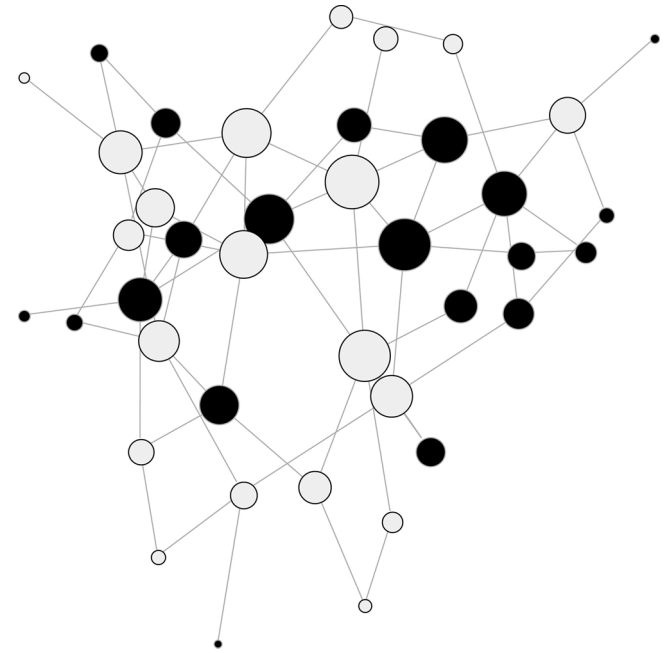


Fig. 2. Example network from Fig. 2 with node sizes proportional to the rank of eigenvector centrality.

- Clustering: A Fisher's exact test applied to the clusters resulting from spectral clustering.
- Centrality: A Wilcoxon–Mann–Whitney test applied to nodal eigenvector centrality.

Even though the size of the experiment is not large, the probability mass function (PMF) statistic required more Monte Carlo samples than were feasible with reasonable computation size. For a single experiment, adding more samples is not too onerous, but within a simulation with 500 replications, computation became untenable.

To compare the randomization and permutation methods with a well-known parametric approach, we tested the hypothesis of no effect using an exponential random graph model (ERGM). Similar to the methods that use R_2, R_1, and R_0, we fit the ERGM with a coefficient for each of $W_{ij}^{(k)}$, $k = 1, 2, 3$. We also include a term for the geometrically weighted edgewise shared partner (GWESP) distribution (Hunter, 2007). This term requires a decay parameter, which can be estimated within the ERGM (Hunter and Handcock, 2006) but is computationally expensive. For each parameter tested, we estimate the decay parameter on 10 networks without including the treatment indicator coefficients, then use the average estimated decay in the models fit on the 500 simulation replications. Again as a necessary computational step, each model is fit using maximum pseudo-likelihood. To test the hypothesis that $\mathbf{Y}$ is independent of $\mathbf{W}$, an F-test is performed to compare the fit of the full ERGM to one that only includes the GWESP term.

We considered five different data generating processes for the networks and parameterized these processes to vary between having no treatment effect and a large treatment effect on some aspect of the network. The five methods are

- An exponential random graph model in which treated units are more likely to form edges.
- A stochastic block model with blocks defined by treatment assignment with varying within and across block edge probabilities.
- A latent space model in which treatment status determines latent positions.

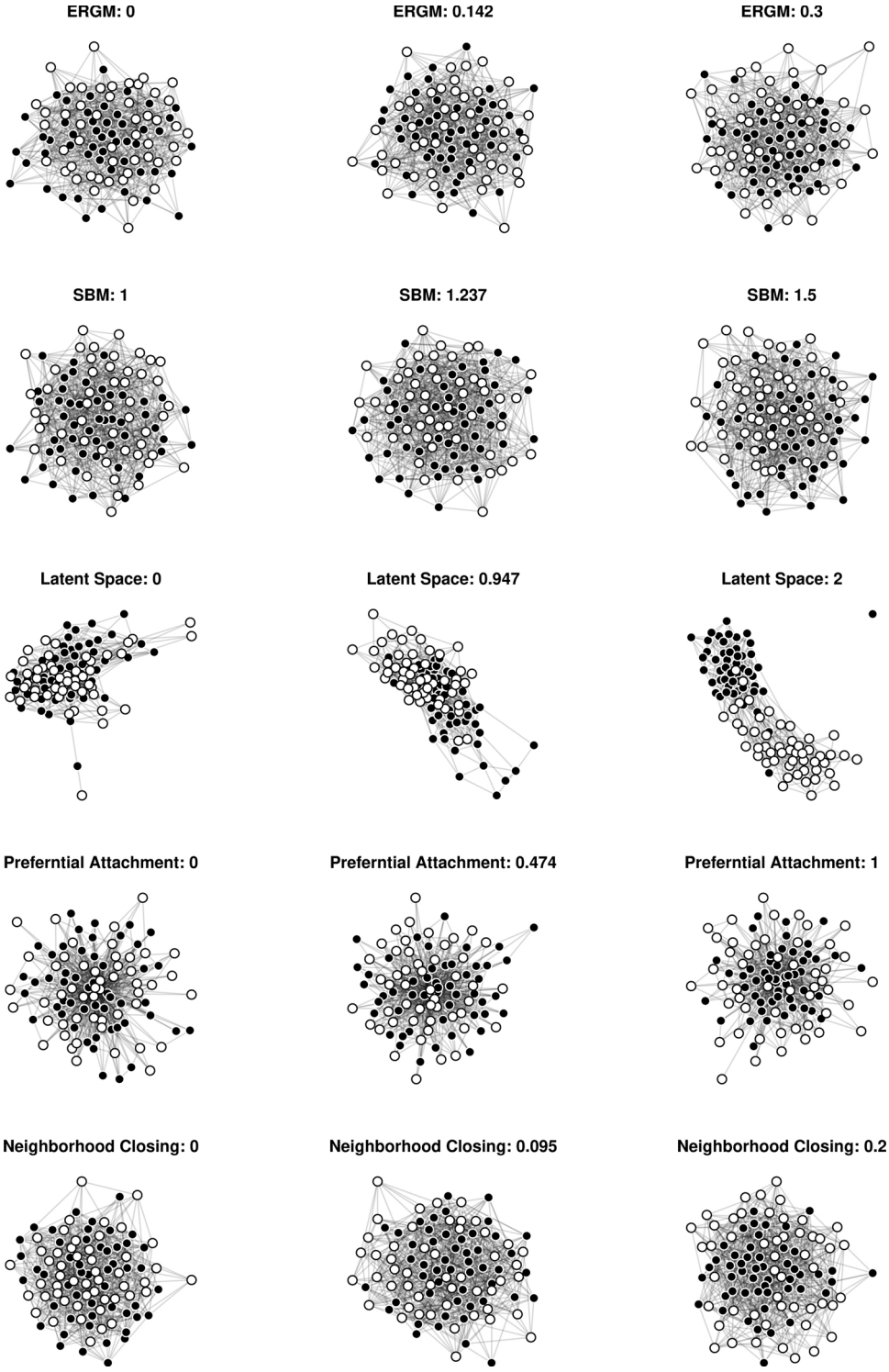


Fig. 3. Example networks for each data generating process used in the simulation studies. For each type of network, the left graph shows no treatment effect, the center graph shows moderate treatment effects, and the right graph shows strong treatment effects. The value after the title gives the exact parameter value, which corresponds to the horizontal axis in Fig. 4.

- A preferential attachment model in which treated units are more desired than control nodes.
- A process in which treated neighbors in an Erdös-Renyi graph are more likely to complete a triangle.

For all data generation methods, we tuned the processes so that graphs had similar densities, with about 15% of the possible dyads forming edges. Fig. 3 shows example networks for each data generating process with the relevant treatment effect parameter set to no treatment effect, a moderate treatment effect, and a large treatment effect. We describe the simulation techniques in more detail in the following paragraphs.

We begin our simulations with perhaps the most common network formation model: an exponential random graph model. To include a treatment effect, the linear index for edge (i, j) is given by $-2 + \beta(Z_i + Z_j - 1)$. The parameter β is varied from 0 to 0.3. As β increases, dyads with treated endpoints will be much more likely to

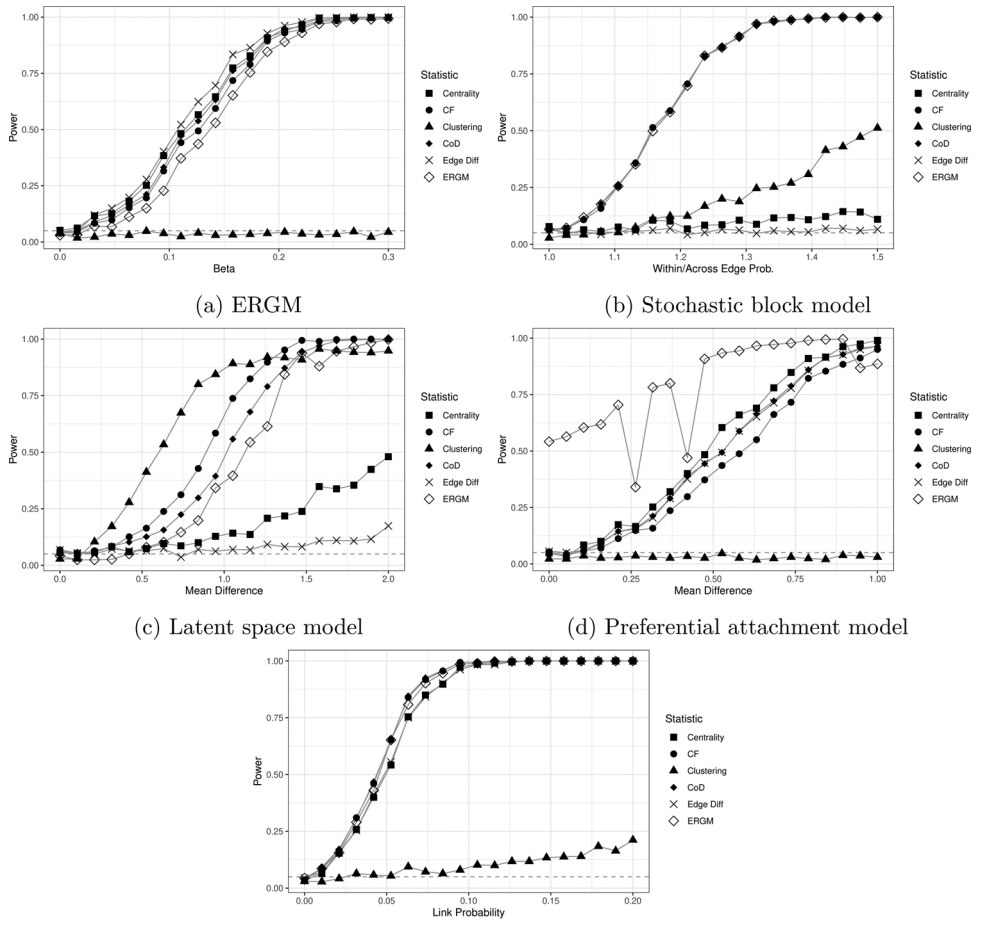


Fig. 4. Power plots for the five permutation test statistics and the ERGM parametric test for a variety of data generation methods. Each model is parameterized by the x-axis. The y-axis is the probability of rejecting the null hypothesis at the $\alpha = 0.05$ level.

form an edge compared to those without any treated nodes, making the treated units both more clustered and having more edges into the control group. By setting the intercept term to $-(2+\beta)$, the overall density of the graph will remain relatively consistent across parameter values. The model also includes a term to increase the number of triangles in the network, though this term does not depend on the treatment assignment. Fig. 4(a) shows the results of the simulation. When $\beta = 0$, the null hypothesis is in fact true, and all methods are able to meet their specified nominal level. With the exception of the clustering statistic, all methods exhibit similar power, rejecting nearly all networks as β approaches 0.3. Among these, edge count difference statistic performs slightly better than the others. The data generation process for these networks made treated units more likely to form edges with both treated and control nodes, and the spectral clustering statistic was unable to separate the group, generating effectively no power in this simulation.

It may seem surprising that the ERGM test is the least powerful among tests with non-negligible power, but note that the ERGM test is specified with a slightly broader model that includes specific terms for the control edges and the geometrically weighted edgewise shared partner distribution, neither of which was used in the simulated networks. In other simulations (not reported), testing using only the coefficient for the treated nodes leads to power comparable to the edge count difference statistic, but caused serious issues with Type I error in other simulations. The broader specification is more generally applicable, at the cost of some power in this specific example.

The second simulation generates the network from a stochastic block model (SBM). In this model, the treatment and control groups define two latent communities. All edges between nodes in the same community occur with probability p_{within}, independently. All across group edges occur with probability p_{across}. We parameterize the simulation by the ratio of these two probabilities $\theta = p_{\text{within}}/p_{\text{across}}$ and vary $\theta \in [1, 1.5]$. In order to keep the overall density of the network similar to the ERGM simulation, we fix the marginal probability of an edge at 0.15 and find p_{within} and p_{across} such that their ratio is θ. Fig. 4(b) shows the results of these simulations with θ on the x-axis. In this simulation, the ERGM, CF and CoD statistics perform virtually identically, all achieving the best power curve of any test. Somewhat surprisingly, the clustering based statistic only exhibits modest power. While we employed a spectral clustering method, it may be that other clustering techniques could better detect the SBM structure and reject the null hypothesis of no effect more frequently. The centrality statistic does not perform well in this simulation, which makes sense as treatment and control groups are symmetric in their edge probabilities. The edge difference test also performed poorly due to the effect of treatment operating equally on both R_2 and R_0. Again, all tests are able to meet their nominal Type I level.

In the third simulation, we use a one-dimensional latent space model. Each node i is given a location on the real line $X_i \sim N(Z_i\mu, 1)$, with μ varying from 0 to 2. The probability of an edge between any two nodes i and j is given by $(9-\mu)^{-1/2}\exp(-(X_i - X_j)^2)$. The leading $(9-\mu)^{-1/2}$ was selected to keep the overall number of edges similar to the ERGM and SBM simulations. Fig. 4(c) shows the clustering statistic generally outperforms the others. The CF statistic just edges out the QAP statistic. The centrality statistic exhibits some power, though it grows slowly as average distance between the treated and control

groups increases. In this simulation, the ERGM based test performs moderately well, with proper Type I error and power not much less than the best performing methods. Again the simple edge count difference statistic displays very little power due to the symmetry of the treated and control latent space distributions.

In the fourth simulation, we generate a "scale free" network where few nodes have very high degree and most nodes have very low degree. Pairs of nodes (i, j) are drawn with probabilities p_i and p_j, respectively, and an edge is formed between i and j. The process is repeated until the density of graph is approximately 0.15, again to match the previous simulations. Nodes with higher probabilities of being sampled will have many more neighbors than those with low probabilities. To assign probabilities, we use the latent positions X_i from the previous simulation. All nodes are ranked such that $a_i = 1$ implies that node i has the highest X_i and $a_i = n$ implies node i has the smallest X_i. Then $p_i \propto 1/a_i$, such that $\sum_{i=1}^{n} p_i = 1$. As the parameter μ, the difference between the means of the treatment and control latent distributions, increases, the most preferred nodes are increasingly composed of the treated group. Fig. 4(d) shows that the centrality statistic performs the best for this data generating process, with QAP, CF, and edge difference statistics also performing well. As in the first simulation, treated nodes will form edges with both treated and control units at a high rate, and accordingly the cluster statistic has no power at any value of θ. While the ERGM test appears to have one of the strongest power curve, this comes at a cost of rejecting the true null hypothesis at about 10 times the allowed Type I error rate. It would appear that the parametric assumptions of the test mistake unrelated network structure for treatment effects in this type of network.

In the fifth simulation, we take an algorithmic approach to network generation. First, we generate an Erdös-Renyi random graph with edge probability θ_1. After assigning treatment and control labels to the nodes, for all treated nodes i and j that do not already have an edge but do share at least one treated neighbor, a new edge is added with probability θ_2, which is varied from 0 to 0.3 over the simulation. Based on θ_2, we select θ_1 so that the overall density is again 0.15 in expectation. Fig. 4(e) shows similar results to the first simulation, with the ERGM fit performing slightly better. With some additional edges with the treated group, the clustering statistic exhibits some power, but much less than any of the other methods.

While not an exhaustive list of ways in which networks could be generated, the five selected models cover many of the most common approaches used in network analysis. Looking across these simulations, we see that both the CF Mahalanobis distance statistic and the coefficient of determination from a QAP regression are frequently the best performing, often having the greatest power or nearly greatest power. If researchers suspect that treatment induces a latent space model or a preferential attachment model, the clustering or centrality statistics would be a better choice. The parametric exponential random graph model generally had competitive power, but struggled in the face of deviations from the parametric assumptions in the latent space and preferential attachment models.

4. Data applications

4.1. Gene wide association study

Tsavachidou et al. (2009) conducted a 2×2 factorial randomized controlled trial to test the effect of selenium and vitamin E to combat the progression of prostate cancer. Both selenium and vitamin E had been identified in a previous observational study of prostate cancer as potentially having positive benefits. Subjects were recruited from patients scheduled to undergo a prostatectomy due to existing prostate cancer. Overall, 39 patients were recruited. After 3–6 weeks of treatment (placebo, selenium, vitamin E, or both), 39 subjects underwent surgery to remove their prostates. Cells were collected and subjected to expression assay. The original study selected cells in three different regions of the excised prostate: epithelial cells, stroma cells, and tumor cells. As only epithelial cells assays are available for all 39 patients, we focus on only those data in this analysis.

After collecting the microarray expression data, Tsavachidou et al. (2009) fit two-way ANOVA models for the two main effects as well as the interaction effect, assuming normally distributed error terms. With nearly 14,000 genes under study, the researchers applied a beta-uniform mixture model to control the false discovery rate at the 2% level. Comparing the placebo to selenium, vitamin E, and combination treatments, the researchers found 2109 differentially expressed genes, with 1329 of those significant comparisons coming from the selenium-placebo contrasts, and concluded that there were significant differences between the treatment conditions with respect to gene expression.

As an alternative to the parametric methods employed in the original publication, we apply the randomization inference network methods proposed in this paper. We create a gene co-expression network in which nodes are subjects and edges are present between subjects that have a similar pattern of gene expression, looking across all genes in the microarray assay. Within each subject, we rank all genes by expression level. Overall rates of expression may vary for subjects for idiosyncratic reasons; transforming expression levels into ranks within subjects allows for a common scale. We then compute the correlation of ranks between subjects. From these correlations, an edge is added between i and j if either i or j is in the other's top ten largest correlations. Fig. 5 shows the resulting network for the 39 subjects and 74 edges. Nodes are labeled by their treatment assignment.

After collapsing the treatment categories to subjects that received any selenium (the selenium and combination therapy groups) and those that did not (the vitamin E and pure control groups), we test the null hypothesis of no effect on the network using the randomization tests discussed in this paper: the difference of edges within the treatment and control groups, the coefficient of determination from a full QAP regression, the Mahalanobis statistic of Chen and Friedman, the probability mass function for the group edge counts, a Fisher's exact test applied to the results of spectral clustering, and a Wilcoxon–Mann–Whitney test applied to node eigenvector centrality. Table 1 reports the p-values for the sharp null hypothesis of no effects for the six statistics. The strongest result was found for the clustering statistic. If treatment truly had no effect, the observed concordance between cluster membership and treatment assignment group observed was extremely unlikely, suggesting that treatment lead to distinct patterns in gene expression. Fig. 6 shows the clusters found when using the clustering statistic. Visually, the treated units (black circles) largely separate from the control units (white squares). The p-value of 0.041 quantifies that this type of pattern would occur in very few random assignments, providing evidence against the sharp null of no effects.

The statistics based on edge counts were somewhat split on the evidence against the null. The two statistics that account the joint distributions of R_2 and R_0 directly, with the CF and PMF statistics, provided some evidence against the sharp null, but the simple edge difference and QAP coefficient of determination did not. The centrality statistic provided almost no evidence against the null, indicating treated and control units were indistinguishable with respect to centrality in the network.

Overall this pattern is most similar to the results reported for the latent space model used in the simulation, in which clustering, CF, and PMF were reasonably powerful, with the other statistics demonstrating substantially less power. Further analysis using latent space models may prove a useful way to capture a lower dimensional representation of the effect of treatment on this network.

4.2. Female representation on corporate boards

Seierstad and Opsahl (2011) studied female representation on 384 corporate boards in Norway over the period of May 2002–August 2011. On alternating months, they compiled lists of corporate boards and

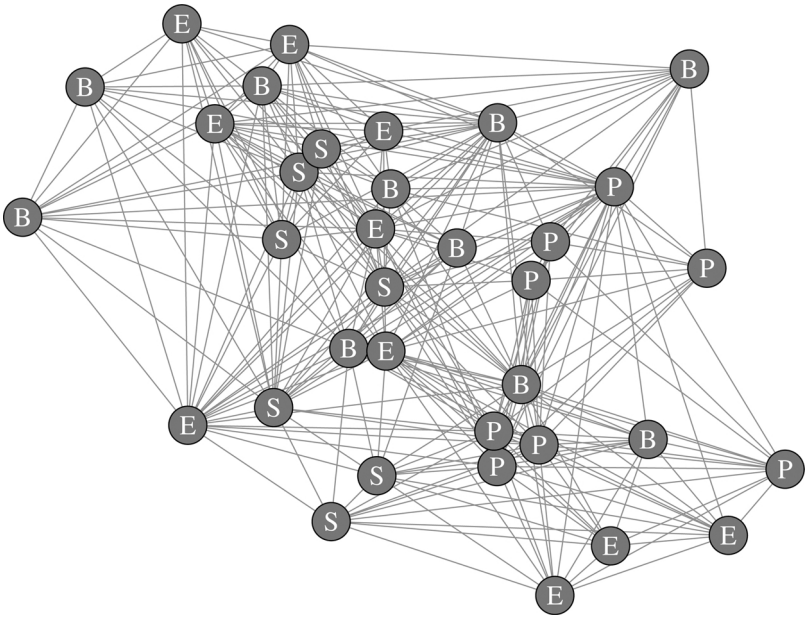


Fig. 5. The network derived from the gene expression data described in Tsavachidou et al. (2009) based on similarity of expression rates. Nodes are labeled by treatment assignment: (p)lacebo, vitamin (e), (s)elenium, or (b)oth.

Table 1
Hypothesis tests of the sharp null of no effects for the network derived from the gene-wide association study with randomly assigned selenium intake (Tsavachidou et al., 2009).

	p-Value
Edge difference	0.812
Coefficient of determination	0.244
Chen and Friedman	0.065
Probability mass function	0.059
Cluster	0.041
Centrality	0.883

matched first names to lists of names that have clear gender reference. Names that could not be easily matched were assigned a gender by investigating corporate web sites. Seierstad and Opsahl (2011) created networks of individuals with a link between any two people who served on the same board in the same month.

We investigate the network created by the union of networks of board members for the period of October 2010 to August 2011, comprising six individual networks. Fig. 7 shows the network with female members in white and male members in black, as well as the degree distribution. Perhaps most striking about this network is the large number of small components and one large central component. By construction, any node in this network must have a degree of at least one. 90% of nodes have a degree of 10 or less, with a few nodes having degree as high as 39.

Table 2 shows the results of testing the null hypothesis that gender labels can be shuffled uniformly at random. The non-linear edge count statistics show strong evidence against the null with the *p*-value for the CF and CoD statistics being equal to 1 in 100,000, the number of Monte Carlo samples used. In other words, following 100,000 Monte Carlo samples, none had higher CF or CoD statistics than those observed. The *p*-value for the PMF statistic was small, though not nearly as definitive. The edge difference statistic did not detect any evidence against the null, nor did the global clustering and centrality statistics. This pattern

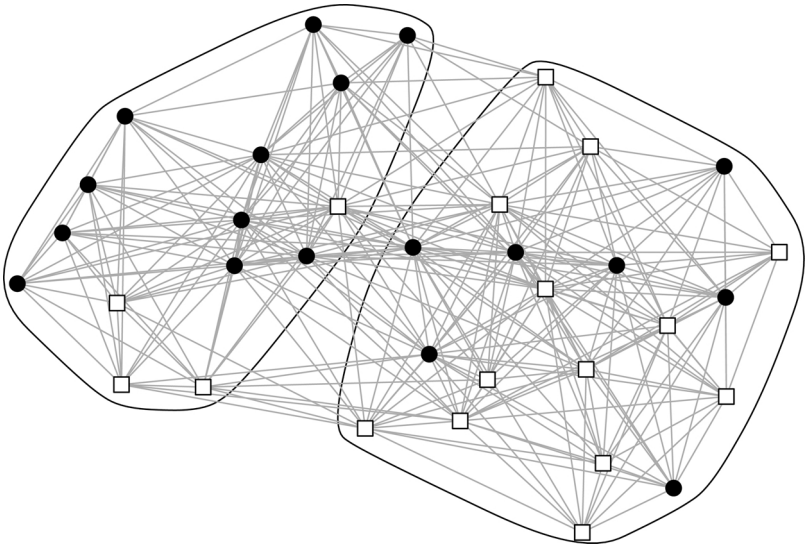
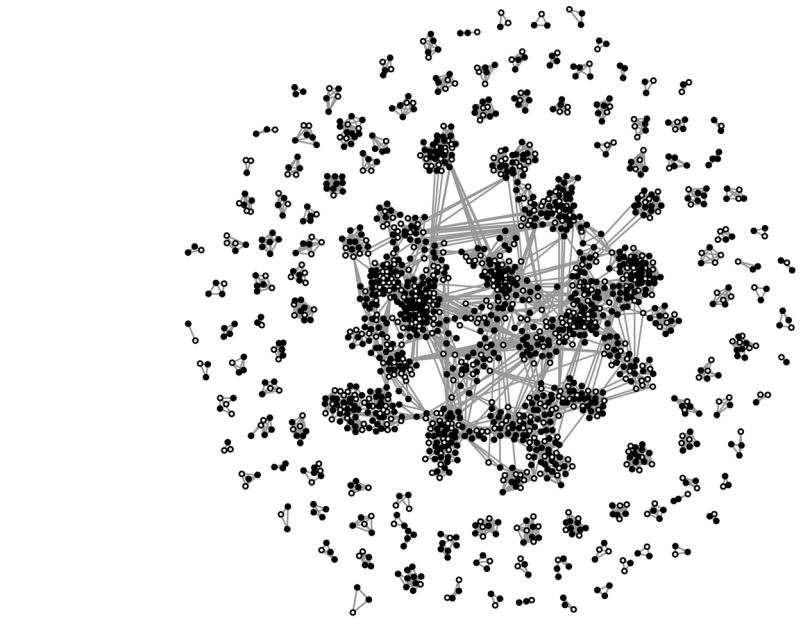
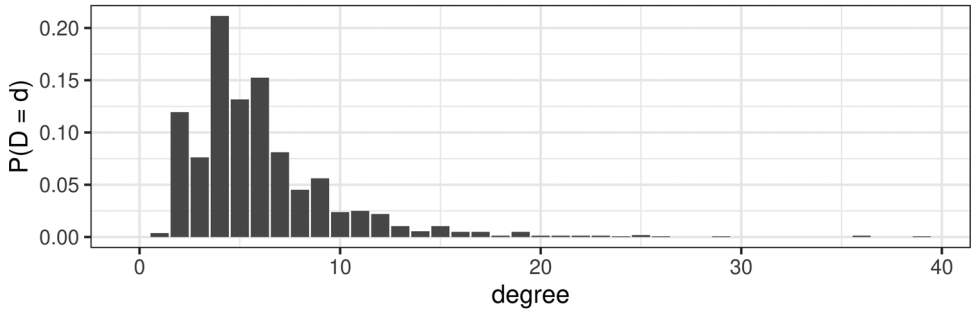


Fig. 6. Network of selenium and placebo subjects with clusters identified. Black circles are control units. White squares are treated units.



(a) Network



(b) Degree Distribution

Fig. 7. Gender co-membership on publicly listed boards in Norway in 2010 and 2011. Panel (a) shows the network itself. White nodes are female members and black nodes are male members. Board members share an edge if both members served on the same board at some point during the study period. Panel (b) shows the degree distribution.

Table 2
Hypothesis tests of the sharp null of no effects for the corporate board co-membership network of Seierstad and Opsahl (2011).

	p-value
Edge difference	0.443
Coefficient of determination	< 0.001
Chen and Friedman	< 0.001
Probability mass function	0.076
Cluster	0.867
Centrality	0.222

of results is suggestive of both the preferential attachment model in which treatment increases the desirability of treated units and the stochastic block model, though neither perfectly captures the pattern of evidence. Norwegian law requires at least one woman per corporate board, and this requirement may generate structure that sits somewhere between the preferential attachment and stochastic block models.

5. Discussion

In this paper we considered the role of randomization and permutation tests in the analysis of social networks. While permutation tests have a long history in network analysis, they are often perceived as being of limited investigative value. We demonstrated the utility of selecting statistics that possess power against meaningful alternatives and introduced several new statistics to provide researchers with more opportunities to test against specific alternatives. Many opportunities exist to harness other network statistics to be used in permutation and randomization tests.

Throughout the paper, we emphasized that randomization tests permit causal interpretations of networks when testing the specific null hypothesis that treatment had no effect. To be clear, the methods investigated in this paper present a particular model of causal events, in which the treatment assignment mechanism is known to precede the network. The potential outcomes model has proven useful in other areas of social and behavioral investigation, and we hope that this paper spurs additional developments under this framework in the network analysis context.

Consistent with literature outside of network analysis, our simulations showed that parametric methods can be quite powerful when the assumptions are met but can also perform poorly when those assumptions are violated. On the other hand, the permutation and randomization tests presented were able to maintain acceptable power, at least across multiple statistics, while consistently meeting expectations for Type I error rates. As discussed in the analysis section, rejecting hypotheses using non-parametric methods can then lead to modeling using parametric forms. This suggests permutation tests have a role to

play as a model selection tool in a larger network analysis research plan.

Acknowledgments

We thank Jake Bowers for useful comments during the development of this paper. We thank two anonymous reviewers and Professor Martin Everett for insightful comments during the reviewing process. This work was supported in part by National Science Foundation grants DMS-1406455 and DMS-1646108.

Appendix A. Moments of R_2, R_0

The test statistic T_{CF} requires computing the expected number of edges within the treated group (μ_2) and the control group (μ_0), as well as the variance–covariance matrix Σ_{20}, with entries σ_{ij}. Let R be the number of edges in the network, n the number of nodes, n_1 the number of units assigned to the treatment condition, and $n_0 = n - n_1$ the number of units assigned to control. Then

$$\mu_2 = R\frac{n_1(n_1-1)}{n(n-1)},$$

$$\mu_0 = R\frac{n_0(n_0-1)}{n(n-1)},$$

$$\sigma_2^2 = \mu_2(1-\mu_2) + C\frac{n_1(n_1-1)(n_1-2)}{n(n-1)(n-2)} + (R(R-1)-C)\frac{n_1(n_1-1)(n_1-2)(n_1-3)}{n(n-1)(n-2)(n-3)},$$

$$\sigma_0^2 = \mu_0(1-\mu_0) + C\frac{n_0(n_0-1)(n_0-2)}{n(n-1)(n-2)} + (R(R-1)-C)\frac{n_0(n_0-1)(n_0-2)(n_0-3)}{n(n-1)(n-2)(n-3)},$$

$$\sigma_{20} = (R(R-1)-C)\frac{n_1n_0(n_1-1)(n_0-1)}{n(n-1)(n-2)(n-3)} - \mu_2\mu_0,$$

where $C = \sum_{i=1}^{n} d_i^2 - \sum_{i=1}^{n} d_i$ and d_i is the degree of node i. A proof of these quantities is given in Chen and Friedman (2017).

Appendix B. QAP derivations

As noted in Dekker et al. (2007), under the restricted permutation of QAP or the methods proposed in this paper, the sample correlation between $\mathbf{Y}$ and $\mathbf{W}^{(k)}$, $k \in \{0, 1, 2\}$ is a linear function of the estimated parameter for $\mathbf{W}^{(k)}$ in a regression on $\mathbf{Y}$, for any valid permutation of $\mathbf{Y}$. In this appendix, we show that function operates through R_k, the total number of edges with treatment assignment $W_{ij} = k$. We also derive the relationship for the coefficient of determination — often called R^2, we use the longer title to avoid confusion with edge totals.

Recall the definitions $\mathbf{Y}'\mathbf{Y} = R$ and $\mathbf{Y}'\mathbf{W}^{(k)} = R_k$, with the property that $R = R_0 + R_1 + R_2$. We begin with the sample correlation between the network (as a vector of unique edges) and $\mathbf{W}^{(2)}$, the indicator for edges with two treated end points:

$$\mathrm{Cor}(\mathbf{Y}, \mathbf{W}^{(2)}) \propto \sum_{i<j} Y_{ij}(W_{ij}^{(2)} - \bar{W}^{(2)}).$$

There are $m = n(n-1)/2$ total units in $\mathbf{Y}$ and $\mathbf{W}^{(2)}$. There are $m_2 = n_1(n_1-1)/2$ entries where $W_{ij}^{(2)} = 1$, so

$$\bar{W}^{(2)} = \frac{m_2}{m}.$$

By definition, $\mathbf{Y}'\mathbf{W}^{(2)} = R_2$, so

$$\mathrm{Cor}(\mathbf{Y}, \mathbf{W}^{(2)}) \propto R_2 - \frac{m_2}{m}R.$$

Under the sharp null hypothesis of no effect, or equivalently the permutation hypothesis, R is a fixed quantity, so using R_2 is equivalent to the sample correlation as a test statistic (i.e., tests that reject for large values of the sample correlation will also reject for large values of R_2). Similar derivations show

$$\mathrm{Cor}(\mathbf{Y}, \mathbf{W}^{(0)}) \propto R_0 - \frac{m_0}{m}R,$$

$$\mathrm{Cor}(\mathbf{Y}, \mathbf{W}^{(1)}) \propto R_1 - \frac{m_1}{n},$$

where $m_0 = n_0(n_0-1)$ and $m_1 = m - m_2 - m_0 = n_1n_0$.

Bivariate QAP can also be used for linear combinations of R_2, R_1, and R_0. Let $V_{ij} = aW_{ij}^{(2)} + bW_{ij}^{(1)} + cW_{ij}^{(0)}$. The correlation of $\mathbf{Y}$ and $\mathbf{V}$ is proportional to

$$\mathrm{Cor}(\mathbf{Y}, \mathbf{V}) \propto \sum_{i<j} Y_{ij}(V_{ij} - \bar{V}),$$

where

$$\bar{V} = \frac{am_2 + bm_1 + cm_0}{m}.$$

By the same logic as the previous computations,

$$\mathrm{Cor}(\mathbf{Y}, \mathbf{V}) \propto aR_2 + bR_1 + cR_0 - \frac{am_2 + bm_1 + cm_0}{m}R.$$

In a regression of $\mathbf{Y}$ on the three indicators $\mathbf{W}^{(0)}$, $\mathbf{W}^{(1)}$, and $\mathbf{W}^{(2)}$ (with no intercept), let $\hat{\mathbf{Y}} = \hat{\beta}_0\mathbf{W}^{(0)} + \hat{\beta}_1\mathbf{W}^{(1)} + \hat{\beta}_2\mathbf{W}^{(2)}$. For any regression, the

coefficient of determination is

$$1 - \frac{(\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}})}{(\mathbf{Y} - \bar{Y}\mathbf{1})'(\mathbf{Y} - \bar{Y}\mathbf{1})}.$$

Expanding the numerator yields,

$$(\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = R - 2\mathbf{Y}'\hat{\mathbf{Y}} + \hat{\mathbf{Y}}'\hat{\mathbf{Y}}.$$

Define the design matrix for the regression as

$$\mathbf{X} = (\mathbf{W}^{(0)} \quad \mathbf{W}^{(1)} \quad \mathbf{W}^{(2)}).$$

Standard least squares results show that

$$\hat{\mathbf{Y}}'\hat{\mathbf{Y}} = [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}]'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\hat{\mathbf{Y}},$$

so overall

$$(\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = R - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

We break this second term into two pieces, $\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{X}'\mathbf{Y}$. Breaking the first term into its constituent parts,

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} m_0 & 0 & 0 \\ 0 & m_1 & 0 \\ 0 & 0 & m_2 \end{pmatrix}$$

and

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{\mathbf{W}^{(0)}}{m_0} & \frac{\mathbf{W}^{(1)}}{m_1} & \frac{\mathbf{W}^{(2)}}{m_2} \end{pmatrix}.$$

Pre-multiplying by $\mathbf{Y}'$,

$$\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{R_0}{m_0} & \frac{R_1}{m_1} & \frac{R_2}{m_2} \end{pmatrix}.$$

Separately, we see that

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} R_0 \\ R_1 \\ R_2 \end{pmatrix}.$$

Putting this all together,

$$\mathbf{Y}'\hat{\mathbf{Y}} = \frac{R_0^2}{m_0} + \frac{R_1^2}{m_1} + \frac{R_2^2}{m_2}.$$

The denominator from the coefficient of determination is

$$(\mathbf{Y} - \bar{Y}\mathbf{1})'(\mathbf{Y} - \bar{Y}\mathbf{1}) = R(1 - \frac{R}{m}).$$

Putting this together, we find the complete coefficient of determination,

$$\left(R - \frac{R^2}{m}\right)^{-1}\left(\frac{R_0^2}{m_0} + \frac{R_1^2}{m_1} + \frac{R_2^2}{m_2} - \frac{R^2}{m}\right).$$

For any test that rejects when the coefficient of determination exceeds some constant c, the test would also reject when

$$\frac{R_0^2}{m_0} + \frac{R_1^2}{m_1} + \frac{R_2^2}{m_2} > cR - (c-1)\frac{R^2}{m}.$$

References

Agresti, A., 1992. A survey of exact inference for contingency tables. Stat. Sci. 7 (1), 131–153.

Agresti, A., 2013. Categorical Data Analysis, third edition. John Wiley & Sons.

Amati, V., Lomi, A., Mira, A., 2018. Social network modeling. Annu. Rev. Stat. Appl. 5 (1), 343–369.

Amelio, A., Pizzuti, C., 2014. Overlapping community discovery methods: a survey. In: Gündüz-Öğüdücü, Ş., Etaner-Uyar, A.Ş. (Eds.), Social Networks: Analysis and Case Studies. Springer, Vienna, pp. 105–125.

Aronow, P.M., Samii, C., 2017. Estimating average causal effects under general interference, with application to a social network experiment. Ann. Appl. Stat. 11 (4), 1912–1947.

Athey, S., Eckles, D., Imbens, G.W., 2018. Exact *p*-values for network interference. J. Am. Stat. Assoc. 113 (521), 230–240.

Baker, F.B., Hubert, L.J., 1981. The analysis of social interaction data: a nonparametric technique. Sociol. Methods Res. 9 (3), 339–361.

Bedi, P., Sharma, C., 2016. Community detection in social networks. Wiley Interdiscipl. Rev. Data Mining Knowledge Discovery 6 (3), 115–135.

Berk, R.A., 2004. Regression Analysis: A Constructive Criticism. Sage Publications, Inc., Thousand Oaks, CA.

Bonacich, P., 1972. Factoring and weighting approaches to status scores and clique identification. J. Math. Sociol. 2 (1), 113–120.

Bonacich, P., 2007. Some unique properties of eigenvector centrality. Social Netw. 29 (4), 555–564.

Borgatti, S.P., 2005. Centrality and network flow. Social Netw. 27 (1), 55–71.

Borgatti, S.P., Everett, M.G., 2006. A graph-theoretic perspective on centrality. Social Netw. 28 (4), 466–484.

Bowers, J., Fredrickson, M.M., Panagopoulos, C., 2013. Reasoning about interference between units: a general framework. Polit. Anal. 21 (1), 97–124.

Chen, H., Friedman, J.H., 2017. A new graph-based two-sample test for multivariate and object data. J. Am. Stat. Assoc. 112 (517), 397–409.

Choi, D.S., 2017. Estimation of monotone treatment effects in network experiments. J. Am. Stat. Assoc. 112 (519), 1147–1155.
Coscia, M., Giannotti, F., Pedreschi, D., 2011. A classification for community discovery methods in complex networks. Stat. Anal. Data Mining: ASA Data Sci. J. 4 (5), 512–546.
Cox, D.R., 2006. Principles of Statistical Inference. Cambridge University Press.
Dekker, D., Krackhardt, D., Snijders, T.A.B., 2007. Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. Psychometrika 72 (4), 563–581.
Dow, M.M., de Waal, F.B., 1989. Assignment methods for the analysis of network subgroup interactions. Social Netw. 11 (3), 237–255.
Fienberg, S.E., 2012. A brief history of statistical models for network analysis and open challenges. J. Comput. Graph. Stat. 21 (4), 825–839.
Fienberg, S.E., Wasserman, S.S., 1981a. Categorical data analysis of single sociometric relations. Sociol. Methodol. 12, 156–192.
Fienberg, S.E., Wasserman, S.S., 1981b. An exponential family of probability distributions for directed graphs: comment. J. Am. Stat. Assoc. 76 (373), 54–57.
Fisher, R.A., 1935. The Design of Experiments. Oliver and Boyd, Edinburgh.
Fortunato, S., 2010. Community detection in graphs. Phys. Rep. 486 (3), 75–174.
Fortunato, S., Castellano, C., 2012. Community structure in graphs. In: Meyers, R.A. (Ed.), Computational Complexity: Theory, Techniques, and Applications. Springer, New York, NY, pp. 490–512.
Frank, O., 1977. Estimation of graph totals. Scand. J. Stat. 4 (2), 81–89.
Frank, O., 1978. Sampling and estimation in large social networks. Social Netw. 1 (1), 91–101.
Freedman, D.A., 2008a. On regression adjustments in experiments with several treatments. Ann. Appl. Stat. 2 (1), 176–196.
Freedman, D.A., 2008b. On regression adjustments to experimental data. Adv. Appl. Math. 40 (2), 180–193.
Freedman, D.A., 2008c. Randomization does not justify logistic regression. Stat. Sci. 23 (2), 237–249.
Freeman, G.H., Halton, J.H., 1951. Note on an exact treatment of contingency, goodness of fit and other problems of significance. Biometrika 38 (1–2), 141–149.
Freeman, L.C., 1978. Centrality in social networks conceptual clarification. Social Netw. 1 (3), 215–239.
Gibbons, J.D., Pratt, J.W., 1975. *P*-values: interpretation and methodology. Am. Stat. 29 (1), 20–25.
Goldenberg, A., Zheng, A.X., Fienberg, S.E., Airoldi, E.M., 2010. A survey of statistical network models. Foundat. Trends Mach. Learn. 2 (2), 129–233.
Good, P.I., 2005. Permutation, Parametric and Bootstrap Tests of Hypotheses, third edition. Springer, New York.
Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., Padmanabhan, K., Samatova, N., 2014. Community detection in large-scale networks: a survey and empirical evaluation. Wiley Interdiscipl. Rev.: Comput. Stat. 6 (6), 426–439.
Hernán, M.A., Robins, J.M., 2019. Causal Inference. Chapman & Hall/CRC.
Hirji, K.F., Johnson, T.D., 1996. A comparison of algorithms for exact analysis of unordered 2 × k contingency tables. Comput. Stat. Data Anal. 21 (4), 419–429.
Holland, P.W., 1986. Statistics and causal inference. J. Am. Stat. Assoc. 81 (396), 945–960.
Holland, P.W., Leinhardt, S., 1981. An exponential family of probability distributions for directed graphs. J. Am. Stat. Assoc. 76 (373), 33–50.
Hunter, D.R., 2007. Curved exponential family models for social networks. Social Netw. 29 (2), 216–230 (Special Section: Advances in Exponential Random Graph (p*) Models).
Hunter, D.R., Handcock, M.S., 2006. Inference in curved exponential family models for networks. J. Comput. Graph. Stat. 15 (3), 565–583.
Hunter, D.R., Krivitsky, P.N., Schweinberger, M., 2012. Computational statistical methods for social network models. J. Comput. Graph. Stat. 21 (4), 856–882.
Imbens, G.W., Rubin, D.B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.
Kolaczyk, E.D., 2009. Statistical Analysis of Network Data: Models and Methods. Springer New York.
Krackhardt, D., 1987. QAP partialling as a test of spuriousness. Social Netw. 9 (2), 171–186.
Krackhardt, D., 1988. Predicting with networks: nonparametric multiple regression analysis of dyadic data. Social Netw. 10 (4), 359–381.
Lehmann, E.L., 1975. Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, Inc., San Francisco.
Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27 (2 (Part 1)), 209–220.
Maritz, J.S., 1981. Distribution-Free Statistical Methods. Chapman and Hall, London.
Matous, P., Wang, P., 2019. External exposure, boundary-spanning, and opinion leadership in remote communities: a network experiment. Social Netw. 56, 10–22.
Nascimento, M.C., de Carvalho, A.C., 2011. Spectral methods for graph clustering – a survey. Eur. J. Oper. Res. 211 (2), 221–231.
Neyman, J.S., 1923. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat. Sci. 5 (4), 465–480 (Originally in Roczniki Nauk Tom X (1923) 1–51 (Annals of Agricultural Sciences). Translated from original Polish by Dambrowska and Speed.).
Nyblom, J., Borgatti, S., Roslakka, J., Salo, M.A., 2003. Statistical analysis of network data – an application to diffusion of innovation. Social Netw. 25 (2), 175–195.
O'Malley, A.J., 2013. The analysis of social network data: an exciting frontier for statisticians. Stat. Med. 32 (4), 539–555.
Radlow, R., Alf, E.F., 1975. An alternate multinomial assessment of the accuracy of the chi-squared test of goodness of fit. J. Am. Stat. Assoc. 70 (352), 811–813.
Rosenbaum, P.R., 2002. Observational Studies, second edition. Springer.
Rosenbaum, P.R., 2007. Interference between units in randomized experiments. J. Am. Stat. Assoc. 102 (477), 191–200.
Rosenbaum, P.R., 2010. Design of Observational Studies. Springer, New York.
Rubin, D.B., 1980. Randomization analysis of experimental data: the Fisher randomization test comment. J. Am. Stat. Assoc. 75 (371), 591–593.
Schaeffer, S.E., 2007. Graph clustering. Comput. Sci. Rev. 1 (1), 27–64.
Seierstad, C., Opsahl, T., 2011. For the few not the many? The effects of affirmative action on presence, prominence, and social capital of women directors in Norway. Scand. J. Manage. 27 (1), 44–54.
Snijders, T.A., 2011. Statistical models for social networks. Annu. Rev. Sociol. 37 (1), 131–153.
Tsavachidou, D., McDonnell, T.J., Wen, S., Wang, X., Vakar-Lopez, F., Pisters, L.L., Pettaway, C.A., Wood, C.G., Do, K.-A., Thall, P.F., Stephens, C., Efstathiou, E., Taylor, R., Menter, D.G., Troncoso, P., Lippman, S.M., Logothetis, C.J., Kim, J., 2009. Selenium and vitamin E: cell type- and intervention-specific tissue effects in prostate cancer. J. Natl. Cancer Inst. 101 (5), 306–320.
Whaley, F.S., 1983. The equivalence of three independently derived permutation procedures for testing the homogeneity of multidimensional samples. Biometrics 39 (3), 741–745.