SIAM Review
Vol. 61, No. 2, pp. 261–276

Network Modularity in the Presence of Covariates*

Beate Ehrhardt†
Patrick J. Wolfe‡

Abstract. We characterize the large-sample properties of network modularity in the presence of covariates, under a natural and flexible null model. This provides for the first time an objective measure of whether or not a particular value of modularity is meaningful. In particular, our results quantify the strength of the relation between observed community structure and the interactions in a network. Our technical contribution is to provide limit theorems for modularity when a community assignment is given by nodal features or covariates. These theorems hold for a broad class of network models over a range of sparsity regimes, as well as for weighted, multiedge, and power-law networks. This allows us to assign p-values to observed community structure, which we validate using several benchmark examples from the literature. We conclude by applying this methodology to investigate a multiedge network of corporate email interactions.

Key words. degree-based network models, limit theorems, network community structure, statistical network analysis

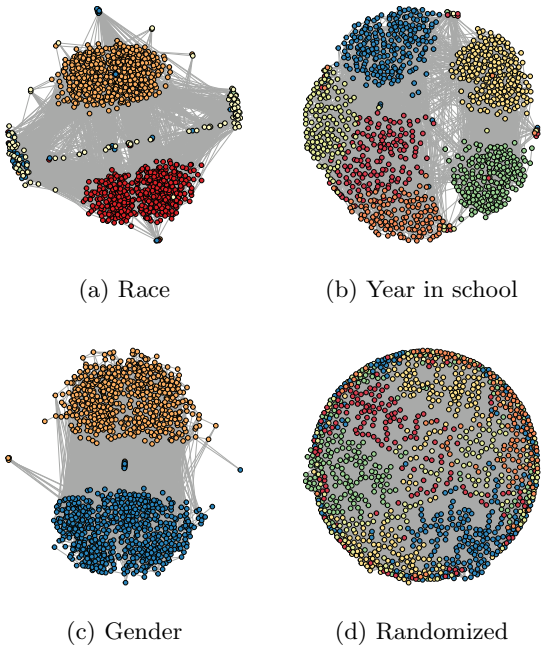
AMS subject classifications. 05C75, 62G20, 91D30

DOI. 10.1137/17M1111528

1. Introduction. A fundamental challenge in modern science is to understand and explain network structure, and in particular, the tendency of nodes in a network to connect in *communities* based on shared characteristics or function. Scientists inevitably observe not only network nodes and their connections, but also additional information in the form of covariates. Most analysis methods fail to exploit this information when attempting to explain network structure, and instead assign communities based solely on the network itself. This leads to a loss of interpretability and presents a barrier to understanding. We solve this problem by showing how to decide whether communities defined by covariates lead to a valid summary of network structure. In the student friendship network shown in Figure 1, for example, this means we can

*Received by the editors June 24, 2017; accepted for publication (in revised form) May 31, 2018; published electronically May 8, 2019.
http://www.siam.org/journals/sirev/61-2/M111152.html
Funding: This work was supported in part by the U.S. Army Research Office under Multidisciplinary University Research Initiative Award 58153-MA-MUR; by the U.S. Office of Naval Research under award N00014-14-1-0819; by the UK Engineering and Physical Sciences Research Council under Mathematical Sciences Established Career Fellowship EP/K005413/1 and Doctoral Training Grant EP/K502959/1; by the UK Royal Society under a Wolfson Research Merit Award; by Marie Curie FP7 Integration Grant PCIG12-GA-2012-334622 within the 7th European Union Framework Program; and by grants from the Simons Foundation and the Isaac Newton Institute for Mathematical Sciences (EP/K032208/1).

†Department of Statistical Science, University College London, London, UK (beate.franke.12@ucl.ac.uk).
‡Departments of Statistics, Computer Science, and Electrical and Computer Engineering, Purdue University, West Lafayette, IN (patrick@purdue.edu).



(a) Race (b) Year in school

(c) Gender (d) Randomized

Fig. 1 *A student friendship network illustrated for four different community assignments, each defined by a covariate* [9, 18, 23].

evaluate whether communities based on common gender, race, or year in school can explain the observed structure of the friendships.

The strength of community structure in networks is most often measured by modularity [19], which is intuitive and practically effective but until now has lacked a sound theoretical basis. We derive modularity from first principles, give it a formal statistical interpretation, and show why it works in practice. Moreover, by acknowledging that different community assignments may explain different aspects of a network's observed structure, we extend the applicability of modularity beyond its typical use to find a single "best" community assignment.

We use covariates to define community assignments and then prove that modularity quantifies how well these covariates explain network structure. We present a fundamental limit theorem for modularity in this context: in the presence of covariates, it behaves like a normal random variable for large networks whenever there is a lack of community structure. This allows us to translate modularity into a probability (a p-value), enabling for the first time its use to draw defensible, repeatable conclusions from network analysis.

Our main technical contribution is a flexible, nonparametric approach to quantify the strength of observed community structure. Most work assumes a single *unobserved* or latent community assignment (e.g., stochastic block models [4, 13, 24], latent space models [11, 12], and random dot product graphs [25, 27]; also, Traud et al. [26] analyze the correlation between covariates and the single, latent community structure). Hoff, Raftery, and Handcock [12], Newman and Clauset [18], and Zhang, Levina, and Zhu [28] all estimate latent community structure, while adjusting for the varying effects of covariates. Fosdick and Hoff [9] simultaneously model covariates and latent

structure, providing a test for independence. By contrast, we derive limit theorems to evaluate *observed* community structure implied by the covariates themselves.

Most recently (after the posting of this manuscript), Newman [17] derived a complementary interpretation of modularity, relating it to maximizing the likelihood of a degree-corrected stochastic block model. Other literature on modularity has typically focused on parametric statistical approaches. For example, the authors of [2] model all edges as equally likely Bernoulli random variables. By contrast, we take a nonparametric approach: using a single parameter per node, we model only the expectation of each edge [6]. This allows for individual node-specific differences but avoids specific distributional assumptions on the edges. Our results apply to a broad class of network models, allowing us to treat (among others) power-law networks [7], weighted networks, and those with multiple edges.

This paper is organized as follows. In section 2, we give a statistical model-based interpretation of modularity. We then present our main result: a methodology to quantify the explanatory power of covariates on the interactions in a network (section 3). Technically, we derive a limit theorem describing the large-sample behavior of modularity in the presence of covariates. In what follows, we explain this result. In section 4, we show that the model underlying modularity is a degree-based model for which we derive large-sample properties. We then deliver a bias-variance decomposition for modularity in section 5. After validating our method on four benchmark examples in section 6, we analyze email interactions in a multiedge corporate email network identifying those covariates that reflect the network structure (section 7). We finish with simulations demonstrating that modularity is robust as a measure of the strength of community structure (section 8).

2. Network Modularity in the Presence of Covariates. Two essential ingredients are necessary to understand modularity in the presence of covariates: first, a framework to allow for a formal interpretation of modularity as a measure of statistical significance; and second, the use of this framework to evaluate a covariate-based community assignment. We now describe each of these ingredients in turn.

First, to interpret modularity as a measure of statistical significance, we must recognize it as an estimator of a population quantity. Let $g(\cdot)$ denote an assignment of nodes into groups (i.e., communities), and write $\delta_{g(i)=g(j)} = 1$ when nodes i and j are assigned to the same group, and 0 otherwise. Denote by A_{ij} the strength of an edge (e.g., a count or a weight) between nodes i and j, and by $d_i = \sum_{j \neq i} A_{ij}$ the degree of the ith node. Then, modularity as defined in [19] is

$$\widehat{Q} = \sum_{j=1}^{n} \sum_{i<j} \left[A_{ij} - \frac{d_i d_j}{\sum_{l=1}^{n} d_l} \right] \delta_{g(i)=g(j)}. \tag{1}$$

Modularity contrasts an observed edge A_{ij} with the ratio $d_i d_j / \sum_l d_l$ whenever nodes i and j are in the same community. Now consider replacing $d_i d_j / \sum_l d_l$ by $\mathbb{E}\, A_{ij}$, the expected value of an edge under a given model:

$$Q = \sum_{j=1}^{n} \sum_{i<j} \left[A_{ij} - \mathbb{E}\, A_{ij} \right] \delta_{g(i)=g(j)}. \tag{2}$$

We recognize Q in (2) as a sum of signed residuals (observed minus expected values) $A_{ij} - \mathbb{E}\, A_{ij}$. If the model for each $\mathbb{E}\, A_{ij}$ posits the *absence* of community structure, then a large positive value of Q indicates the *presence* of such structure (more within-group edges than expected). Figure 1 illustrates this effect: the visible community

structure in Figures 1(a)–(c) is obscured in Figure 1(d) when communities are assigned at random. Moreover, using $d_i d_j / \sum_l d_l$ as a proxy for $\mathbb{E}\, A_{ij}$, we see that modularity $\widehat{Q}$ as defined in (1) is an estimator of Q in (2). We will return to this point in the next section.

Second, to interpret covariate-based community structure, we must recognize that different community assignments reveal different structural aspects of a network. Figures 1(a)–(c) illustrate this point using a student friendship network grouped by gender, race, and year in school. Covariates such as these define distinct community assignments, each of which relates the covariate in question to the observed network structure.

A key insight is that rather than maximizing modularity to obtain a single "best" community assignment, we may instead use modularity to measure the strength of an observed community structure. If a particular community assignment is given by a covariate, then modularity allows us to quantify the explanatory value of this covariate for the observed structure of the network.

3. Main Result: A Limit Theorem for Modularity. Our main result is a practical tool to understand objectively whether a covariate captures the structure of the interactions in a network. Technically, we derive a theorem quantifying the large-sample behavior of modularity in the above setting. In particular, if the null model of Definition 2 below is in force, then modularity in the presence of covariates behaves like a normal random variable. This enables us to associate a p-value with any observed community structure, quantifying how unlikely it is (under the null) to observe a community structure *at least as extreme as* the one we observe.

THEOREM 1 (central limit theorem for modularity). *Suppose the null model of Definition 2 below is in force, and consider a sequence of networks where for each n we observe a fixed (nonrandom) group assignment $g(1), g(2), \ldots, g(n)$. Then as long as the number of groups grows strictly more slowly than n, there exist constants b and s for each n such that as $n \to \infty$,*

$$\frac{\widehat{Q} - b}{s} \xrightarrow{d} \text{Normal}(0, 1).$$

Proof. Proofs of all results are given in the appendices. ∎

Thus, when appropriately shifted and scaled, modularity converges in distribution to a standard normal random variable. In what follows we explain this result and give explicit formulations for b and s^2 (see (4) and (5) below).

4. The Network Model Underlying Modularity. To understand Theorem 1, we must establish a technical foundation for modularity in the presence of covariates. Different models for the network edges A_{ij} will imply different estimators for Q in (2). Estimating Q using $\widehat{Q}$ in (1), we indirectly assume a model for the absence of community structure, where nodes connect independently based on the product of their individual propensities to form connections [6, 20, 21].

DEFINITION 2 (the network model underlying modularity). *Consider an undirected, random graph on n nodes without self-loops. We model its (possibly weighted) edges $A_{ij} \geq 0$ as independent random variables with expectations given by the product of node-specific parameters $\pi_1, \pi_2, \ldots, \pi_n > 0$:*

$$\mathbb{E}\, A_{ij} = \pi_i \pi_j, \quad 1 \leq i < j \leq n.$$

Furthermore, considering a sequence of such networks as n grows, we assume they are well behaved asymptotically:

1. *No single node dominates the network:* $\max_i \pi_i/\bar{\pi}$, *with* $\bar{\pi} = \frac{1}{n}\sum_{l=1}^{n} \pi_l$, *is bounded asymptotically.*
2. *The network is not too sparse:* $\min_i \pi_i \cdot \sqrt{n}$ *diverges as* n *grows.*
3. *The expectation of each edge* $\mathbb{E}\, A_{ij}$ *does not diverge too quickly as* n *grows:* $\max_i \pi_i/\sqrt{n}$ *goes to* 0.
4. *The variance of each edge does not vary too much from its expectation:* $\operatorname{Var} A_{ij}/\mathbb{E}\, A_{ij}$ *is bounded from above and away from* 0 *asymptotically.*
5. *The skewness of each edge* A_{ij} *is controlled: the third central moment* $\mathbb{E}\left[(A_{ij} - \mathbb{E}\, A_{ij})^3\right]$ *divided by the variance* $\operatorname{Var} A_{ij}$ *is bounded asymptotically.*

We make no further assumptions on the distribution of A_{ij}, and so our results apply in many settings, including weighted networks and those with multiple edges. Assumptions 1–3 are structural: the first excludes star-like networks; the second ensures that the network is not too sparse; and the third controls the growth of $\mathbb{E}\, A_{ij}$ with n in the weighted or multiedge setting. Assumptions 4 and 5 are technical; they exclude extreme behavior of the edge variables. For instance, both are fulfilled whenever $A_{ij} \sim \text{Bernoulli}\,(\pi_i\pi_j)$ or $A_{ij} \sim \text{Poisson}\,(\pi_i\pi_j)$.

Each parameter π_i describes the relative popularity of node i. Thus, to fit the degree-based model of Definition 2 to a network, we estimate the parameters π_i using the node's degrees d_i as follows [6, 20, 21]:

$$\hat{\pi}_i = \frac{d_i}{\sqrt{\sum_{l=1}^{n} d_l}}, \quad 1 \le i \le n. \tag{3}$$

The estimator $\hat{\pi}_i$ is both more natural and more computationally efficient than the corresponding maximum likelihood estimator for π_i, which follows from the theory of generalized linear models and cannot be written explicitly in closed form. In many settings the difference between these estimators is provably small [21], and so properties of maximum likelihood estimation can also be expected to hold for (3).

Most importantly, we show that any finite collection of estimators defined by (3) tends toward a multivariate normal distribution when n is large and Definition 2 is in force. This generalizes a univariate result in [20] which assumes $\text{Bernoulli}\,(\pi_i\pi_j)$ edges and a power-law degree distribution.

THEOREM 3 (multivariate central limit theorem for (3)). *Assume the model of Definition* 2 *and any finite set of estimators from* (3). *Relabeling the indices of these estimators from* 1 *to* r *without loss of generality, we have that as* $n \to \infty$,

$$\sqrt{\sum_{l=1}^{n} \mathbb{E}\, d_l}\left(\frac{\hat{\pi}_1 - \pi_1}{\sqrt{\operatorname{Var} d_1}}, \ldots, \frac{\hat{\pi}_r - \pi_r}{\sqrt{\operatorname{Var} d_r}}\right) \xrightarrow{d} \text{Normal}(0, \boldsymbol{I_r}).$$

Furthermore, $\sqrt{n \operatorname{Var} d_i / \sum_{l=1}^{n} \mathbb{E}\, d_l}$ *is bounded asymptotically and can be consistently estimated if* $A_{ij} \sim \text{Bernoulli}\,(\pi_i\pi_j)$ *or* $\text{Poisson}\,(\pi_i\pi_j)$ *by substituting* $\hat{\pi}$ *for* π *in* $\operatorname{Var} d_i$ *and* $\mathbb{E}\, d_i$.

From Definition 2 and (3), it is natural to define

$$\widehat{\mathbb{E}\, A_{ij}} = \hat{\pi}_i\hat{\pi}_j = \frac{d_i d_j}{\sum_{l=1}^{n} d_l}, \quad 1 \le i < j \le n.$$

Substituting $\widehat{\mathbb{E}\,A}_{ij}$ for $\mathbb{E}\,A_{ij}$ in (2), we immediately recognize modularity $\widehat{Q}$ as defined in (1). Thus, modularity implicitly assumes the degree-based model of Definition 2.

Moreover, $\widehat{\mathbb{E}\,A}_{ij} - \mathbb{E}\,A_{ij}$ converges in probability to zero under the model of Definition 2 (see the appendices). As a consequence of Theorem 3, we then obtain a central limit theorem for $\widehat{\mathbb{E}\,A}_{ij}$.

COROLLARY 4. *As $n \to \infty$ under the model of Definition 2,*

$$\frac{\widehat{\mathbb{E}\,A}_{ij} - \mathbb{E}\,A_{ij}}{\sqrt{\left(\pi_j^2 \operatorname{Var} d_i + \pi_i^2 \operatorname{Var} d_j\right) / \sum_{l=1}^n \mathbb{E}\,d_l}} \xrightarrow{d} \text{Normal}(0,1).$$

Furthermore, $\sqrt{[n/\,\mathbb{E}\,A_{ij}] \cdot \left(\pi_j^2 \operatorname{Var} d_i + \pi_i^2 \operatorname{Var} d_j\right) / \sum_{l=1}^n \mathbb{E}\,d_l}$ is bounded asymptotically, and can be consistently estimated by substituting $\hat{\pi}$ for π if $A_{ij} \sim \text{Bernoulli}\,(\pi_i\pi_j)$ or $A_{ij} \sim \text{Poisson}\,(\pi_i\pi_j)$.

This result leads to the first of two key insights as to why modularity, when appropriately shifted and scaled, behaves like a Normal$(0,1)$ random variable. Recall that $\widehat{Q}$ (see (1)) is an estimator for its population counterpart Q (see (2)), in which $\widehat{\mathbb{E}\,A}_{ij}$ estimates $\mathbb{E}\,A_{ij}$. Comparing (1) and (2), and approximating $\widehat{\mathbb{E}\,A}_{ij}$ by $\mathbb{E}\,d_i d_j / \sum_{l=1}^n \mathbb{E}\,d_l$, we obtain

$$\mathbb{E}(\widehat{Q} - Q) \approx \sum_{j=1}^n \sum_{i<j} \left(\mathbb{E}\,A_{ij} - \frac{\mathbb{E}\,d_i d_j}{\sum_{l=1}^n \mathbb{E}\,d_l}\right) \delta_{g(i)=g(j)}.$$

Under the model of Definition 2, this difference cancels to first order (see the appendices), yielding an approximate bias term of

$$b = \sum_{j=1}^n \sum_{i<j} \frac{\mathbb{E}\,A_{ij}\left(\mathbb{E}\,d_i + \mathbb{E}\,d_j - \sum_{l=1}^n \pi_l^2\right)}{\sum_{l=1}^n \mathbb{E}\,d_l} \delta_{g(i)=g(j)}. \tag{4}$$

This is precisely the shift term appearing in Theorem 1.

5. Modularity Reflects Within- and Between-Group Edges. Figure 2 illustrates the second main insight into the limiting behavior of modularity: its variability reduces asymptotically to that of a centered sum of within- and between-group edges.

More specifically, every network degree $d_i = \sum_{j \neq i} A_{ij}$ decomposes into within- and between-group components:

$$d_i = d_i^w + d_i^b;$$
$$d_i^w = \sum_{j \neq i} A_{ij}\delta_{g(i)=g(j)}, \qquad d_i^b = \sum_{j \neq i} A_{ij}\delta_{g(i)\neq g(j)}.$$

This decomposition is surprisingly powerful, in part because the model of Definition 2 asserts that d_i^w and d_i^b are statistically independent for any fixed group assignment $g(1), g(2), \ldots, g(n)$. After separating the systematic bias term b in modularity from its random variation, we obtain the following decomposition.

THEOREM 5 (bias-variance decomposition for modularity). *Under the model of Definition 2 and for a fixed (nonrandom) group assignment $g(1), g(2), \ldots, g(n)$, it holds that*

$$\widehat{Q} - b = \sum_{i=1}^n \alpha_i \left[d_i^w - \mathbb{E}\,d_i^w\right] + \sum_{i=1}^n \beta_i \left[d_i^b - \mathbb{E}\,d_i^b\right] + \epsilon,$$

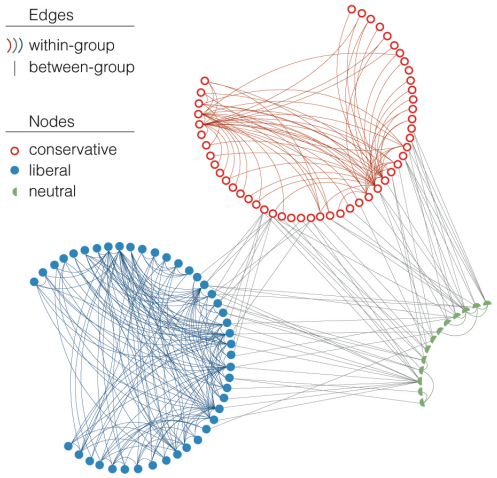


Fig. 2 *Within- and between-group edges in a network of political books frequently purchased together, where groups are defined by political alignment* [16]. *Note that only within-group edges appear in Q (see* (2)*); by contrast, both types of edges contribute to modularity $\widehat{Q}$ (see* (1)*).*

where ϵ is a random error term, $\alpha_i = 1/2 + \beta_i$, and

$$\beta_i = \left[\frac{1}{2}\frac{\sum_{l=1}^{n} \mathbb{E}\, d_l^w}{\sum_{l=1}^{n} \mathbb{E}\, d_l} - \frac{\mathbb{E}\, d_i^w}{\mathbb{E}\, d_i}\right], \quad 1 \le i \le n.$$

Theorem 5 quantifies the random variability inherent in modularity under the model of Definition 2. It establishes that a main term contributing to the variability of $\widehat{Q} - b$ in this setting is a linear combination of centered within- and between-group degrees (d_i^w, d_i^b), which for each i are statistically independent. The weights α_i and β_i associated with this linear combination are determined by the global proportion of expected within-group edges in the network, relative to the local proportion of expected within-group edges specific to node i.

Combining these two insights, we first shift modularity $\widehat{Q}$ by its approximate bias b and then scale it by the variance s^2 of $\sum_{i=1}^{n} \alpha_i \left[d_i^w - \mathbb{E}\, d_i^w\right] + \sum_{i=1}^{n} \beta_i \left[d_i^b - \mathbb{E}\, d_i^b\right]$:

$$s^2 = \sum_{j=1}^{n} \sum_{i<j} \left[\delta_{g(i)=g(j)} + \beta_i + \beta_j\right]^2 \operatorname{Var} A_{ij}. \tag{5}$$

Recalling Theorem 5, we then know that we are left with a linear combination of centered within- and between-group degrees that are now also scaled by s. This leads directly to a central limit theorem for modularity $\widehat{Q}$ as stated in Theorem 1:

$$\frac{\widehat{Q} - b}{s} \xrightarrow{d} \text{Normal}(0, 1).$$

6. Applying the Limit Theorem to Benchmark Examples. Having established a central limit theorem for modularity in the presence of covariates, we now show how to apply this result in practice. To turn our theory into a methodology suitable for a specific network dataset, we first need to elicit a model for the data based on Definition 2. We then fit this model, leading ultimately to a p-value based on Theorem 1. We now illustrate the complete analysis procedure for four binary networks

Table 1 *Four benchmark network datasets.*

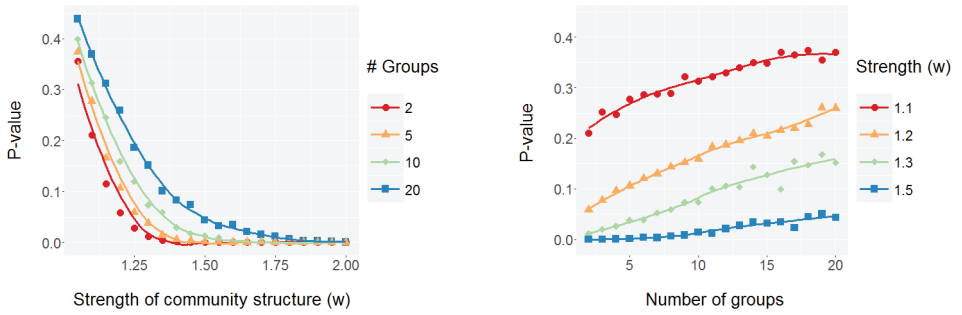
Dataset	Covariate	Nodes	Groups	Degree percentiles		
				25%	50%	75%
Books [16]	Political alignment	105	3	5	6	9
Jazz bands [10]	Recording location	198	17	16	25	39
Blogs [1]	Political alignment	1224	2	3	13	36
Coauthors [15]	Subject category	36297	7	2	5	10

Table 2 *Analysis of the data of Table* 1*, using modularity derived from covariate-based community assignments.*

Dataset	Simulated under the null					Data as observed		
(covariates as	$\widehat{Q}$	$(\widehat{Q}-\hat{b})/\hat{s}$		p-value		$\widehat{Q}$	$(\widehat{Q}-\hat{b})/\hat{s}$	p-value
in Table 1)	mean	mean	std.	mean	std.			
Books	2.60	0.02	1.01	0.51	0.29	189	21	$< 10^{-6}$
Jazz bands	6.67	0.01	1.02	0.51	0.29	552	29	$< 10^{-6}$
Blogs	23.20	0.01	1.04	0.50	0.30	6812	118	$< 10^{-6}$
Coauthors	14.64	0.00	1.00	0.50	0.29	73614	472	$< 10^{-6}$

which, along with their covariates, frequently serve as benchmarks for community detection [8, 16]. Tables 1 and 2 summarize all data and results.

1. First, we must further specify the null model of Definition 2, so that the parameter s^2 in (5) can be estimated. This can be done either by assuming sets of the variances $\operatorname{Var} A_{ij}$ to be equal, or by assuming a distribution for the edges A_{ij}. Since the benchmark networks we consider here are binary ($A_{ij} \in \{0,1\}$), we model their edges as
$$A_{ij} \sim \text{Bernoulli}(\pi_i \pi_j).$$
2. Second, we must assess whether the five asymptotic assumptions of Definition 2 appear to hold for our data and whether the number K of communities is sufficiently smaller than n (i.e., we assume $K/n \to 0$). Assumptions 3–5 are automatically satisfied for Bernoulli edges, and since all $K \ll n$ we are left to assess assumptions 1 ($\max_i \pi_i/\bar{\pi}$ bounded) and 2 ($\min_i \pi_i \cdot \sqrt{n}$ growing). We do this by substituting $\hat{\pi}_i$ for π_i, noting that $\max_i \hat{\pi}_i/\bar{\hat{\pi}} = \max_i d_i/\bar{d}$ and $\min_i \hat{\pi}_i \cdot \sqrt{n} = \min_i d_i/\sqrt{\bar{d}}$. Replacing $\min_i d_i$, $\bar{d}$, and $\max_i d_i$, respectively, by the first, second, and third degree quartiles as shown in Table 1, we observe that for all four benchmark networks, these ratios are of order one. This indicates that these networks are neither too star-like nor too sparse for Theorem 1 to apply.
3. Third, we estimate the parameters b and s necessary to shift and scale $\widehat{Q}$ in accordance with Theorem 1. To obtain an estimator $\hat{b}$, we substitute $\hat{\pi}$ for π in (4). The estimator $\hat{s}$ depends on the assumption added in step 1 above. Here, with $A_{ij} \sim \text{Bernoulli}\,(\pi_i \pi_j)$, we have
$$\operatorname{Var} A_{ij} = \pi_i \pi_j \left(1 - \pi_i \pi_j\right).$$
Then, $\hat{s}$ follows directly by substituting $\hat{\pi}$ for π in (5).
4. Finally, we compute and interpret the resulting approximate p-value. We first decide whether we want to test for an assortative or a disassortative community structure. We then define community assignments $g(1), g(2), \ldots, g(n)$



(a) By number of groups

(b) By strength of community structure

Fig. 3 *The relationship between p-value and the strength of community structure: p-values averaged over* 1000 *simulated networks each drawn from a degree-corrected stochastic block model* [14] *with the parameters* $\hat{\pi}_i$ *and size* n *of the network of books (Table* 1*).*

based on a covariate, and calculate $\widehat{Q}$ as per (1). We next estimate $(\widehat{Q} - b)/s$ using $\hat{b}$ and $\hat{s}$. Then, by Theorem 1, we compute an approximate one-sided p-value as follows:

$$\Pr\left(Z \geq \left|\frac{\widehat{Q} - \hat{b}}{\hat{s}}\right|\right), \quad Z \sim \text{Normal}(0,1). \tag{6}$$

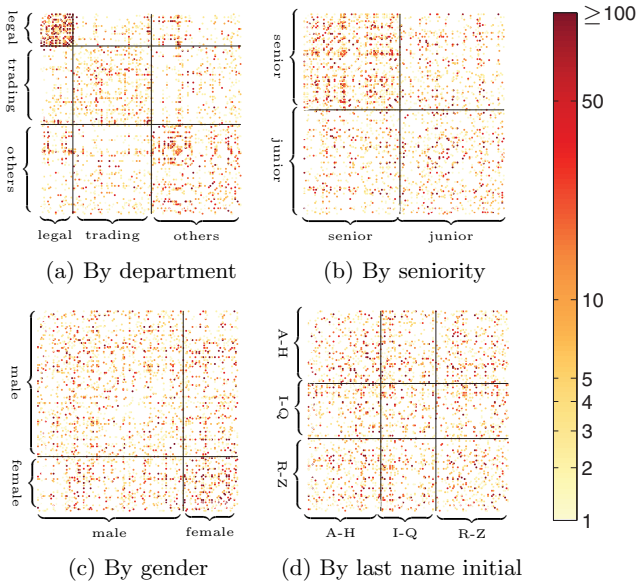
A small p-value implies that the observed value of modularity (or any larger value) is unlikely under the null.

Table 2 shows the results of applying this procedure to four benchmark datasets: a network of books [16] where books are connected if they have frequently been purchased together, categorized by political affiliation (see Figure 2); a network of jazz bands [10] where bands are connected if they have at least one band member in common, categorized by recording location; a network of political commentary websites (blogs) [1] where blogs are connected if they refer to each other, categorized by political affiliation; and a network of physicists [15] where physicists are connected if they have coauthored a manuscript, categorized by manuscript subject category.

The first conclusion of our benchmark analysis is as follows: when we fit the null model of Definition 2 to each of these four networks, and then simulate from the fitted model (parametric bootstrap), each simulated network results in (via (6)) a p-value with empirical mean near $1/2$ and standard deviation near $1/\sqrt{12}$. This empirical result aligns with Theorem 1, which predicts the p-values to be uniformly distributed with exactly that mean and standard deviation in the limit.

Our second conclusion is that, when using the observed data rather than simulated data under the null, each of the covariates leads (again via (6)) to a very small p-value ($< 10^{-6}$; see Table 2). This suggests that the data as observed are extremely unlikely under the null. Furthermore, since the null itself cannot explain any community structure, the conclusion we obtain agrees with the use of these covariates by other researchers as ground truth in community detection settings.

Figure 3 illustrates that the relationship between p-value and the strength of the community structure depends strongly on the effective sample size, here represented by the number of groups. Based on the parameters $\hat{\pi}_i$ and size n of the network of books (Table 1), we simulate from a degree-corrected stochastic block model [14]:



(a) By department (b) By seniority

(c) By gender (d) By last name initial

Fig. 4 *Multiedges A_{ij} in the Enron corporate email dataset (*153 *employees,* 32261 *pairwise email exchanges), grouped according to four different covariate-based community assignments. Shading indicates the number of emails exchanged.*

Table 3 *Analysis of the data of Figure* 4*, using modularity derived from multiple covariate-based community assignments.*

Covariate (no. groups)	$\widehat{Q}$	$\frac{\widehat{Q}-\hat{b}}{\hat{s}}$	p-value	
			Eq. (6)	Bootstrap
Department (3)	11454	6.17	$< 10^{-6}$	$< 10^{-6}$
Seniority (3)	6346	3.14	9×10^{-4}	8×10^{-6}
Gender (2)	5013	2.36	9×10^{-3}	2×10^{-3}
First name initial (17)	971	0.74	2×10^{-1}	2×10^{-1}
Last name initial (3)	−667	−0.46	7×10^{-1}	7×10^{-1}

$\mathbb{E}\, A_{ij} = \mathrm{w}\pi_i\pi_j$ when i and j are in the same community and $\mathbb{E}\, A_{ij} = \pi_i\pi_j$, otherwise. In Figure 3(a), we show for a fixed number of groups that as the strength of community structure w increases, the p-value decreases. However, for a fixed strength of community structure w, the p-value increases as the number of groups increases (see Figure 3(b)). These simulations demonstrate that the p-value quantifies the plausibility of the observed data under the null, but not the strength of community structure.

7. Evaluating Communities in a Multiedge Email Network. We now illustrate how our methodology can identify covariates that reflect a network's community structure. This analysis goes beyond the four benchmark examples considered above, where we validated our methodology but did not reach any new data-analytic conclusions. Here we evaluate the effects of employee *seniority*, *gender*, and *company department* on community structure in a multiedge corporate email network (see Figure 4). Table 3 summarizes all results, showing that each of these covariates results in a small p-value, while covariates based on grouping the *first* or *last name initials* of the em-

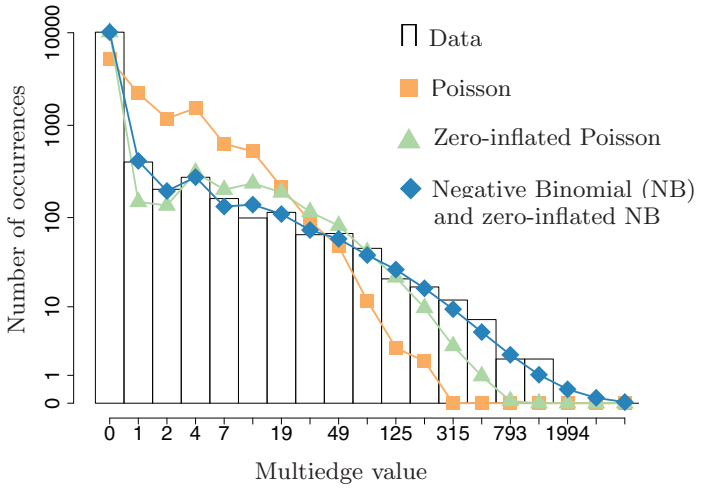


Fig. 5 *Observed versus expected email counts for maximum likelihood fits of four different models satisfying Definition* 2.

Table 4 *Goodness-of-fit versus model complexity for the models in Figure* 5 *(starting from the one-parameter model* Poisson (λ)*, relative to a saturated negative binomial (NB) model with* $r \to \infty$*).*

Model for the multiedges A_{ij}	Degrees of freedom	Residual deviance	Relative change
Poisson	153	142031	−39%
Zero-inflated Poisson	154	57070	−37%
Negative Binomial (NB)	154	12671	−19%
Zero-inflated NB	155	12671	0%

ployees do not. We will return to this analysis in more detail below, after describing the data and eliciting a suitable model.

This network and its covariates form a substantially richer dataset than those treated above. The data come from the Enron corporation [22]: as part of a U.S. government investigation following allegations of fraud, the email activities of senior employees from 1998–2002 were made public. Following the analysis in [22], we exclude all emails that have been sent en masse (to more than five recipients), leading to 32261 pairwise email exchanges between 153 employees. To model this network we will use the full flexibility afforded by Definition 2, following the four steps described in the previous section to determine a p-value corresponding to each covariate.

Step 1. To construct a suitable model for the observed multiedges A_{ij}, we compare four distributions satisfying the assumptions of Definition 2: Poisson$(\pi_i\pi_j)$, NegativeBinomial$(\pi_i\pi_j, r)$ with common shape parameter r, and zero-inflated versions of both. Figure 5 shows how well these distributions model the multiedges. Even without zero-inflation, the negative binomial (NB) distribution yields a good fit, particularly in the right tail. A formal model comparison via suitable likelihood ratio tests [5] confirms this: as Table 4 shows, the NB achieves the best balance between fitting the observed data (residual deviance) and model complexity (degrees of

freedom). We thus choose the model

$$A_{ij} \sim \text{NegativeBinomial}(\pi_i \pi_j, r). \tag{7}$$

Step 2. To verify the assumptions of Definition 2 for our data, we first assess assumptions 1 and 2 exactly as before. Computing quartiles Q_1–Q_3 of the degrees—$68, 200, 564$—we see that Q_3/Q_2 and $Q_1/\sqrt{Q_2}$ are both of order one. Assumption 3 ($\max_i \pi_i/\sqrt{n}$ shrinking) can be analogously assessed via $Q_3/(n\sqrt{Q_2})$. Assumptions 4 and 5 require $\operatorname{Var} A_{ij}/\mathbb{E}\, A_{ij} = 1 + \pi_i\pi_j/r$ and $\mathbb{E}\left[(A_{ij} - \mathbb{E}\, A_{ij})^3\right]/\operatorname{Var} A_{ij} = 1 + 2\pi_i\pi_j/r$ to be bounded. To assess this, we observe that a maximum likelihood estimate of r [5] yields $\hat{r} = 0.047$, while the first three quartiles of $\widehat{\mathbb{E}\, A_{ij}}$ are, respectively, $0.16, 0.59, 2.1$. The ratio of the number of communities K over n is below 0.02 for all covariates, but first name initial with $K/n = 0.1111$ (see Table 3 for values of K).

Step 3. To estimate b and s in Theorem 1, we substitute $\hat{\pi}_i$ for π_i in (4) and (5) exactly as before. Recall, however, that to estimate s we also require an estimate of $\operatorname{Var} A_{ij}$ in (5). Under the parametrization of (7), it follows that

$$\operatorname{Var} A_{ij} = \pi_i \pi_j \left(1 + \pi_i \pi_j / r\right). \tag{8}$$

Thus, $\operatorname{Var} A_{ij}$ can be estimated by substituting $\hat{\pi}_i$ for π_i and $\hat{r}$ for r in (8). This yields the required estimators $\hat{b}$ and $\hat{s}$.

Step 4. To calculate p-values, we must first compute $(\widehat{Q} - \hat{b})/\hat{s}$ for each covariate. Since we analyze more than one covariate, one must adjust the p-values for multiple comparisons. Below we consider a conservative Bonferroni correction. However, when considering many covariates, we recommend looking into more advanced multiple comparison adjustments, e.g., controlling the false discovery rate [3]. Prior to our analysis, we would expect that employee gender, seniority, and department might reflect aspects of community structure in email interactions. By contrast, we would expect covariates based on the first or last name of each individual to be noninformative. Figure 4 illustrates, in decreasing order of $(\widehat{Q} - \hat{b})/\hat{s}$, the observed structure of our data when grouped by covariate.

Table 3 reports two approximate p-values per covariate, in contrast to the previous section. The first of these derives (via (6)) from Theorem 1, which shows the limiting distribution of $(\widehat{Q} - \hat{b})/\hat{s}$ under the assumed model to be a standard normal. The second is based on 10^7 replicates of the parametric bootstrap, whereby we fit an NB model to the data and then simulate from the fitted values to obtain an empirical finite-sample distribution. Table 3 indicates that our asymptotic theory is somewhat conservative in this setting, leading as it does here to larger p-values than the bootstrap.

Finally, considering these p-values in more detail, we see from Table 3 that for the covariates of department, gender, and seniority, all p-values fall below 1% (leading to a corrected total of 5% after adjusting for multiple comparisons). In contrast, we obtain large p-values for first and last name covariates. This matches our expectations that department, gender, and seniority are likely to have an impact on email interactions, while there is no obvious reason why this should hold for name-related covariates.

8. Beyond the Theory: Robustness of Modularity. Simulations indicate that modularity as a measure of community structure (Theorem 1) is robust against devi-

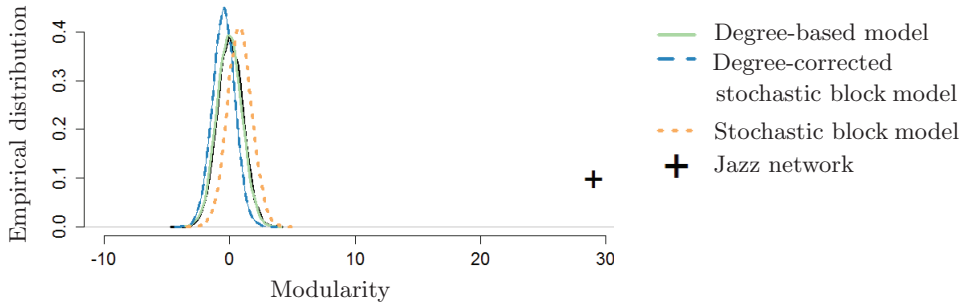
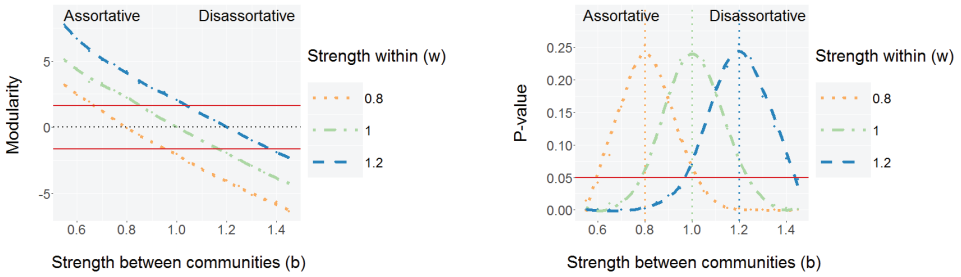


Fig. 6 *Robustness of modularity to the degree-based model assumption: modularity for simulated networks with increasing disagreement with the degree-based model, contrasted by the observed modularity of the network of jazz bands (see Table 1).*

ations from the degree-based model assumption. We have seen in the last two sections that the observed networks with informative covariates led to high modularity and low p-values. Since we checked all other assumptions, the reason can either be an informative community structure or a poor model fit of the degree-based model to the data. In Figure 6, we illustrate the impact on modularity of an increasing disagreement between simulated networks and the degree-based model measured in mean squared error (MSE). The networks are simulated from a degree-based model (MSE = 1435, solid green line), a degree-corrected stochastic block model (MSE = 2293, dashed blue line), and a stochastic block model (MSE = 2839, dotted orange line). For each model, we used the size, the expected density, and the covariate of the network of jazz bands (see Table 1). We see in all three cases that modularity is well approximated by a Normal$(0, 1)$ random variable despite the increasingly poor model fit and the fact that the observed modularity value of the network of jazz bands is highly unlikely under any of the empirical distributions.

In practice, we can assess the distribution of modularity for an observed network and its covariate using simulations. We compute modularity for the observed edges and the observed community structure. To ensure independence between the community structure and the edges, we randomly permute the rows of the adjacency matrix and apply the same permutation to the columns. We repeat the procedure for 1000 random permutations and assess the empirical distribution of modularity. For Theorem 1 to hold, the empirical distribution needs to be approximately Normal$(0, 1)$. Applying this empirical assessment to the networks in Table 1 showed that modularity is well approximated by a Normal$(0, 1)$ random variable in all four cases.

In contrast to the limitations of the original modularity $\widehat{Q}$, modularity as introduced here (i.e., $(\widehat{Q} - \hat{b})/\hat{s}$) can be used to assess the strength of both assortative and disassortative community structures. Figure 7 illustrates how modularity and the corresponding p-values change as we move from an assortative to a disassortative network. Modularity measures the divergence of the observed network from the degree-based model. For assortative networks, we observe more edges within communities than would be expected, leading to high modularity values. In contrast, in disassortative networks we observe fewer edges and low modularity values. Since under the degree-based model modularity is asymptotically normal distributed—a symmetric distribution—we see low p-values for both assortative and disassortative networks. The networks in Figure 7 are simulated from a degree-corrected stochastic block model [14]—$\mathbb{E}\, A_{ij} = w\pi_i\pi_j$ when i and j are in the same community, and



(a) Modularity

(b) p-values

Fig. 7 *Modularity and its corresponding p-value for assortative ($b \ll w$) and disassortative ($b \gg w$) networks simulated from a degree-corrected stochastic block model* [14]. *The red solid lines indicate when the community structure is significant. The dotted lines indicate when there is no community structure present ($b = w$).*

$\mathbb{E}\, A_{ij} = b\pi_i\pi_j$ otherwise—where the parameters $\hat{\pi}_i$, size n, and number of groups agree with the network of books (see Table 1).

9. Discussion. Networks have richer and more varied structure than can be described by a single "best" community assignment. To reflect this, we have introduced an approach which exploits the structural information captured by covariates, each of which may describe different aspects of community structure in the data. In contrast to community *detection* per se, this approach allows us to assess the significance of a given, interpretable community assignment with respect to the observed network structure. As described in the data analysis examples above, our method leads to the identification of structurally significant community assignments, ultimately yielding a better understanding of the network under study.

In technical terms, we have established a central limit theorem for modularity under a nonparametric null model, yielding p-values to assess the significance of observed community structure. The model we introduce shows explicitly how modularity measures variability in the data that cannot be explained solely by node-specific propensities for connection. What is more, modularity has more explanatory power than a classical (chi-squared) goodness-of-fit statistic: by aggregating the estimated *signed* residuals $A_{ij} - d_i d_j / \sum_l d_l$ within every network community, it measures the global tendency of a given community assignment to explain the observed network structure.

To advance the state of the art in network analysis, we as a research community must use this explanatory power to understand the effects of multiple observed communities on network structure, incorporating continuous covariates and combinations of covariates. Our work here represents a first step in this direction: we use the explanatory power of modularity to assess the significance of observed community structure relative to a null model. This opens the door to more advanced uses of multiple observed community assignments within formal statistical modeling frameworks. This is an important next step, since we see clear evidence here that multiple groupings may explain different aspects of a network's community structure.

Acknowledgments. The authors thank Leon Danon for sharing the data on jazz musicians from [10] and María Dolores Alfaro Cuevas for producing Figure 2. The

first author thanks Pierre-André Maugis, Sofia Olhede, Mason Porter, and Gesine Reinert for helpful discussions and feedback on the manuscript.

REFERENCES

[1] L. A. Adamic and N. Glance, *The political blogosphere and the* 2004 *U.S. election: Divided they blog*, in Proceedings of the 3rd International Workshop on Link Discovery, ACM Press, New York, 2005, pp. 36–43. (Cited on pp. 268, 269)

[2] E. Arias-Castro and N. Verzelen, *Community detection in dense random networks*, Ann. Statist., 42 (2014), pp. 940–969. (Cited on p. 263)

[3] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, J. Roy. Statist. Soc. Ser. B, 57 (1995), pp. 289–300. (Cited on p. 272)

[4] P. J. Bickel and P. Sarkar, *Hypothesis testing for automated community detection in networks*, J. Roy. Statist. Soc. Ser. B, 78 (2016), pp. 253–273. (Cited on p. 262)

[5] A. C. Cameron and P. K. Trivedi, *Econometric models based on count data: Comparisons and applications of some estimators and tests*, J. Appl. Econometrics, 1 (1986), pp. 29–53. (Cited on pp. 271, 272)

[6] F. Chung and L. Lu, *The average distances in random graphs with given expected degrees*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 15879–15882. (Cited on pp. 263, 264, 265)

[7] A. Clauset, C. R. Shalizi, and M. E. J. Newman, *Power-law distributions in empirical data*, SIAM Rev., 51 (2009), pp. 661–703, https://doi.org/10.1137/070710111. (Cited on p. 263)

[8] J. Duch and A. Arenas, *Community detection in complex networks using extremal optimization*, Phys. Rev. E, 72 (2005), art. 027104. (Cited on p. 268)

[9] B. K. Fosdick and P. D. Hoff, *Testing and modeling dependencies between a network and nodal attributes*, J. Amer. Statist. Assoc., 110 (2015), pp. 1047–1056. (Cited on p. 262)

[10] P. M. Gleiser and L. Danon, *Community structure in jazz*, Adv. Complex Syst., 6 (2003), pp. 565–573. (Cited on pp. 268, 269, 274)

[11] M. S. Handcock, A. E. Raftery, and J. M. Tantrum, *Model-based clustering for social networks*, J. Roy. Statist. Soc. Ser. A, 170 (2007), pp. 301–354. (Cited on p. 262)

[12] P. D. Hoff, A. E. Raftery, and M. S. Handcock, *Latent space approaches to social network analysis*, J. Amer. Statist. Assoc., 97 (2002), pp. 1090–1098. (Cited on p. 262)

[13] P. W. Holland, K. B. Laskey, and S. Leinhardt, *Stochastic blockmodels: First steps*, Soc. Netw., 5 (1983), pp. 109–137. (Cited on p. 262)

[14] B. Karrer and M. E. J. Newman, *Stochastic blockmodels and community structure in networks*, Phys. Rev. E, 83 (2011), 016107. (Cited on pp. 269, 273, 274)

[15] M. E. J. Newman, *The structure of scientific collaboration networks*, Proc. Natl. Acad. Sci. USA, 98 (2001), pp. 404–409. (Cited on pp. 268, 269)

[16] M. E. J. Newman, *Modularity and community structure in networks*, Proc. Natl. Acad. Sci. USA, 103 (2006), pp. 8577–8582. (Cited on pp. 267, 268, 269)

[17] M. E. J. Newman, *Equivalence between modularity optimization and maximum likelihood methods for community detection*, Phys. Rev. E, 94 (2016), art. 052315. (Cited on p. 263)

[18] M. E. J. Newman and A. Clauset, *Structure and inference in annotated networks*, Nature Commun., 7 (2016), art. 11863. (Cited on p. 262)

[19] M. E. J. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Phys. Rev. E, 69 (2004), art. 026113. (Cited on pp. 262, 263)

[20] S. C. Olhede and P. J. Wolfe, *Degree-Based Network Models*, preprint, https://arxiv.org/abs/1211.6537, 2012. (Cited on pp. 264, 265)

[21] P. O. Perry and P. J. Wolfe, *Null Models for Network Data*, preprint, https://arxiv.org/abs/1201.5871, 2012. (Cited on pp. 264, 265)

[22] P. O. Perry and P. J. Wolfe, *Point process modelling for directed interaction networks*, J. Roy. Statist. Soc. Ser. B, 75 (2013), pp. 821–849. (Cited on p. 271)

[23] M. D. Resnick, P. S. Bearman, R. W. Blum, K. E. Bauman, K. M. Harris, J. Jones, J. Tabor, T. Beuhring, R. E. Sieving, M. Shew, M. Ireland, L. H. Bearinger, and J. R. Udry, *Protecting adolescents from harm: Findings from the National Longitudinal Study on Adolescent Health*, J. Amer. Med. Assoc., 278 (1997), pp. 823–832. (Cited on p. 262)

[24] T. A. B. Snijders and K. Nowicki, *Estimation and prediction for stochastic blockmodels for graphs with latent block structure*, J. Classification, 14 (1997), pp. 75–100. (Cited on p. 262)

[25] D. L. Sussman, M. Tang, and C. E. Priebe, *Consistent latent position estimation and vertex classification for random dot product graphs*, IEEE Trans. Pattern Anal. Mach. Intell., 36 (2014), pp. 48–57. (Cited on p. 262)

[26] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, *Comparing community structure to characteristics in online collegiate social networks*, SIAM Rev., 53 (2011), pp. 526–543, https://doi.org/10.1137/080734315. (Cited on p. 262)
[27] S. Young and E. Scheinerman, *Random dot product graph models for social networks*, in Algorithms and Models for the Web-Graph, A. Bonato and F. R. K. Chung, eds., Lecture Notes in Comput. Sci. 4863, Springer-Verlag, Berlin, 2007, pp. 138–149. (Cited on p. 262)
[28] Y. Zhang, E. Levina, and J. Zhu, *Community detection in networks with node features*, Electron. J. Stat., 10 (2016), pp. 3153–3178. (Cited on p. 262)