The Annals of Applied Probability
2015, Vol. 25, No. 6, 3465–3510
DOI: 10.1214/14-AAP1080

COMMUNITY DETECTION IN SPARSE RANDOM NETWORKS

BY NICOLAS VERZELEN[1] AND ERY ARIAS-CASTRO[2]

INRA and University of California, San Diego

We consider the problem of detecting a tight community in a sparse random network. This is formalized as testing for the existence of a dense random subgraph in a random graph. Under the null hypothesis, the graph is a realization of an Erdős–Rényi graph on N vertices and with connection probability p_0; under the alternative, there is an unknown subgraph on n vertices where the connection probability is $p_1 > p_0$. In Arias-Castro and Verzelen [*Ann. Statist.* **42** (2014) 940–969], we focused on the asymptotically *dense* regime where p_0 is large enough that $np_0 > (n/N)^{o(1)}$. We consider here the asymptotically *sparse* regime where p_0 is small enough that $np_0 < (n/N)^{c_0}$ for some $c_0 > 0$. As before, we derive information theoretic lower bounds, and also establish the performance of various tests. Compared to our previous work [*Ann. Statist.* **42** (2014) 940–969], the arguments for the lower bounds are based on the same technology, but are substantially more technical in the details; also, the methods we study are different: besides a variant of the scan statistic, we study other tests statistics such as the size of the largest connected component, the number of triangles, and the number of subtrees of a given size. Our detection bounds are sharp, except in the Poisson regime where we were not able to fully characterize the constant arising in the bound.

1. Introduction. Community detection refers to the problem of identifying communities in networks, for example, circles of friends in social networks, or groups of genes in graphs of gene co-occurrences [Bickel and Chen (2009), Girvan and Newman (2002), Lancichinetti and Fortunato (2009), Newman (2006), Newman and Girvan (2004), Reichardt and Bornholdt (2006)]. Although fueled by the increasing importance of graph models and network structures in applications, and the emergence of large-scale social networks on the internet, the topic is much older in the social sciences, and the algorithmic aspect is very closely related to graph partitioning, a longstanding area in computer science. We refer the reader to the comprehensive survey paper of Fortunato (2010) for more examples and references.

Received August 2013; revised September 2014.

[1]Supported in part by the French Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration).

[2]Supported in part by a grant from the Office of Naval Research (N00014-13-1-0257).

MSC2010 subject classifications. 05C80, 62C20.

Key words and phrases. Community detection, detecting a dense subgraph, minimax hypothesis testing, Erdős–Rényi random graph, planted subgraph problem, scan statistic, largest connected component.

By community detection, we mean here something slightly different. Indeed, instead of aiming at extracting the community (or communities) from within the network, we simply focus on deciding whether or not there is a community at all. Therefore, instead of considering a problem of graph partitioning, or clustering, we consider a problem of testing statistical hypotheses. We observe an undirected graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ with $N := |\mathcal{V}|$ nodes. Without loss of generality, we take $\mathcal{V} = [N] := \{1, \dots, N\}$. The corresponding adjacency matrix is denoted $\mathbf{W} = (W_{i,j}) \in \{0,1\}^{N \times N}$, where $W_{i,j} = 1$ if, and only if, $(i, j) \in \mathcal{E}$, meaning there is an edge between nodes $i, j \in \mathcal{V}$. Note that $\mathbf{W}$ is symmetric, and we assume that $W_{ii} = 0$ for all i. Under the null hypothesis, the graph $\mathcal{G}$ is a realization of $\mathbb{G}(N, p_0)$, the Erdős–Rényi random graph on N nodes with probability of connection $p_0 \in (0,1)$; equivalently, the upper diagonal entries of $\mathbf{W}$ are independent and identically distributed with $\mathbb{P}(W_{i,j} = 1) = p_0$ for any $i \neq j$. Under the alternative, there is a subset of nodes indexed by $S \subset \mathcal{V}$ such that $\mathbb{P}(W_{i,j} = 1) = p_1$ for any $i, j \in S$ with $i \neq j$, while $\mathbb{P}(W_{i,j} = 1) = p_0$ for any other pair of nodes $i \neq j$. We assume that $p_1 > p_0$, implying that the connectivity is stronger between nodes in S, so that S is an assortative community. The subset S is not known, although in most of the paper we assume that its size $n := |S|$ is known. Let H_0 denote the null hypothesis, which consists of $\mathbb{G}(N, p_0)$ and is therefore simple. And let H_S denote the alternative where S is the anomalous subset of nodes. We are testing H_0 versus $H_1 := \bigcup_{|S|=n} H_S$. We consider an asymptotic setting where

$$N \to \infty, \qquad n = n(N) \to \infty, \qquad n/N \to 0, \qquad n/\log N \to \infty, \tag{1}$$

meaning the graph is large in size, and the subgraph is comparatively small, but not too small. Also, the probabilities of connection, $p_0 = p_0(N)$ and $p_1 = p_1(N)$, may change with N—in fact, they will tend to zero in most of the paper.

Despite its potential relevance to applications, this problem has received considerably less attention. We mention the work of Wang et al. (2008) who, in a somewhat different model, propose a test based on a statistic similar to the modularity of Newman and Girvan (2004); the test is evaluated via simulations. Sun and Nobel (2008) consider the problem of detecting a clique, a problem that we addressed in detail in our previous paper [Arias-Castro and Verzelen (2014)], and which is a direct extension of the "planted clique problem" [Alon, Krivelevich and Sudakov (1998), Dekel, Gurel-Gurevich and Peres (2011), Feige and Ron (2010)]. Rukhin and Priebe (2012) consider a test based on the maximum number of edges among the subgraphs induced by the neighborhoods of the vertices in the graph; they obtain the limiting distribution of this statistic in the same model we consider here, with p_0 and p_1 fixed, and n is a power of N, and in the process show that their test reduces to the test based on the maximum degree. Closer in spirit to our own work, Butucea and Ingster (2011) study this testing problem in the case where p_0 and p_1 are fixed. A dynamic setting is considered in Heard et al. (2010), Mongiovì et al. (2013), Park, Priebe and Youssef (2013) where the goal is to detect changes in the graph structure over time.

1.1. *Hypothesis testing.* We start with some concepts related to hypothesis testing. We refer the reader to Lehmann and Romano (2005) for a thorough introduction to the subject. A test ϕ is a function that takes $\mathbf{W}$ as input and returns $\phi = 1$ to claim there is a community in the network, and $\phi = 0$ otherwise. The (worst-case) risk of a test ϕ is defined as

$$\gamma_N(\phi) = \mathbb{P}_0(\phi = 1) + \max_{|S|=n} \mathbb{P}_S(\phi = 0), \tag{2}$$

where $\mathbb{P}_0$ is the distribution under the null H_0 and $\mathbb{P}_S$ is the distribution under H_S, the alternative where S is anomalous. We say that a sequence of tests (ϕ_N) for a sequence of problems $(\mathbf{W}_N)$ is asymptotically powerful (resp., powerless) if $\gamma_N(\phi_N) \to 0$ (resp., $\to 1$). We will often speak of a test being powerful or powerless when in fact referring to a sequence of tests and its asymptotic power properties. Then, practically speaking, a test is asymptotically powerless if it does not perform substantially better than any method that ignores the adjacency matrix $\mathbf{W}$, that is, guessing. We say that the hypotheses merge asymptotically if

$$\gamma_N^* := \inf_{\phi} \gamma_N(\phi) \to 1,$$

and that the hypotheses separate completely asymptotically if $\gamma_N^* \to 0$, which is equivalent to saying that there exists a sequence of asymptotically powerful tests. Note that if $\liminf \gamma_N^* > 0$, no sequence of tests is asymptotically powerful, which includes the special case where the two hypotheses are contiguous.

Our general objective is to derive the detection boundary for the problem of community detection. On the one hand, we want to characterize the range of parameters (n, N, p_0, p_1) such that either all tests are asymptotically powerless ($\gamma_N^* \to 1$) or no test is asymptotically powerful ($\liminf \gamma_N^* > 0$). On the other hand, we want to introduce asymptotically minimax optimal tests, that is, tests ϕ satisfying $\gamma_N(\phi) \to 0$ whenever $\gamma_N(\phi) \to 0$ or $\limsup \gamma_N^* < 1$ whenever $\limsup \gamma_N^* < 1$.

1.2. *Our previous work.* We recently considered this testing problem in Arias-Castro and Verzelen (2014), focusing on the *dense* regime where $\log(1 \vee (np_0)^{-1}) = o(\log(N/n))$ or equivalently $p_0 \geq n^{-1}(n/N)^{o(1)}$. (For $a, b \in \mathbb{R}$, $a \wedge b$ denotes the minimum of a and b and $a \vee b$ denotes their maximum.) We obtained information theoretic lower bounds, and we proposed and analyzed a number of methods, both when p_0 is known and when it is unknown. (None of the methods we considered require knowledge of p_1.) In particular, a combination of the total degree test based on

$$W := \sum_{1 \leq i < j \leq N} W_{i,j}, \tag{3}$$

and the scan test based on

$$W_n^* := \max_{|S|=n} W_S, \qquad W_S := \sum_{i,j \in S, i<j} W_{i,j}, \tag{4}$$

was found to be asymptotically minimax optimal when p_0 is known and when n is not too small, specifically $n/\log N \to \infty$. This extends the results that Butucea and Ingster (2011) obtained for p_0 and p_1 fixed (and p_0 known). In that same paper, we also proposed and studied a convex relaxation of the scan test, based on the largest n-sparse eigenvalue of $\mathbf{W}^2$, inspired by related work of Berthet and Rigollet (2013).

1.3. *Contribution*. Continuing our work, in the present paper we focus on the *sparse* regime where

$$p_0 \le \frac{1}{n}\left(\frac{n}{N}\right)^{c_0} \qquad \text{for some constant } c_0 > 0. \tag{5}$$

Obviously, (5) implies that $np_0 \le 1$. We define

$$\lambda_0 = Np_0, \qquad \lambda_1 = np_1, \tag{6}$$

and note that λ_0 and λ_1 may vary with N. Our results can be summarized as follows.

REGIME 1 ($\lambda_0 = (N/n)^\alpha$ with fixed $0 < \alpha < 1$). Compared to the setting in our previous work [Arias-Castro and Verzelen (2014)], the total degree test (3) remains a contender, scanning over subsets of size exactly n as in (4) does not seem to be optimal anymore, all the more so when p_0 is small. Instead, we scan over subsets of a wider range of sizes, using

$$W_n^{\ddagger} = \sup_{k=n/u_N}^{n} \frac{W_k^*}{k}, \tag{7}$$

where $u_N = \log\log(N/n)$. We call this the broad scan test. In analogy with our previous results in Arias-Castro and Verzelen (2014), we find that a combination of the total degree test (3) and the broad scan test based on (7) is asymptotically optimal when $\lambda_0 \to \infty$, in the following sense. Suppose $n = N^\kappa$ with $0 < \kappa < 1$. When $\kappa > \frac{1+\alpha}{2+\alpha}$, the total degree test is asymptotically powerful when $\lambda_1 \gg \frac{N^{(1+\alpha)/2}}{n^{1+\alpha}}$ and the two hypotheses merge asymptotically when $\lambda_1 \ll \frac{N^{(1+\alpha)/2}}{n^{1+\alpha}}$. [For two sequences of reals, (a_N) and (b_N), we write $a_N \ll b_N$ to mean that $a_N = o(b_N)$.] When $\kappa < \frac{1+\alpha}{2+\alpha}$, that is for smaller n, there exists a sequence of increasing functions ψ_n (defined in Theorem 1) such that the broad scan test is asymptotically powerful when $\liminf(1-\alpha)\psi_n(\lambda_1) > 1$ and the hypotheses merge asymptotically when $\limsup(1-\alpha)\psi_n(\lambda_1) < 1$. Furthermore, as $n \to \infty$, $\psi_n(\lambda) \asymp \lambda$ when $\lambda \ge 1$ remains fixed, while $\psi_n(1) \to 1$, and $\psi_n(\lambda) \sim \lambda/2$ for $\lambda \to \infty$. As a consequence, the broad scan test is asymptotically powerful when λ_1 is larger than (up to a numerical) $(1-\alpha)^{-1}$. See Table 1 for a visual summary. [For two real sequences, (a_N) and (b_N), we write $a_N \prec b_N$ to mean that $a_N = O(b_N)$, and $a_N \asymp b_N$ when $a_N \prec b_N$ and $a_N \succ b_N$.]

TABLE 1

Detection boundary and near-optimal algorithms in the regime $\lambda_0 = (N/n)^\alpha$ with $0 < \alpha < 1$ and $n = N^\kappa$ with $0 < \kappa < 1$. Here, "undetectable" means that the hypotheses merge asymptotically, while "detectable" means that there exists an asymptotically powerful test

κ	$\kappa < \frac{1+\alpha}{2+\alpha}$	$\kappa > \frac{1+\alpha}{2+\alpha}$
Undetectable	$\lambda_1 \prec (1-\alpha)^{-1}$; exact equation in (46)	$\lambda_1 \ll \frac{N^{(1+\alpha)/2}}{n^{1+\alpha}}$
Detectable	$\lambda_1 \succ (1-\alpha)^{-1}$; exact equation in (14)	$\lambda_1 \gg \frac{N^{(1+\alpha)/2}}{n^{1+\alpha}}$
Optimal test	Broad scan test	Total degree test

When $N^{-o(1)} \le \lambda_0 \le (N/n)^{o(1)}$ and $n = N^\kappa$ with $1/2 < \kappa < 1$, the total degree test is optimal, in the sense that it is asymptotically powerful for $\lambda_1^2/\lambda_0 \gg n^2/N$, while the hypotheses merge asymptotically for $\lambda_1^2/\lambda_0 \ll n^2/N$. This is why we assume in the remainder of this discussion that $n = N^\kappa$ with $0 < \kappa < 1/2$.

REGIME 2 ($\lambda_0 \to \infty$ with $\log(\lambda_0) = o[\log(N/n)]$). When $\kappa < \frac{1}{2}$, the broad scan test is asymptotically powerful when $\liminf \lambda_1 > 1$ and the hypotheses merge asymptotically when $\limsup \lambda_1 < 1$. See the first line of Table 2 for a visual summary.

REGIME 3 ($\lambda_0 > 0$ and $\lambda_1 > 0$ are fixed). The Poisson regime where λ_0 and λ_1 are assumed fixed is depicted on Figure 1. When $\lambda_1 > 1$, the broad scan test is asymptotically powerful. When $\lambda_0 > e$ and $\lambda_1 < 1$, no test is able to fully separate the hypotheses. In fact, for any fixed (λ_0, λ_1) a test based on the number of triangles has some nontrivial power [depending on (λ_0, λ_1)], implying that the two hypotheses do not completely merge in this case. The case where $\lambda_0 < e$ is not completely settled. No test is able to fully separate the hypotheses if $\lambda_1 < \sqrt{\lambda_0/e}$. The largest connected component test is optimal up to a constant when $\lambda_0 < 1$ and

TABLE 2

Detection boundary and near-optimal algorithms in the regimes $\lambda_0 \to \infty$ and $\lambda_0 \to 0$ and $n = N^\kappa$ with $0 < \kappa < 1/2$. For $1/2 < \kappa < 1$, the detection boundary occurs at $\lambda_1 \asymp N^{1/2}/n^2$ and is achieved by the total degree test

λ_0	$1 \ll \lambda_0 \ll (\frac{N}{n})^{o(1)}$	$\frac{1}{N^{o(1)}} \le \lambda_0 = o(1)$
Undetectable	$\limsup \lambda_1 < 1$	$\limsup \frac{\log(\lambda_1^{-1})}{\log(\lambda_0^{-1})} > \kappa$
Detectable	$\liminf \lambda_1 > 1$	$\liminf \frac{\log(\lambda_1^{-1})}{\log(\lambda_0^{-1})} < \kappa$
Optimal test	Largest CC test	Broad scan test

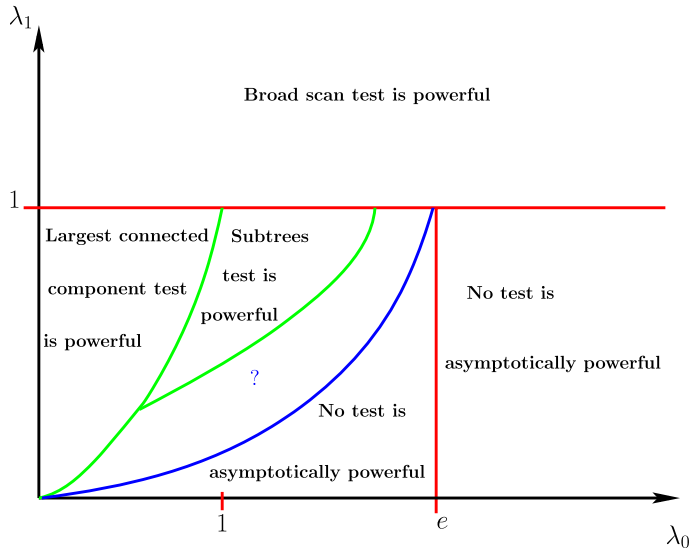

FIG. 1. *Detection diagram in the Poisson asymptotic where λ_0 and λ_1 are fixed and $n = N^\kappa$ with $0 < \kappa < 1/2$.*

a test based on counting subtrees of a certain size bridges the gap in constants for $1 \le \lambda_0 < e$, but not completely. When λ_0 is bounded from above and $\lambda_1 = o(1)$, the two hypotheses merge asymptotically.

REGIME 4 ($\lambda_0 = o(1)$ with $\log(1/\lambda_0) = o[\log(N)]$). Finally, when $\lambda_0 \to 0$, the largest connected component test is asymptotically optimal. See Table 2.

1.4. *Methodology for the lower bounds.* Compared to our previous work [Arias-Castro and Verzelen (2014)], the derivation of the various lower bounds here rely on the same general approach. Let $\mathbb{G}(N, p_0; n, p_1)$ denote the random graph obtained by choosing S uniformly at random among subsets of nodes of size n, and then generating the graph under the alternative with S being the anomalous subset. When deriving a lower bound, we first reduce the composite alternative to a simple alternative, by testing $H_0 : \mathbb{G}(N, p_0)$ versus $\bar{H}_1 := \mathbb{G}(N, p_0; n, p_1)$. Let L denote the corresponding likelihood ratio, that is, $L = \sum_{|S|=n} L_S / \binom{N}{n}$, where L_S is the likelihood ratio for testing H_0 versus H_S. Then these hypotheses merge in the asymptote if, and only if, $L \to 1$ in probability under H_0. A variant of the so-called "truncated likelihood" method, introduced by Butucea and Ingster (2011), consists in proving that $\mathbb{E}_0(\tilde{L}) \to 1$ and $\mathbb{E}_0(\tilde{L}^2) \to 1$, where $\tilde{L}$ is a truncated likelihood of the form $\tilde{L} = \sum_{|S|=n} L_S \mathbb{1}_{\Gamma_S} / \binom{N}{n}$, where Γ_S is a carefully chosen event. (For a set or event A, $\mathbb{1}_A$ denotes the indicator function of A.) An important difference with our previous work is the more delicate choice of Γ_S, which here relies

more directly on properties of the graph under consideration. We mention that we use a variant to show that H_0 and $\bar{H}_1$ do *not* separate in the limit. This could be shown by proving that the two graph models $\mathbb{G}(N, p_0)$ and $\mathbb{G}(N, p_0; n, p_1)$ are contiguous. The "small subgraph conditioning" method of Robinson and Wormald (1992, 1994) [see the more recent exposition in Wormald (1999)] was designed for that purpose. For example, this is the method that Mossel, Neeman and Sly (2012) use to compare a Erdős–Rényi graph with a stochastic block model[3] with two blocks of equal size. This method does not seem directly applicable in the situations that we consider here, in part because the second moment of the likelihood ratio, meaning $\mathbb{E}[L^2]$, tends to infinity at the limit of detection.

1.5. *Content*. The remaining of the paper is organized as follows. In Section 2, we introduce some notation and some concepts in probability and statistics, including concepts related to hypothesis testing and some basic results on the binomial distribution. In Section 3, we study a variant of the scan test and the test based on the size of the largest connected component. We refer the reader to the extended version [Verzelen and Arias-Castro (2013)] for other tests, such as the test based on the number of triangles, which always has some power in the Poisson regime, and a test based on counting the number of subtrees of a given size, which partially bridges the gap in constants in the Poisson regime that are near-optimal in different regimes. In Section 4, we state and prove information theoretic lower bounds on the difficulty of the detection problem. In Section 5, we discuss the situations where p_0 and/or n are unknown, as well as open problems.

2. Preliminaries. In this section, we first define some general assumptions and some notation, although more notation will be introduced as needed. We then list some general results that will be used multiple times throughout the paper.

2.1. *Assumptions and notation*. We recall that $N \to \infty$ and the other parameters such as n, p_0, p_1 may change with N, and this dependency is left implicit. Unless otherwise specified, all the limits are with respect to $N \to \infty$. We assume that $N^2 p_0 \to \infty$, for otherwise the graph (under the null hypothesis) is so sparse that number of edges remains bounded. Similarly, we assume that $n^2 p_1 \to \infty$, for otherwise there is a nonvanishing chance that the community (under the alternative) does not contain any edges. Throughout the paper, we assume that n and p_0 are both known, and discuss the situation where they are unknown in Section 5.

Define

$$\alpha = \frac{\log \lambda_0}{\log(N/n)}, \tag{8}$$

[3]This is a popular model of a network with communities, also known as the planted partition model. In this model, the nodes belong to blocks: nodes in the same block connect with some probability p_{in}, while nodes in different blocks connect with probability p_{out}.

which varies with N, and notice that $p_0 = \frac{\lambda_0}{N}$ with $\lambda_0 = (\frac{N}{n})^\alpha$. The dense regime considered in Arias-Castro and Verzelen (2014) corresponds to $\liminf \alpha \geq 1$. Here, we focus on the sparse regime where $\limsup \alpha < 1$. The case where $\alpha \to 0$ includes the Poisson regime where λ_0 is constant.

Recall that $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is the (undirected, unweighted) graph that we observe, and for $S \subset \mathcal{V}$, let $\mathcal{G}_S$ denote the subgraph induced by S in $\mathcal{G}$.

We use standard notation such as $a_N \sim b_N$ when $a_N/b_N \to 1$; $a_N = o(b_N)$ when $a_N/b_N \to 0$; $a_N = O(b_N)$, or equivalently $a_N \prec b_N$, when $\limsup_N |a_N/b_N| < \infty$; $a_N \asymp b_N$ when $a_N = O(b_N)$ and $b_N = O(a_N)$. We extend this notation to random variables. For example, if A_N and B_N are random variables, then $A_N \sim B_N$ if $A_N/B_N \to 1$ in probability.

For $x \in \mathbb{R}$, define $x_+ = x \vee 0$ and $x_- = (-x) \vee 0$, which are the positive and negative parts of x. For an integer n, let

$$n^{(2)} = \binom{n}{2} = \frac{n(n-1)}{2}. \tag{9}$$

Because of its importance in describing the tails of the binomial distribution, the following function—which is the relative entropy or Kullback–Leibler divergence of $\mathrm{Bern}(q)$ to $\mathrm{Bern}(p)$—will appear in our results:

$$H_p(q) = q \log\left(\frac{q}{p}\right) + (1-q) \log\left(\frac{1-q}{1-p}\right), \qquad p, q \in (0,1). \tag{10}$$

We let $H(q)$ denote $H_{p_0}(q)$.

2.2. *Calibration of a test.* We say that the test that rejects for large values of a (real-valued) statistic $T = T_N(\mathbf{W}_N)$ is asymptotically powerful if there is a critical value $t = t(N)$ such that the test $\{T \geq t\}$ has risk (2) tending to 0. The choice of t that makes this possible may depend on p_1. In practice, t is chosen to control the probability of type I error, which does not necessitate knowledge of p_1 as long as T itself does not depend on p_1, which is the case of all the tests we consider here. Similarly, we say that the test is asymptotically powerless if, for any sequence of reals $t = t(N)$, the risk of the test $\{T \geq t\}$ is at least 1 in the limit.

We prefer to leave the critical values implicit as their complicated expressions do not offer any insight into the theoretical difficulty or the practice of testing for the presence of a dense subgraph. Indeed, if a method can run efficiently, then most practitioners will want to calibrate it by simulation (permutation or parametric bootstrap, when p_0 is unknown). Besides, the interested reader will be able to obtain the (theoretical) critical values by a cursory examination of the proofs.

2.3. *Some general results.* Remember the definition of the entropy function in (10). The following is a simple concentration inequality for the binomial distribution.

LEMMA 1 (Chernoff's bound). *For any positive integer n, any $q, p \in (0,1)$, we have*

$$\mathbb{P}\big(\mathrm{Bin}(n,p) \geq qn\big) \leq \exp\big(-nH_p(q)\big). \tag{11}$$

Here are some asymptotics for the entropy function.

LEMMA 2. *Define $h(x) = x\log x - x + 1$. For $0 < p \leq q < 1$, we have*

$$0 \leq H_p(q) - ph(q/p) \leq O\Big(\frac{q^2}{1-q}\Big).$$

The following are standard bounds on the binomial coefficients. Recall that $e = \exp(1)$.

LEMMA 3. *For any integers $1 \leq k \leq n$,*

$$\Big(\frac{n}{k}\Big)^k \leq \binom{n}{k} \leq \Big(\frac{en}{k}\Big)^k. \tag{12}$$

Let $\mathrm{Hyp}(N,m,n)$ denotes the hypergeometric distribution counting the number of red balls in n draws from an urn containing m red balls out of N.

LEMMA 4. $\mathrm{Hyp}(N,m,n)$ *is stochastically smaller than* $\mathrm{Bin}(n,\rho)$, *where* $\rho := \frac{m}{N-m}$.

3. Some near-optimal tests. In this section, we consider several tests and establish their performances. We start by recalling the result we obtained for the total degree test, based on (3), in our previous work [Arias-Castro and Verzelen (2014)]. Recalling the definition of λ_0 and λ_1 in (6), define

$$\zeta := \frac{(p_1 - p_0)^2}{p_0}\frac{n^4}{N^2} = \frac{(\lambda_1 - \lambda_0 n/N)^2}{\lambda_0}\frac{n^2}{N}. \tag{13}$$

PROPOSITION 1 (Total degree test). *The total degree test is asymptotically powerful if $\zeta \to \infty$, and asymptotically powerless if $\zeta \to 0$.*

In view of Proposition 1, the setting becomes truly interesting when $\zeta \to 0$, which ensures that the naive total degree test is indeed powerless.

3.1. *The broad scan test.* In the denser regimes that we considered in Arias-Castro and Verzelen (2014), the (standard) scan test based on W_n^* defined in (4) played a major role. In the sparser regimes we consider here, the broad scan test based on $W_n^\ddagger$ defined in (7) has more power. Assume that $\liminf \lambda_1 > 1$, so that $\mathcal{G}_S$ is supercritical under H_S. Then it is preferable to scan over the largest connected component in $\mathcal{G}_S$ rather than scan $\mathcal{G}_S$ itself.

LEMMA 5. *For any $\lambda > 1$, let η_λ denote the smallest solution of the equation $\eta = \exp(\lambda(\eta - 1))$. Let $\mathcal{C}_m$ denote a largest connected component in $\mathbb{G}(m, \lambda/m)$ and assume that $\lambda > 1$ is fixed. Then, in probability, $|\mathcal{C}_m| \sim (1 - \eta_\lambda)m$ and $W_{\mathcal{C}_m} \sim \frac{\lambda}{2}(1 - \eta_\lambda^2)m$.*

PROOF. The bounds on the number of vertices in the giant component is well known [Van der Hofstad (2012), Theorem 4.8], while the lower bound on the number of edges comes from Pittel and Wormald (2005), Note 5. □

By Lemma 5, most of the edges of $\mathcal{G}_S$ lie in its giant component, which is of size roughly $(1 - \eta_{\lambda_1})n$. This informally explains why a test based on $W^*_{n(1-\eta_{\lambda_1})}$ is more promising that the standard scan test based on W^*_n.

In the details, the exact dependency of the optimal subset size to scan over seems rather intricate. This is why in $W^\ddagger_n$ we scan over subsets of size $n/u_N \le k \le n$. [Recall that $u_N = \log\log(N/n)$, although the exact form of u_N is not important.] For any subset $S \subset \mathcal{V}$, let

$$W^*_{k,S} = \max_{T \subset S, |T|=k} W_T.$$

Note that $W^*_{k,\mathcal{V}} = W^*_k$ defined in (4). Recall the definition of the exponent α in (8).

THEOREM 1 (Broad scan test). *The scan test based on $W^\ddagger_n$ is asymptotically powerful if*

$$\limsup \alpha \le 1 \quad \textit{and} \quad \liminf(1 - \alpha) \sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W^*_{k,S}]}{k} > 1; \tag{14}$$

or

$$\alpha \to 0 \quad \textit{and} \quad \liminf \lambda_1 > 1. \tag{15}$$

Note that the quantity $\sup_{k=n/u_N}^n \mathbb{E}_S[W^*_{k,S}]/k$ does not depend on p_0 or α. We shall prove in the next section that the power of the broad scan test is essentially optimal: if

$$\limsup \alpha < 1 \quad \text{and} \quad \limsup(1 - \alpha) \sup_{k=n/u_N}^{n} \mathbb{E}_S\big[W^*_{k,S}\big]/k < 1,$$

or $\alpha \to 0$ and $\limsup \lambda_1 < 1$, then no test is asymptotically powerful [at least when $n^2 = o(N)$, so that the total degree test is powerless]. Regarding (14), we could not get a closed-form expression of this supremum. Nevertheless, we show in the proof that

$$\liminf \sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W^*_{k,S}]}{k} \ge \liminf \frac{\lambda_1}{2}(1 + \eta_{\lambda_1}), \tag{16}$$

where η_λ is defined in Lemma 5. Moreover, we establish the following upper bound.

LEMMA 6.

$$\liminf \sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W^*_{k,S}]}{k} \le \liminf \frac{\lambda_1}{2} + 1 + \sqrt{1+\lambda_1}. \tag{17}$$

If $\lambda_1 \to \infty$, *then*

$$\sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W^*_{k,S}]}{k} \sim \lambda_1/2.$$

PROOF. Fix $\varepsilon > 0$ and define $x := 2[(1+\varepsilon) + \sqrt{(1+\varepsilon)^2 + \lambda_1(1+\varepsilon)}]$. First, we control the deviations of $W^*_{k,S}$. Define $q_k = (\lambda_1 + x)/(k-1)$ and notice that $q_k \ge p_1$ for $n/u_N \le k \le n$. Since $\log(1+t) \le t$ for any $t > -1$, we have

$$H_{p_1}(q_k) := q_k \log\left(\frac{q_k}{p_1}\right) + (1-q_k)\log\left(\frac{1-q_k}{1-p_1}\right) \ge q_k \log\left(\frac{q_k}{p_1}\right) - q_k + p_1.$$

Applying a union bound and Chernoff's inequality (11), we control the deviations of $W^*_{k,S}$:

$$\mathbb{P}_S[W^*_{k,S} \ge k^{(2)} q_k] \le \binom{n}{k} \exp[-k^{(2)} H_{p_1}(q_k)] \le \exp[kA_k],$$

where

$$A_k := \log\left(\frac{en}{k}\right) - \frac{k-1}{2}\left(q_k \log\left(\frac{q_k}{p_1}\right) - q_k + p_1\right).$$

Observe that x is larger than 2. As a consequence, we obtain

$$\begin{aligned} A_k &= 1 + \log\left(\frac{n}{k}\right) - \frac{\lambda_1 + x}{2}\log\left(\frac{n(\lambda_1+x)}{(k-1)\lambda_1}\right) + \frac{\lambda_1 + x}{2} - \frac{\lambda_1(k-1)}{2n} \\ &\le 1 + \frac{x}{2} - \frac{\lambda_1 + x}{2}\log\left(\frac{\lambda_1 + x}{\lambda_1}\right) - \frac{\lambda_1}{2}\left[\frac{k-1}{n} - 1 - \log\left(\frac{k-1}{n}\right)\right] \\ &\le 1 - \frac{x^2}{4(\lambda_1 + x)}, \end{aligned}$$

where we used in the last line the inequalities $t - \log t - 1 \ge 0$ and $\log(1-t) \le -t - t^2/2$, valid for any $t \ge 0$. By definition of x, we have $x^2/(4(\lambda_1 + x)) = 1 + \varepsilon$. In conclusion, we have proved that for any integer k between n/u_N and n

$$\mathbb{P}_S\left[\frac{W^*_{k,S}}{k} \ge \frac{\lambda_1 + x}{2}\right] \le \exp[-k\varepsilon]. \tag{18}$$

Let us now control the lower deviations of $\frac{1}{k}W^*_{k,S}$ using Lemma 7,

$$\mathbb{P}_S\left[\frac{W^*_{k,S}}{k} \le \mathbb{E}_S\left[\frac{W^*_{k,S}}{k}\right] - \left(\mathbb{E}_S\left[\frac{W^*_{k,S}}{k}\right]\right)^{1/2} \frac{8}{k^{1/2}}\right] \le 2^{-8}.$$

For k large enough, $\exp[-k\varepsilon] \le 1/2$, which therefore implies that

$$\mathbb{E}_S\left[\frac{W^*_{k,S}}{k}\right] \le \left(\mathbb{E}_S\left[\frac{W^*_{k,S}}{k}\right]\right)^{1/2} \frac{8}{(n/u_N)^{1/2}} + \frac{\lambda_1 + x}{2}, \tag{19}$$

since $k \ge n/u_N$. Taking the supremum over k and letting n go to infinity, we conclude that

$$\begin{aligned}\liminf \bigvee_{k=n/u_N}^{n} \mathbb{E}_S\left[\frac{W^*_{k,S}}{k}\right] &\le \liminf \frac{\lambda_1 + x}{2} \\ &= \liminf \frac{\lambda_1}{2} + (1+\varepsilon) + \sqrt{(1+\varepsilon)^2 + \lambda_1(1+\varepsilon)}.\end{aligned}$$

Then letting ε going to zero allows us to prove the first result.

Now assume that $\lambda_1 \to \infty$. From (19), we deduce that

$$\limsup \lambda_1^{-1} \bigvee_{k=n/u_N}^{n} \mathbb{E}_S\left[\frac{W^*_{k,S}}{k}\right] \le \frac{1}{2}.$$

On the other hand,

$$\bigvee_{k=n/u_N}^{n} \mathbb{E}_S\left[\frac{W^*_{k,S}}{k}\right] \ge \frac{\mathbb{E}_S[W^*_{n,S}]}{n} = \lambda_1 \frac{n-1}{2n} \sim \frac{\lambda_1}{2}.$$

This completes the proof. □

Hence, assuming α and λ_1 are fixed and positive, the broad scan test is asymptotically powerful when $(1-\alpha)\frac{\lambda_1}{2}(1+\eta_{\lambda_1}) > 1$. In contrast, the scan test was proved to be asymptotically powerful when $(1-\alpha)\frac{\lambda_1}{2} > 1$ [Arias-Castro and Verzelen (2014), Proposition 3], so that we have improved the bound by a factor larger than $1+\eta_{\lambda_1}$ and smaller than one $1 + 2\lambda_1^{-1}(1+\sqrt{1+\lambda_1})$. When α converges to one, it was proved in Arias-Castro and Verzelen (2014) that the minimax detection boundary corresponds to $(1-\alpha)\lambda_1/2 \sim 1$ [at least when $n^2 = o(N)$]. Thus, for α going to one, both the broad scan test and the scan test have comparable power and are essentially optimal. In the dense case, the broad scan test and the scan test have also comparable powers as shown by the next result which is the counterpart of Arias-Castro and Verzelen (2014), Proposition 3.

PROPOSITION 2. *Assume that p_0 is bounded away from one. The broad scan test is powerful if*

$$\liminf \frac{nH(p_1)}{2\log(N/n)} > 1.$$

The proof is essentially the same as the corresponding result for the scan test itself. See Arias-Castro and Verzelen (2014).

PROOF OF THEOREM 1. First, we control $W_n^{\ddagger}$ under the null hypothesis. For any positive constant $c_0 > 0$, we shall prove that

$$\mathbb{P}_0\big[(1-\alpha)W_n^{\ddagger} \geq 1 + c_0\big] = o(1). \tag{20}$$

Under conditions (14) and (15), α is smaller than for N large enough. Consider any integer $k \in [n/u_N, n]$, and let $q_k = 2(1 + c_0)/[(k-1)(1-\alpha)]$. Recall that $k^{(2)} = k(k-1)/2$. Applying a union bound and Chernoff's bound (Lemma 1), we derive that

$$\begin{aligned}\mathbb{P}_0\left[W_k^* \geq \frac{1+c_0}{1-\alpha}k\right] &\leq \binom{N}{k}\exp\big[-k^{(2)}H(q_k)\big] \\ &\leq \exp\left[k\left\{\log(eN/k) - \frac{k-1}{2}H(q_k)\right\}\right].\end{aligned}$$

We apply Lemma 2 knowing that $q_k/p_0 \to \infty$, and use the definition of α in (8), to control the entropy as follows:

$$\begin{aligned}\frac{k-1}{2}H(q_k) &\sim \frac{k-1}{2}q_k \log\left[\frac{q_k}{p_0}\right] \\ &= \frac{1+c_0}{1-\alpha}\big[\log(N/n) - \log\lambda_0 + O(\log u_N)\big] \\ &\sim (1+c_0)\log(N/n),\end{aligned}$$

since $\log(u_N) = o(\log(N/n))$. Consequently,

$$\mathbb{P}_0\left[W_k^* \geq \frac{1+c_0}{1-\alpha}k\right] \leq \exp\big[-kc_0\log(N/n)\big(1+o(1)\big)\big],$$

where the $o(1)$ is uniform with respect to k. Applying a union bound, we conclude that

$$\mathbb{P}_0\big[(1-\alpha)W_n^{\ddagger} \geq 1+c_0\big] \leq \sum_{k=n/u_N}^{n} \exp\big[-kc_0\log(N/n)\big(1+o(1)\big)\big] = o(1).$$

We now lower bound $W_n^{\ddagger}$ under the alternative hypothesis. First, assume that (14) holds, so that there exist a positive constant c and a sequence of integers $k_n \geq n/u_N$ such that $\mathbb{E}_S[W_{k_n,S}^*] \geq k_n(1+c)/(1-\alpha)$ eventually. In particular, $\mathbb{E}_S[W_{k_n,S}^*] \to \infty$. We then use (22) in the following concentration result for $W_{k,S}^*$.

LEMMA 7. *For an integer $0 \leq k \leq n$, define $\mu_{k,S}^* = \mathbb{E}_S[W_{k,S}^*]$. We have the following deviation inequalities*:

$$\mathbb{P}_S\big[W_{k,S}^* \geq \mu_{k,S}^* + t\big] \leq \exp\left[-\frac{\log(2)}{4}\left\{t \wedge \frac{t^2}{8\mu_{k,S}^*}\right\}\right] \qquad \forall t > 8\big(1 \vee \sqrt{\mu_{k,S}^*}\big) \tag{21}$$

and

$$\mathbb{P}_S\big[W^*_{k,S} \le \mu^*_{k,S} - t\big] \le \exp\Big[-\log(2)\frac{t^2}{8\mu^*_{k,S}}\Big] \qquad \forall t > 4\sqrt{\mu^*_{k,S}}. \tag{22}$$

It follows from Lemma 7 that, with probability going to one under $\mathbb{P}_S$,

$$W^\ddagger_n \ge \frac{W^*_{k_n}}{k_n} \ge \frac{W^*_{k_n,S}}{k_n} \ge \frac{1+c/2}{1-\alpha}.$$

Taking $c_0 = c/4$ in (20) allows us to conclude that the test based on $W^\ddagger_n$ with threshold $\frac{1+c/2}{1-\alpha}$ is asymptotically powerful.

Now, assume that (15) holds. Because $W^\ddagger_n$ is stochastically increasing in λ_1 under $\mathbb{P}_S$, we may assume that $\lambda_1 > 1$ is fixed. We use a different strategy which amounts to scanning the largest connected component of $\mathcal{G}_S$. Let $\mathcal{C}^S_{\max}$ be a largest connected component of $\mathcal{G}_S$.

For a small $c > 0$ to be chosen later, assume that $(1-c)n(1-\eta_{\lambda_1}) \le |\mathcal{C}^S_{\max}| \le (1+c)n(1-\eta_{\lambda_1})$ and $W_{\mathcal{C}^S_{\max}} \ge (1-c)\frac{n\lambda_1}{2}(1-\eta^2_{\lambda_1})$, which happens with high probability under $\mathbb{P}_S$ by Lemma 5. Note that, because $\lambda_1 > 1$, we have $\eta_{\lambda_1} < 1$ and, therefore, $|\mathcal{C}^S_{\max}| \asymp n$. Consequently, when computing $W^\ddagger_n$ we scan $\mathcal{C}^S_{\max}$, implying that

$$W^\ddagger_n \ge \frac{W_{\mathcal{C}^S_{\max}}}{|\mathcal{C}^S_{\max}|} \ge \frac{(1-c)(\lambda_1/2)(1-\eta^2_{\lambda_1})n}{(1+c)(1-\eta_{\lambda_1})n} \ge \frac{1-c}{1+c}\frac{\lambda_1}{2}(1+\eta_{\lambda_1}).$$

Since c above may be taken as small as we wish, and in view of (20), it suffices to show that $\lambda_1(1+\eta_{\lambda_1}) > 2$. Since η_λ converges to one when λ goes to one, we have $\lim_{\lambda\to 1}\lambda(1+\eta_\lambda) = 2$. Consequently, it suffices to show that the function $f : \lambda \mapsto \lambda(1+\eta_\lambda)$ is increasing on $(1,\infty)$. By definition of η_λ, we have $\eta_\lambda < 1/\lambda$ (since $e^{-\lambda} < 1/\lambda$) and $\eta'(\lambda) = \eta_\lambda(\eta_\lambda - 1)/(1-\lambda\eta_\lambda)$. Consequently, $f'(\lambda) = 2 + \frac{\eta_\lambda - 1}{1-\lambda\eta_\lambda}$. Hence, $f'(\lambda)$ is positive if $\eta_\lambda < (2\lambda-1)^{-1} := a_\lambda$. Recall that η_λ is the smallest solution of the equation $x = \exp[\lambda(x-1)]$, the largest solution being $x = 1$. Furthermore, we have $x \ge \exp[\lambda(x-1)]$ for any $x \in [\eta_\lambda, 1]$. To conclude, it suffices to prove $a_\lambda > e^{\lambda(a_\lambda - 1)}$. This last bound is equivalent to

$$\lambda - \frac{1}{2} - \frac{1}{2(2\lambda-1)} - \log(2\lambda-1) > 0.$$

The function on the LHS is null for $\lambda = 1$. Furthermore, its derivative $\frac{4(\lambda-1)^2}{(2\lambda-1)^2}$ is positive for $\lambda > 1$, which allows us to conclude. □

PROOF OF LEMMA 7. The proof is based on moment bounds for functions of independent random variables due to Boucheron et al. (2005) that generalize the Efron–Stein inequality.

Recall that $\mathcal{G}_S = (S, \mathcal{E}_S)$ is the subgraph induced by S. Fix some integer $k \in [0, n]$. For any $(i, j) \in \mathcal{E}_S$, define the graph $\mathcal{G}_S^{(i,j)}$ by removing (i, j) from the edge set of $\mathcal{G}_S$. Let $W_T^{(i,j)}$ be defined as W_T but computed on $\mathcal{G}_S^{(i,j)}$, and then let $W_{k,S}^{*(i,j)} = \max_{T \subset S, |T|=k} W_T^{(i,j)}$. Observe that $0 \le W_{k,S}^* - W_{k,S}^{*(i,j)} \le 1$ and that $W_{k,S}^{*(i,j)}$ is a measurable function of $\mathcal{E}_S^{(i,j)}$, the edges set of $\mathcal{G}_S^{(i,j)}$. Let $T^* \subset S$ be a subset of size k such that $W_{k,S}^* = W_{T^*}$. Then we have

$$\sum_{(i,j)\in\mathcal{E}_S} (W_{k,S}^* - W_{k,S}^{*(i,j)}) \le \sum_{(i,j)\in\mathcal{E}_S} (W_{T^*} - W_{T^*}^{(i,j)}) = W_{T^*} = W_{k,S}^*,$$

where the first equality comes from the fact that $W_{T^*} - W_{T^*}^{(i,j)} = \mathbb{1}_{\{(i,j)\in\mathcal{E}_{T^*}\}}$.

Applying [Boucheron et al. (2005), Corollary 1], we derive that, for any real $q \ge 2$,

$$[\mathbb{E}_S\{(W_{k,S}^* - \mathbb{E}_S[W_{k,S}^*])_+^q\}]^{1/q} \le \sqrt{2q\mathbb{E}_S[W_{k,S}^*]} + q;$$
$$[\mathbb{E}_S\{(W_{k,S}^* - \mathbb{E}_S[W_{k,S}^*])_-^q\}]^{1/q} \le \sqrt{2q\mathbb{E}_S[W_{k,S}^*]}.$$

Take some $t > 8(1 \vee \sqrt{\mathbb{E}_S[W_{k,S}^*]})$. For any $q \ge 2$, we have by Markov's inequality

$$\mathbb{P}_S[W_{k,S}^* \ge \mathbb{E}_S[W_{k,S}^*] + t] \le \left(\frac{\sqrt{2q\mathbb{E}_S[W_{k,S}^*]} + q}{t}\right)^q.$$

The choice $q = \frac{t}{4} \wedge \frac{t^2}{32\mathbb{E}_S[W_{k,S}^*]}$ is larger than 2 and leads to (21). Similarly, if we take some $t > 4\sqrt{\mathbb{E}_S[W_{k,S}^*]}$, and apply Markov's inequality, we get

$$\mathbb{P}_S[W_{k,S}^* \le \mathbb{E}_S[W_{k,S}^*] - t] \le \left(\frac{\sqrt{2q\mathbb{E}_S[W_{k,S}^*]}}{t}\right)^q.$$

The choice $q = \frac{t^2}{8\mathbb{E}_S[W_{k,S}^*]} \ge 2$ leads to (22). □

3.2. *The largest connected component.* This test rejects for large values of the size (number of nodes) of the largest connected component in $\mathcal{G}$, which we denoted $\mathcal{C}_{\max}$. We focus on the subcritical regime, $\limsup \lambda_0 < 1$, where the test is most relevant. We refer the reader to the extended online version [Verzelen and Arias-Castro (2013)] for a study of this test in the supercritical regime.

Define

$$I_\lambda = \lambda - 1 - \log(\lambda). \tag{23}$$

THEOREM 2 (Subcritical largest connected component test). *Assume that* $\log\log(N) = o(\log n)$, $\limsup \lambda_0 < 1$, *and* $I_{\lambda_0}^{-1}\log(N) \to \infty$. *The largest connected component test is asymptotically powerful when* $\liminf \lambda_1 > 1$ *or*

$$\lambda_0 \le \lambda_1 e^{1-\lambda_1} \text{ eventually} \quad \text{and} \quad \liminf \frac{I_{\lambda_0}}{\lambda_0 + I_{\lambda_1} - \lambda_0 e^{I_{\lambda_1}}} \frac{\log(n)}{\log(N)} > 1. \tag{24}$$

If we further assume that $n^2 = o(N)$, *then the largest connected component test is asymptotically powerless when* $\lambda_1 < 1$ *for all* n *and*

$$\lambda_0 \ge \lambda_1 e^{1-\lambda_1} \text{ eventually} \quad \text{or} \quad \limsup \frac{I_{\lambda_0}}{\lambda_0 + I_{\lambda_1} - \lambda_0 e^{I_{\lambda_1}}} \frac{\log(n)}{\log(N)} < 1. \tag{25}$$

If we assume that both λ_0 and λ_1 go to zero, then condition (24) is equivalent to

$$\liminf \frac{I_{\lambda_0}}{I_{\lambda_1}} \frac{\log(n)}{\log(N)} > 1, \tag{26}$$

which corresponds to the optimal detection boundary in this setting, as shown in Theorem 3.

The technical hypothesis $\log\log(N) = o(\log n)$ is only used for convenience when analyzing the case $\lambda_1 \to 1$. The condition $I_{\lambda_0}^{-1}\log(N) \to \infty$ implies that λ_0 can only converge to zero slower than any power of N. Although it is possible to analyze the test in the very sparse setting where λ_0 goes to zero polynomially fast, we did not do so to keep the exposition focused on the more interesting regimes.

PROOF OF THEOREM 2. That the test is powerful when $\liminf \lambda_1 > 1$ derives from the well-known phase transition phenomenon of Erdős–Rényi graphs. Let $\mathcal{C}_m$ denote a largest connected component of $\mathbb{G}(m, \lambda/m)$ and assume that $\lambda \in (0, \infty)$ is fixed. By Van der Hofstad (2012), Theorems 4.4, 4.5 and 4.8, in probability, we know that

$$|\mathcal{C}_m| \sim \begin{cases} I_\lambda^{-1}\log m, & \text{if } \lambda < 1, \\ (1-\eta_\lambda)m, & \text{if } \lambda > 1, \end{cases}$$

where η_λ is defined as in Lemma 5. When $\lambda > 1$, the result is actually contained in Lemma 5.

Hence, under the null with $\limsup \lambda_0 < 1$, the largest connected component of $\mathcal{G}$ is of order $\log(N)$ with probability going to one. Under the alternative H_S with $\liminf \lambda_1 > 1$, the graph $\mathcal{G}_S$ contains a giant connected component whose size of order n with probability going to one. Recalling that $\log(N) = o(n)$ allows us to conclude.

Now suppose that (24) holds. We assume that the sequence λ_1 is always smaller or equal to 1, that $I_{\lambda_1}^{-1} = O(\log(n)/\log(N))$ and that $\log(I_{\lambda_1}^{-1} \vee 1) = o(\log n)$, meaning that λ_1 does not converge too fast to 1. We may do so while keeping condition (24) true because the distribution of $|\mathcal{C}_{\max}|$ under $\mathbb{P}_S$ is stochastically

increasing with λ_1, because $\limsup \lambda_0 < 1$, $I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}} \sim I_{\lambda_1}(1-\lambda_0)$ for $\lambda_1 \to 1$, and because $\log\log(N) = o(\log n)$.

By hypothesis (24), there exists a constant $c' > 0$, such that

$$\tau := \liminf \frac{I_{\lambda_0} \log(n)}{(I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}}) \log(N)} \geq 1 + c'.$$

To upper-bound the size of $\mathcal{C}_{\max}$ under $\mathbb{P}_0$, we use the following.

LEMMA 8. *Let $\mathcal{C}_m$ denote a largest connected component of $\mathbb{G}(m, \lambda/m)$ and assume that $\lambda < 1$ for all m and $\log[I_\lambda^{-1} \vee 1] = o(\log(m))$. Then, for any sequence u_m satisfying*

$$\liminf \frac{u_m I_\lambda}{\log m} > 1,$$

we have

$$\mathbb{P}\big(|\mathcal{C}_m| \geq u_m\big) = o(1).$$

PROOF. This lemma is a slightly modified version of Van der Hofstad (2012), Theorem 4.4, the main difference being that λ was fixed in the original statement. Details are omitted. □

Define $c = (c' \wedge 1)/4$. Applying Lemma 8, $|\mathcal{C}_{\max}| \leq t_0 := I_{\lambda_0}^{-1} \log(N)(1+c)$, with probability going to one under $\mathbb{P}_0$.

We now need to lower-bound the size of $\mathcal{C}_{\max}$ under $\mathbb{P}_S$. Define

$$k_0 = (1-c)\log(n)\big[I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}}\big]^{-1}, \qquad k = \lceil k_0 \rceil,$$

$$q_0 = (1-c)\log(n)\frac{1 - \lambda_0 e^{I_{\lambda_1}}}{I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}}}, \qquad q = \lfloor q_0 \rfloor.$$

The denominator of k_0 is positive since $\lambda_0 e^{I_{\lambda_1}} \leq 1$ and

$$I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}} \geq I_{\lambda_1} + e^{-I_{\lambda_1}}\big(1 - e^{I_{\lambda_1}}\big) = I_{e^{-I_{\lambda_1}}} > 0. \tag{27}$$

We note that $k = O(\log n)$, unless the denominator of k_0 goes to zero, which is only possible when I_{λ_1} goes to zero (implying $\lambda_1 \to 1$), in which case

$$k \sim \log(n)\big[I_{\lambda_1}(1-\lambda_0)\big]^{-1} = O\big[I_{\lambda_1}^{-1} \vee 1\big]\log(n) = O\big[\log(N)\big], \tag{28}$$

since, in this case, (24) implies that $I_{\lambda_1}^{-1} = O(\log(n)/\log(N))$, and we also have $\limsup \lambda_0 < 1$ by assumption. So (28) holds in any case.

We shall prove that among the connected components of $\mathcal{G}_S$ of size larger than q, there exists at least one component whose size in $\mathcal{G}$ is larger than k. By definition of c, we have $\liminf k/t_0 \geq \tau(1-c)/(1+c) \geq (1+c')(1-c)/(1+c) > 1$,

and the connected component test is therefore powerful. The main arguments rely on the second moment method and on the comparison between cluster sizes and branching processes. Before that, recall that $t_0 \to \infty$, so that $\log(n) I^{-1}_{e^{-I_{\lambda_1}}} \succ k_0 \to \infty$, which in turn implies $I_{\lambda_1} = o(\log(n))$.

LEMMA 9. *Fix any $c > 0$. Consider the distribution $\mathbb{G}(m, \lambda/m)$ and assume that λ satisfies*

$$\limsup \lambda \le 1, \qquad \log\bigl[I_\lambda^{-1} \vee 1\bigr] = o\bigl(\log(m)\bigr), \qquad I_\lambda^{-1} \log m \to \infty.$$

For any sequence $q = a \log(m)$ with $a \le I_\lambda^{-1}(1-c)$, let $Z_{\ge q}$ denote the number of nodes belonging to a connected component whose size is larger than q. With probability going to one, we have

$$Z_{\ge q} \ge m^{1 - a I_\lambda - o(1)}. \tag{29}$$

PROOF. This lemma is a simple extension of the second moment method argument [equations (4.3.34) and (4.3.35)] in the proof of Van der Hofstad (2012), Theorem 4.5, where λ is fixed, while here it may vary with m, and in particular, may converge to 1. We leave the details to the reader. □

Observe that

$$\frac{q}{(1-c) I_{\lambda_1}^{-1} \log(n)} \le \frac{I_{\lambda_1} - \lambda_0 I_{\lambda_1} e^{I_{\lambda_1}}}{I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}}} \le 1 - \lambda_0 \frac{1 - e^{I_{\lambda_1}} + I_{\lambda_1} e^{I_{\lambda_1}}}{I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}}} \le 1,$$

using the fact that $x e^x - e^x + 1 \ge 0$ for any $x \ge 0$. Thus, we can apply Lemma 9 to $\mathcal{G}_S$. And by Lemma 8, the largest connected component of $\mathcal{G}_S$ has size smaller than $2 I_{\lambda_1}^{-1} \log(n)$ with probability tending to one. Hence, $\mathcal{G}_S$ contains more than

$$\frac{n^{1+o(1)} e^{-q I_{\lambda_1}}}{2 I_{\lambda_1}^{-1} \log n} = n e^{-q I_{\lambda_1} - o(\log(n))}$$

connected components of size larger than q. [We used the fact that $\log(I_{\lambda_1}^{-1} \vee 1) = o(\log n)$.] If $k_0 - q_0 \le 1$, then applying Lemma 9 to $q+2$ (instead of q) allows us to conclude that there exists a connected component of size at least k. This is why we assume in the following that $\liminf k_0 - q_0 > 1$. By definition of k_0 and q_0, $k_0 - q_0 \ge 1$, implies that

$$\log(n)\lambda_0 \ge \frac{1}{1-c} e^{-I_{\lambda_1}} \bigl(I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}}\bigr) \ge \frac{1}{1-c} e^{-I_{\lambda_1}} I_{e^{-I_{\lambda_1}}}$$

by (27). Thus, $\liminf k_0 - q_0 > 1$ implies that for n large enough $\log(n)\lambda_0 \ge \lambda_1 I_{\lambda_1} e$, and consequently

$$I_{\lambda_0} \le O(1) - \log(\lambda_0) \le o\bigl(\log(n)\bigr) + I_{\lambda_1} + \log\bigl[I^{-1}_{e^{-I_{\lambda_1}}}\bigr] = o\bigl(\log(n)\bigr) \tag{30}$$

since $I_{\lambda_1} = o(\log(n))$, $-\log(I_{\lambda_1}) \le o(\log(n))$ and $I^{-1}_{e^{-I_\lambda}} = O[(e^{-I_\lambda} - 1)^{-2}] = O[I_\lambda^{-2}]$.

Let $\{\mathcal{C}_S^{(i)}, i \in \mathcal{I}\}$ denote the collection of connected components of size larger than q in $\mathcal{G}_S$. For any such component $\mathcal{C}_S^{(i)}$, we extract any subconnected component $\tilde{\mathcal{C}}_S^{(i)}$ of size q. Recall that, with probability going to one,

$$|\mathcal{I}| \ge n^{1-o(1)} e^{-q I_{\lambda_1}}. \tag{31}$$

For any node x, let $\mathcal{C}(x)$ denote the connected component of x in $\mathcal{G}$, and let $\mathcal{C}_{-S}(x)$ denote the connected component of x in the graph $\mathcal{G}_{-S}$ where all the edges in $\mathcal{G}_S$ have been removed. Then let

$$U_i := \bigcup_{x \in \tilde{\mathcal{C}}_S^{(i)}} \mathcal{C}_{-S}(x), \qquad i \in \mathcal{I}; \qquad V = \sum_{i \in \mathcal{I}} \mathbb{1}_{\{|U_i| \ge k\}}.$$

Since $V \ge 1$ implies that the largest connected component of $\mathcal{G}$ is larger than k, it suffices to prove that V is larger than one with probability going to one. Observe that conditionally to $|\mathcal{I}|$, the distribution of $(|U_i|,\ i \in \mathcal{I})$ is independent of $\mathcal{G}_S$. Again, we use a second moment method based on a stochastic comparison between connected components and binomial branching processes.

LEMMA 10. *The following bounds hold*:

$$\mathbb{P}_S\big[|U_i| \ge k\big] \ge \left(\frac{k}{k-q}\right)^{k-q} e^{-\lambda_0 q - I_{\lambda_0}(k-q)} n^{-o(1)},$$

$$\operatorname{Var}_S[V|\mathcal{G}_S] \le |\mathcal{I}|\mathbb{P}_S\big[|U_i| \ge k\big] + \frac{|\mathcal{I}|^2 q^2}{N} \mathbb{E}_S\big[|U_i| \mathbb{1}_{\{U_i \ge k\}}\big], \tag{32}$$

$$\mathbb{P}_S\big[|U_i| \ge k\big] \le \mathbb{E}_S\big[|U_i| \mathbb{1}_{\{U_i \ge k\}}\big] \le \left(\frac{k}{k-q}\right)^{k-q} e^{-\lambda_0 q - I_{\lambda_0}(k-q)} n^{o(1)}. \tag{33}$$

Before proceeding to the proof of Lemma 10, we finish proving that $V \ge 1$ with probability going to one. Let us define $\mu_k := (\frac{k}{k-q})^{k-q} e^{-\lambda_0 q - I_{\lambda_0}(k-q)}$. Applying Chebyshev's inequality, we derive from Lemma 10

$$V \ge |\mathcal{I}|\mu_k n^{-o(1)} - O_{\mathbb{P}_S}\big[(|\mathcal{I}|\mu_k)^{1/2} n^{o(1)}\big] - O_{\mathbb{P}_S}\big[|\mathcal{I}|(\mu_k/N)^{1/2} n^{o(1)}\big].$$

In order to conclude, we only to need to prove that $|\mathcal{I}|\mu_k \ge n^{c-o(1)}$ since $(|\mathcal{I}|\mu_k)^{1/2}/|\mathcal{I}|(\mu_k/N)^{1/2} = \sqrt{N/|\mathcal{I}|} \ge 1$. Relying on (31), we derive

$$\begin{aligned} |\mathcal{I}|\mu_k &\ge n^{1-o(1)} \left(\frac{k}{k-q}\right)^{k-q} e^{-\lambda_0 q - q I_{\lambda_1} - I_{\lambda_0}(k-q)} \\ &\ge n^{1-o(1)} \left(\frac{k_0}{k_0-q_0}\right)^{k_0-q_0} e^{-\lambda_0 q_0 - q_0 I_{\lambda_1} - I_{\lambda_0}(k_0-q_0) - 2I_{\lambda_0}} \end{aligned}$$

$$\begin{aligned}
&\geq n^{1-o(1)}\lambda_0^{-(k_0-q_0)}e^{-\lambda_0 q_0-k_0 I_{\lambda_1}-I_{\lambda_0}(k_0-q_0)}\\
&\geq n^{1-o(1)}e^{-k_0\lambda_0-k_0 I_{\lambda_1}}e^{k_0-q_0}\\
&\geq n^{1-o(1)}\exp\bigl[-k_0\bigl(\lambda_0+I_{\lambda_1}-\lambda_0 e^{I_{\lambda_1}}\bigr)\bigr]=n^{c-o(1)},
\end{aligned}$$

where we use (30) and $\frac{k_0}{k_0-q_0}=\lambda_0^{-1}e^{-I_{\lambda_1}}$ in the third line, the definition $I_{\lambda_0}=\lambda_0-\log(\lambda_0)-1$ in the fourth line and the definitions of k_0 and q_0 in the last line.

PROOF OF LEMMA 10. We shall need the two following lemmas.

LEMMA 11 (Upper bound on the cluster sizes). *Consider the distribution $\mathbb{G}(m,\lambda/m)$ and a collection $\mathcal{J}$ of nodes. For each $k\geq|\mathcal{J}|$,*

$$\mathbb{P}\Bigl[\Bigl|\bigcup_{x\in\mathcal{J}}\mathcal{C}(x)\Bigr|\geq k\Bigr]\leq\mathbb{P}_{m,\lambda/m}[T_1+\cdots+T_{|\mathcal{J}|}\geq k],$$

where $T_1,T_2,\ldots$ denote the total progenies of i.i.d. binomial branching processes with parameters m and λ/m. For each $|\mathcal{J}|\leq k\leq m$,

$$\mathbb{P}\Bigl[\Bigl|\bigcup_{x\in\mathcal{J}}C(x)\Bigr|\geq k\Bigr]\geq\mathbb{P}_{m-k,\lambda/m}[T_1+\cdots+T_{|\mathcal{J}|}\geq k],$$

where $T_1,T_2,\ldots$ denote the total progenies of i.i.d. binomial branching processes with parameters $m-k$ and λ/m.

Lemma 11 is a slightly modified version of Van der Hofstad (2012), Theorems 4.2 and 4.3, the only difference being that $|\mathcal{J}|=1$ in the original statement. The proof is left to the reader. The following result is proved in Van der Hofstad (2012), Section 3.5.

LEMMA 12 (Law of the total progeny). *Let $T_1,\ldots,T_r$ denote the total progenies of r i.i.d. branching processes with offspring distribution X. Then*

$$\mathbb{P}[T_1+\cdots+T_r=k]=\frac{r}{k}\mathbb{P}[X_1+\cdots+X_k=k-r],$$

where (X_i), $i=1,\ldots,k$ are i.i.d. copies of X.

Consider any subset $\mathcal{J}$ of node of size q. Under the null hypothesis, $|U_i|=|\bigcup_{x\in\tilde{\mathcal{C}}_S^{(i)}}\mathcal{C}_{-S}(x)|$ is stochastically dominated by $Z:=|\bigcup_{x\in\mathcal{J}}\mathcal{C}(x)|$. Let T_q be sum of the total progenies of q independent binomial branching processes with parameters $N-n+q-k$ and p_0. By Lemma 11, we derive

$$\mathbb{P}_S\bigl[|U_i|\geq k\bigr]\geq\mathbb{P}_0[Z\geq k]\geq\mathbb{P}_{N-n+q-k,p_0}[T_q\geq k]\geq\mathbb{P}_{N-n+q-k,p_0}[T_q=k].$$

Let $X_1, X_2, \dots$ denote independent binomial random variables with parameters $N-n+q-k$ and p_0. Relying on Lemma 12 and the lower bound $\binom{s}{r} \geq \frac{(s-r)^r}{r!} \geq (re)^{-1}(\frac{(s-r)e}{r})^r$, we derive

$$\begin{aligned}
&\mathbb{P}_{N-n+q-k,p_0}[T_q = k] \\
&\quad = \frac{q}{k}\mathbb{P}_{N-n+q-k,p_0}[X_1 + \cdots + X_k = k-q] \\
&\quad = \frac{q}{k}\binom{k(N-n+q-k)}{k-q} p_0^{k-q}(1-p_0)^{k(N-n+q-k)-k+q} \\
&\quad \succ \frac{q}{k^2}\left[\frac{ek(N-n-2(k-q))}{k-q}\right]^{k-q}\left(\frac{\lambda_0}{N}\right)^{k-q} e^{-\lambda_0 k - kO(n/N)} \\
&\quad \succ \frac{q}{k^2} e^{-I_{\lambda_0}(k-q)} e^{-\lambda_0 q}\left(\frac{k}{k-q}\right)^{k-q} e^{-kO(n/N)} \\
&\quad \succ \left(\frac{k}{k-q}\right)^{k-q} e^{-\lambda_0 q - I_{\lambda_0}(k-q)} n^{o(1)},
\end{aligned}$$

where we used (28) with $n\log(N)/N = o(\log(n))$ in the last line.

Let us now prove (33). The first inequality is Markov's. For the second, by Lemma 11, U_i is stochastically dominated by $\tilde{T}_q$, the sum of the total progenies of q independent binomial branching processes with parameters N and p_0, so that

$$\mathbb{E}_S\big[|U_i|\mathbb{1}_{\{U_i \geq k\}}\big] = \sum_{r=k}^{N} \mathbb{P}_S[U_i \geq r] \leq \sum_{r=k}^{\infty} \mathbb{P}_{N,p_0}[\tilde{T}_q \geq r] \leq \sum_{r=k}^{\infty} r\mathbb{P}_{N,p_0}[\tilde{T}_q = r].$$

We use Lemma 12 to control the deviation of $\tilde{T}_q$. Below $X_1, X_2, \dots$ denote independent binomial random variables with parameter N and p_0:

$$\begin{aligned}
\sum_{r=k}^{\infty} r\mathbb{P}_{N,p_0}[\tilde{T}_q = r] &\leq \sum_{r=k}^{\infty} r\frac{q}{r}\mathbb{P}_{N,p_0}[X_1 + \cdots + X_r = r-q] \\
&\leq \sum_{r=k}^{\infty} q\exp\left[-Nr\,H_{p_0}\left(\frac{r-q}{Nr}\right)\right],
\end{aligned} \tag{34}$$

by Chernoff's inequality since

$$\frac{r-q}{Nr} \geq \frac{k-q}{Nk} \geq \frac{k_0-q_0}{Nk_0} = \frac{\lambda_0 e^{I_{\lambda_1}}}{N} > \frac{\lambda_0}{N} = p_0.$$

By Lemma 2, $H_{p_0}(a) \geq a\log(a/p_0) - a + p_0$. Thus, we arrive at

$$\begin{aligned}
\mathbb{E}_S\big[|U_i|\mathbb{1}_{\{U_i \geq k\}}\big] &\leq \sum_{r=k}^{\infty} q\exp\left[-(r-q)\log\left(\frac{r-q}{r\lambda_0}\right) + r - q - r\lambda_0\right] \\
&\leq q\sum_{r=k}^{\infty}\exp[A_r],
\end{aligned} \tag{35}$$

where $A_r := -(r-q)I_{\lambda_0} - q\lambda_0 - (r-q)\log(\frac{r-q}{r})$. Differentiating the function A_r with respect to r, we get

$$\begin{aligned}\frac{dA_r}{dr} &= -I_{\lambda_0} - \log\left(\frac{r-q}{r}\right) - 1 + \frac{r-q}{r} \\ &\le -I_{\lambda_0} - \log\left(\frac{k-q}{k}\right) - 1 + \frac{k-q}{k} \\ &\le -I_{\lambda_0} - \log\left(\frac{k_0-q_0}{k_0}\right) - 1 + \frac{k_0-q_0}{k_0} \\ &= -\lambda_0 - I_{\lambda_1} + \lambda_0 e^{I_{\lambda_1}},\end{aligned}$$

which is negative as argued below the definition of k. Consequently, A_r is a decreasing function of r. Define r_1 as the smallest integer such that $\log((r-q)/r) \ge -I_{\lambda_0}/2$. Since $\limsup \lambda_0 < 1$, it follows that $r_1 = O(q)$. Coming back to (35), we derive

$$\begin{aligned}\mathbb{E}_S\big[|U_i|\mathbb{1}_{\{U_i\ge k\}}\big] &\le q(r_1-k)_+ \exp[A_k] + q\sum_{r=r_1}^{\infty}\exp[A_r] \\ &\le qe^{A_k}\left[(r_1-k)_+ + \sum_{r=r_1}^{\infty} e^{-(r-q)[I_{\lambda_0}-\log((r-q)/r)]}\right] \\ &\le qe^{A_k}\left[(r_1-k)_+ + \sum_{r=r_1}^{\infty} e^{-(r-q)I_{\lambda_0}/2}\right] \le e^{A_k}O\big(k^2\big),\end{aligned} \tag{36}$$

since $\limsup \lambda_0 < 1$. From (28), we know that $k = O(\log(N)) = n^{o(1)}$, which allows us to prove (33).

Turning to the proof of (32), we have the decomposition

$$\begin{aligned}&\mathrm{Var}_S[V|\mathcal{G}_S] \\ &\quad\le |\mathcal{I}|\mathbb{P}_S[U_i \ge k] + \sum_{i\ne i'\in\mathcal{I}}\big\{\mathbb{P}_S\big[|U_i|\ge k, |U_{i'}|\ge k\big] - \mathbb{P}_S^2\big[|U_i|\ge k\big]\big\} \\ &\quad\le |\mathcal{I}|\mathbb{P}_S[U_i \ge k] + |\mathcal{I}|^2\mathbb{P}_S\big[|U_i|\ge k, U_i\cap U_{i'}\ne\varnothing\big] \\ &\qquad + |\mathcal{I}|^2\big\{\mathbb{P}_S\big[|U_i|\ge k, |U_{i'}|\ge k, U_i\cap U_{i'}=\varnothing\big] - \mathbb{P}_S^2\big[|U_i|\ge k\big]\big\}.\end{aligned} \tag{37}$$

The last term is nonpositive. Indeed,

$$\begin{aligned}&\mathbb{P}_S\big[|U_i|\ge k, |U_{i'}|\ge k, U_i\cap U_{i'}=\varnothing\big] - \mathbb{P}_S^2\big[|U_i|\ge k\big] \\ &\quad= \sum_{r=k}^{N}\mathbb{P}_S\big[|U_i|=r\big]\big(\mathbb{P}_S\big[|U_{i'}|\ge k, U_i\cap U_{i'}=\varnothing\big||U_i|=r\big] - \mathbb{P}_S\big[|U_{i'}|\ge k\big]\big) \\ &\quad\le \sum_{r=k}^{N}\mathbb{P}_S\big[|U_i|=r\big]\big(\mathbb{P}_S\big[|U_{i'}|\ge k|U_i\cap U_{i'}=\varnothing, |U_i|=r\big] - \mathbb{P}_S\big[|U_{i'}|\ge k\big]\big),\end{aligned}$$

where the last difference is negative, as the graph is now smaller once we condition on $|U_i| \geq 1$ and $U_i \cap U_{i'} = \varnothing$. Consider the second term in (37):

$$\mathbb{P}_S\big[|U_i| \geq k, U_i \cap U_{i'} \neq \varnothing\big] = \sum_{r=k}^{N} \mathbb{P}_S\big[|U_i| = r\big]\mathbb{P}_S\big[U_i \cap U_{i'} \neq \varnothing \big| |U_i| = r\big].$$

By symmetry and a union bound, we derive

$$\mathbb{P}_S\big[U_i \cap U_{i'} \neq \varnothing \big| |U_i| = r\big] \leq q^2 \mathbb{P}_S\big[y \in \mathcal{C}_{-S}(x) \big| |U_i| = r\big],$$

for some $x \in \tilde{\mathcal{C}}_S^{(i)}$ and $y \in \tilde{\mathcal{C}}_S^{(i')}$. Since the graph $\mathcal{G}_{-S}$ is not symmetric, the probability that a fixed node z belongs to $\mathcal{C}_{-S}(x)$ conditionally to $|\mathcal{C}_{-S}(x)|$ is smaller for $z \in S \setminus \{i\}$ than for $z \in S^c$. It follows that

$$\mathbb{P}_S\big[y \in \mathcal{C}_{-S}(x) \big| |U_i| = r\big] \leq \mathbb{E}_S\bigg[\frac{|\mathcal{C}_{-S}(x)| - 1}{N-1}\Big| |U_i| = r\bigg].$$

Since $|\mathcal{C}_{-S}(x)| \leq r$, we conclude

$$\mathbb{P}_S\big[|U_i| \geq k, U_i \cap U_{i'} \neq \varnothing\big] \leq \sum_{r=k}^{N} \mathbb{P}_S\big[|U_i| = r\big]\frac{q^2 r}{N} = \frac{q^2}{N}\mathbb{E}_S\big[|U_i| \mathbb{1}_{\{U_i \geq k\}}\big]. \qquad \square$$

Let us continue with the proof of Theorem 2, now assuming that $\lambda_1 < 1$, that condition (25) holds, and that $n^2 = o(N)$. We assume in the sequel that $I_{\lambda_1} \leq -\log(\lambda_0)$, meaning that λ_1 is not too small. We may do so while keeping condition (25) true, because the distribution of $|\mathcal{C}_{\max}|$ under $\mathbb{P}_S$ is increasing with respect to λ_1 and because for $I_{\lambda_1} = -\log(\lambda_0)$, (25) is equivalent to $\limsup \log(n)/\log(N) < 1$, which is always true since $n^2 = o(N)$. Similarly, we assume that $I_{\lambda_1} = o(\log(n))$ while keeping condition (25) true since for I_{λ_1} going to infinity, (25) is equivalent to $\limsup \frac{I_{\lambda_0}\log(n)}{I_{\lambda_1}\log(N)} < 1$ and since $I_{\lambda_0}^{-1}\log(N) \to \infty$. By condition (25), there exists a constant $c > 0$ such that

$$\limsup \frac{I_{\lambda_0}}{\lambda_0 + I_{\lambda_1} - \lambda_0 e^{I_{\lambda_1}}} \frac{\log(n)}{\log(N)} < 1 - c. \tag{38}$$

We shall prove that with probability $\mathbb{P}_S$ going to one, the largest connected component of $\mathcal{G}$ does not intersect S. As the distribution of the statistic under the alternative dominates the distribution under the null, this will imply that the largest connected component test is asymptotically powerless.

Denote $\mathcal{A}$ the event that, for all $(x, y) \in S$, there is no path between x and y with all other nodes in S^c. For any subset T, denote $\mathcal{C}_T(x)$ the connected component of x in $\mathcal{G}_T$, and recall that $\mathcal{C}(x)$ is a shorthand for $\mathcal{C}_{\mathcal{V}}(x)$. By symmetry, we have

$$\mathbb{P}_S\big[\mathcal{A}^c\big] \leq n^2 \mathbb{P}_0\big[y \in \mathcal{C}_{-S}(x)\big] \leq n^2 \mathbb{P}_0\big[y \in \mathcal{C}(x)\big],$$

since the probability of the edges outside $\mathcal{G}_S$ under $\mathbb{P}_S$ is the same as under $\mathbb{P}_0$. Again, by symmetry

$$\mathbb{P}_0\big[y \in \mathcal{C}(x)\big] = \mathbb{E}_0\big[\mathbb{P}_0\big[y \in C(x)\big]|\mathcal{C}(x)|\big] \le \mathbb{E}_0\left[\frac{|\mathcal{C}(x)|}{N-1}\right] \le \frac{1}{(N-1)(1-\lambda_0)},$$

as the expected size of a cluster is dominated by the expected progeny of a branching process with parameters N and p_0 (Lemma 11) and the expected progeny of a subcritical branching process having mean offspring $\mu < 1$ is $(1-\mu)^{-1}$ [Van der Hofstad (2012), Theorem 3.5]. Thus,

$$\mathbb{P}_S\big[\mathcal{A}^c\big] = O\big(n^2/N\big) = o(1). \tag{39}$$

Define

$$k := (1-c)^{1/2}\log(N) I_{\lambda_0}^{-1}. \tag{40}$$

Since $\limsup \lambda_0 < 1$ and since $\log\log(N) = o[\log(n)]$, it follows that $k \asymp \log(N) = n^{o(1)}$. By Lemma 9, $|\mathcal{C}_{\max}|$ is larger or equal to k with probability $\mathbb{P}_S$ (and $\mathbb{P}_0$) going to one. Thus, it suffices to prove that $\mathbb{P}_S[\bigvee_{x\in S}|\mathcal{C}(x)| \ge k] \to 0$. Observe that

$$\mathbb{P}_S\left[\bigvee_{x\in S}|\mathcal{C}(x)| \ge k\right] \le n\mathbb{P}_S\big[\{|C(x)| \ge k\} \cap \mathcal{A}\big] + \mathbb{P}_S\big[\mathcal{A}^c\big],$$

so that, by (39), we only need to prove that $n\mathbb{P}_S[\{|C(x)| \ge k\} \cap \mathcal{A}] = o(1)$. Under the event $\mathcal{A}$, $\mathcal{C}(x) \cap S$ is exactly the connected component $\mathcal{C}_S(x)$ of x in $\mathcal{G}_S$. Furthermore, $\mathcal{C}(x)$ is the union of $\mathcal{C}_{-S}(y)$ over $y \in \mathcal{C}_S(x)$. Consequently, we have the decomposition

$$\begin{aligned}\mathbb{P}_S\big[\{|\mathcal{C}(x)| \ge k\} \cap \mathcal{A}\big] &\le \mathbb{P}_S\big[|\mathcal{C}_S(x)| \ge k\big] \\ &\quad + \sum_{q=1}^{k-1} \mathbb{P}_S\big[|\mathcal{C}_S(x)| = q\big]\mathbb{P}_S\big[\mathcal{B}_q||\mathcal{C}_S(x)| = q\big],\end{aligned}$$

where $\mathcal{B}_q := \{|\bigcup_{y\in\mathcal{C}_S(x)}\mathcal{C}_{-S}(y)| \ge k\}$. By Lemma 11, the distribution of $|\mathcal{C}_S(x)|$ is stochastically dominated by the total progeny distribution of a binomial branching process with parameters $(n, \lambda_1/n)$. Denote by $\mathcal{J}$ any set of nodes of size q. Since, conditionally to $|\mathcal{C}_S(x)| = q$, the event $\mathcal{B}_q$ is increasing and only depends on the edges outside $\mathcal{G}_S$, we have

$$\mathbb{P}_S\big[\mathcal{B}_q||\mathcal{C}_S(x)| = q\big] \le \mathbb{P}_0\left[\left|\bigcup_{y\in\mathcal{J}}\mathcal{C}(y)\right| \ge k\right],$$

which is in turn, by Lemma 11, smaller than the probability that the total progeny of q independent branching processes with parameters $(N, \lambda_0/N)$ is larger than k.

Relying on the law of the total progeny of branching processes (Lemma 12) and Lemma 11, we get

$$\mathbb{P}_S\big[|\mathcal{C}_S(x)| = q\big] \le \frac{1}{q}\mathbb{P}\big[\mathrm{Bin}(nq, \lambda_1/n) = q-1\big],$$

$$\mathbb{P}_S\big[\mathcal{B}_q \big| |\mathcal{C}_S(x)| = q\big] \le \sum_{r=k}^{\infty} \frac{q}{r}\mathbb{P}\big[\mathrm{Bin}(Nr, \lambda_0/N) = r-q\big].$$

Working out the density of the binomial random variable, we derive

$$\mathbb{P}_S\big[|\mathcal{C}_S(x)| = q\big] \le \binom{nq}{q-1} p_1^{q-1}(1-p_1)^{nq-q+1} \prec \frac{1}{\lambda_1} e^{-I_{\lambda_1} q},$$

and for $q \le (1-\lambda_0)k$, we get

$$\mathbb{P}_S\big[\mathcal{B}_q \big| |\mathcal{C}_S(x)| = q\big] \le \frac{q}{k}\exp\left[-NkH_{p_0}\left(\frac{k-q}{Nk}\right)\right],$$

which is exactly the term (34), which has been proved in (36) to be smaller than

$$O\big(k^2\big)\left(\frac{k}{k-q}\right)^{k-q} e^{-(k-q)I_{\lambda_0} - q\lambda_0}.$$

Let define

$$B_\ell := e^{-I_{\lambda_1}\ell - \ell\lambda_0 - (k-\ell)I_{\lambda_0}}\left(\frac{k}{k-\ell}\right)^{k-\ell}.$$

Gathering all these bounds, we get

$$\begin{aligned}\mathbb{P}_S\big[\{|C(x)| \ge k\} \cap \mathcal{A}\big] &\prec \frac{e^{-I_{\lambda_1}k}}{\lambda_1} + \sum_{q=\lceil(1-\lambda_0)k\rceil}^{k-1} \frac{e^{-I_{\lambda_1}q}}{\lambda_1} + O\left(\frac{k^2}{\lambda_1}\right)\sum_{q=1}^{\lfloor(1-\lambda_0)k\rfloor} B_q \\ &\prec \frac{k^3}{\lambda_1}\left[e^{-I_{\lambda_1}(1-\lambda_0)k} + \bigvee_{q=1}^{k} B_q\right] \prec n^{o(1)} \sup_{q\in[0;k]} B_q,\end{aligned}$$

where we observe that $e^{-I_{\lambda_1}(1-\lambda_0)k} = B_{(1-\lambda_0)k}$ and we use $k = n^{o(1)}$ and $I_{\lambda_1} = o(\log(n))$. By differentiating $\log(B_q)$ as a function of q, we obtain the maximum

$$\sup_{q\in[0;k]} B_q \le \begin{cases} e^{-kI_{\lambda_0}}, & \text{if } \lambda_0 e^{I_{\lambda_1}} > 1, \\ e^{-I_{\lambda_1}k}\exp\big[\lambda_0 k\big(e^{I_{\lambda_1}} - 1\big)\big], & \text{else.}\end{cases}$$

Recall that we assume $\lambda_0 e^{I_{\lambda_1}} \le 1$ so that

$$\begin{aligned}\mathbb{P}_S\big[\{|C(x)| \ge k\} \cap \mathcal{A}\big] &\prec n^{o(1)}\exp\big[-k\{\lambda_0 + I_{\lambda_1} - \lambda_0 e^{I_{\lambda_1}}\}\big] \\ &\prec n^{-(1-c)^{-1/2}+o(1)},\end{aligned}$$

by definition (40) of k and condition (38). We conclude that

$$n\mathbb{P}_S\big[\{|C(x)| \ge k\} \cap \mathcal{A}\big] = o(1). \qquad \square$$

3.3. *Other tests.* In the extended version [Verzelen and Arias-Castro (2013)], we consider other tests.

The number of k-trees. Consider the test that rejects for large values of the number of subtrees of size k in $\mathcal{G}$. This test happens to partially bridge the gap in constants between what the broad scan test and largest connected component test can achieve in the regime where λ_0 is constant. In more detail, we find in Verzelen and Arias-Castro (2013) that even in the supercritical Poisson regime with $1 < \lambda_0 < e$, there exist subcritical communities $\lambda_1 < 1$ that are asymptotically detectable with probability going to one. The condition $\lambda_1 > \sqrt{\lambda_0/e}$ will be shown to be minimal in Theorem 4.

The number of triangles. Consider the test that rejects for large values of the number of triangles in $\mathcal{G}$. This is an emblematic test among those based on counting patterns, as it is the simplest and the least costly to compute. As such, the number of triangles in a graph is an important topological characteristic, with applications in the study of real-life networks. For example, Maslov, Sneppen and Zaliznyak (2004) use the number of triangles to quantify the amount of clustering in the Internet. It is easy to see—and formally established in Verzelen and Arias-Castro (2013)—that this test as nontrivial power in the Poisson regime where both λ_0 and λ_1 are fixed.

4. Information theoretic lower bounds. In this section, we state and prove lower bounds on the risk of *any* test whatsoever. In most cases, we find sufficient conditions under which the null and alternative hypotheses merge asymptotically, meaning that all tests are asymptotically powerless. In other cases, we find sufficient conditions under which no test is asymptotically powerful.

To derive lower bounds, it is standard to reduce a composite hypothesis to a simple hypothesis. This is done by putting a prior on the set of distributions that define the hypothesis. In our setting, we assume that p_0 is known so that the null hypothesis is simple, corresponding to the Erdős–Rényi model $\mathbb{G}(N, p_0)$. The alternative $H_1 := \bigcup_{|S|=n} H_S$ is composite and "parameterized" by subsets of nodes of size n. We choose as prior the uniform distribution over these subsets, leading to the simple hypothesis $\bar{H}_1$ comprising of $\mathbb{G}(N, p_0; n, p_1)$ defined earlier. The corresponding risk for H_0 versus $\bar{H}_1$ is

$$\bar{\gamma}_N(\phi) = \mathbb{P}_0(\phi = 1) + \frac{1}{\binom{N}{n}} \sum_{|S|=n} \mathbb{P}_S(\phi = 0).$$

Note that $\gamma_N(\phi) \geq \bar{\gamma}_N(\phi)$ for any test ϕ. Our choice of prior was guided by invariance considerations: the problem is invariant with respect to a relabeling of the nodes. In our setting, this implies that $\gamma_N^* = \bar{\gamma}_N^*$, or equivalently, that there exists a test invariant with respect to permutation of the nodes that minimizes the worst-case risk [Lehmann and Romano (2005), Lemma 8.4.1]. Once we have a simple

versus simple hypothesis testing problem, we can express the risk in closed form using the corresponding likelihood ratio. Let $\bar{\mathbb{P}}_1$ denote the distribution of $\mathbf{W}$ under $\bar{H}_1$, meaning $\mathbb{G}(N, p_0; n, p_1)$. The likelihood ratio for testing $\mathbb{P}_0$ versus $\bar{\mathbb{P}}_1$ is

$$L = \frac{1}{\binom{N}{n}} \sum_{|S|=n} L_S, \tag{41}$$

where L_S is the likelihood for testing $\mathbb{P}_0$ versus $\mathbb{P}_S$. Then the test $\phi^* = \{L > 1\}$ is the unique test that minimizes $\bar{\gamma}_N$, and

$$\bar{\gamma}_N(\phi^*) = \bar{\gamma}_N^* = 1 - \tfrac{1}{2}\mathbb{E}_0|L - 1|.$$

For each subset $S \subset \mathcal{V}$ of size n, let Γ_S be a decreasing event, that is, a decreasing subset of adjacency matrices, and define the truncated likelihood as

$$\tilde{L} = \frac{1}{\binom{N}{n}} \sum_{|S|=n} L_S \mathbb{1}_{\Gamma_S}. \tag{42}$$

We have

$$\begin{aligned}\mathbb{E}_0|L - 1| &\le \mathbb{E}_0|\tilde{L} - 1| + \mathbb{E}_0(L - \tilde{L}) \\ &\le \sqrt{\mathbb{E}_0[\tilde{L}^2] - 1 + 2\big(1 - \mathbb{E}_0[\tilde{L}]\big)} + \big(1 - \mathbb{E}_0[\tilde{L}]\big),\end{aligned}$$

using the Cauchy–Schwarz inequality and the fact that $\mathbb{E}_0 L = 1$ since it is a likelihood. Hence, for all tests to be asymptotically powerless, it suffices that $\limsup \mathbb{E}_0[\tilde{L}^2] \le 1$ and $\liminf \mathbb{E}_0[\tilde{L}] \ge 1$. Note that

$$\mathbb{E}_0[\tilde{L}] = \frac{1}{\binom{N}{n}} \sum_{|S|=n} \mathbb{P}_S(\Gamma_S).$$

In all our examples, $\mathbb{P}_S(\Gamma_S)$ is only a function of $|S|$, and since all the sets we consider have same size n, $\mathbb{E}_0[\tilde{L}] \to 1$ is equivalent to $\mathbb{P}_S(\Gamma_S) \to 1$.

4.1. *All tests are asymptotically powerless.* We start with some sufficient conditions under which all tests are asymptotically powerless. Recall α in (8) and ζ in (13). We require that $\zeta \to 0$ below to prevent the total degree test from having any power (see Proposition 1).

THEOREM 3. *Assume that $\zeta \to 0$. Then all tests are asymptotically powerless in any of the following situations*:

$$\lambda_0 \to 0, \lambda_1 \to 0, \qquad \limsup \frac{I_{\lambda_0}}{I_{\lambda_1}} \frac{\log n}{\log N} < 1; \tag{43}$$

$$0 < \liminf \lambda_0 \le \limsup \lambda_0 < \infty, \lambda_1 \to 0; \tag{44}$$

$$\lambda_0 \to \infty \quad \textit{with } \alpha \to 0, \qquad \limsup \lambda_1 < 1; \tag{45}$$

$$0 < \liminf \alpha \le \limsup \alpha < 1, \qquad \limsup(1 - \alpha) \sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W_{k,S}^*]}{k} < 1. \tag{46}$$

We recall here the first few steps that we took in Arias-Castro and Verzelen (2014) to derive analogous lower bounds in the denser regime where $\liminf \alpha \geq 1$. We start with some general identities. We have

$$L_S := \exp\big(\theta W_S - \Lambda(\theta) n^{(2)}\big), \tag{47}$$

with

$$\theta := \theta_{p_1}, \qquad \theta_q := \log\left(\frac{q(1-p_0)}{p_0(1-q)}\right) \tag{48}$$

and

$$\Lambda(\theta) := \log\big(1 - p_0 + p_0 e^{\theta}\big),$$

which is the cumulant generating function of $\mathrm{Bern}(p_0)$.

In all cases, the events Γ_S satisfy

$$\Gamma_S \subset \bigcap_{k > k_{\min}} \big\{W_T \leq w_k, \forall T \subset S \text{ such that } |T| = k\big\}, \tag{49}$$

where $k_{\min}$ and w_k vary according to the specific setting.

To prove that $\mathbb{E}_0 \tilde{L}^2 \leq 1 + o(1)$, we proceed as follows. We have

$$\begin{aligned}\mathbb{E}_0\big[\tilde{L}^2\big] &= \frac{1}{\binom{N}{n}^2} \sum_{|S_1|=n} \sum_{|S_2|=n} \mathbb{E}_0(L_{S_1} L_{S_2} \mathbb{1}_{\Gamma_{S_1}} \mathbb{1}_{\Gamma_{S_2}}) \\ &= \frac{1}{\binom{N}{n}^2} \sum_{|S_1|=n} \sum_{|S_2|=n} \mathbb{E}_0\big[\exp\big(\theta(W_{S_1} + W_{S_2}) - 2\Lambda(\theta) n^{(2)}\big) \mathbb{1}_{\Gamma_{S_1} \cap \Gamma_{S_2}}\big].\end{aligned}$$

Define

$$W_{S \times T} = \sum_{i \in S, j \in T} W_{i,j},$$

and note that $W_S = \frac{1}{2} W_{S \times S}$. We use the decomposition

$$W_{S_1} + W_{S_2} = W_{S_1 \times (S_1 \setminus S_2)} + W_{S_2 \times (S_2 \setminus S_1)} + 2W_{S_1 \cap S_2}, \tag{50}$$

the independence of the random variables on the RHS of (50), and the FKG inequality to get

$$\mathbb{E}_0\big(e^{\theta(W_{S_1} + W_{S_2}) - 2\Lambda(\theta) n^{(2)}} \mathbb{1}_{\Gamma_{S_1} \cap \Gamma_{S_2}}\big) \leq \mathrm{I} \cdot \mathrm{II} \cdot \mathrm{III}, \tag{51}$$

where $K = |S_1 \cap S_2|$,

$$\begin{aligned}\mathrm{I} &:= \mathbb{E}_0\left[\exp\left(\theta W_{S_1 \times (S_1 \setminus S_2)} - \frac{\Lambda(\theta)}{2}(n-K)(n+K-1)\right)\right] = 1, \\ \mathrm{II} &:= \mathbb{E}_0\left[\exp\left(\theta W_{S_2 \times (S_2 \setminus S_1)} - \frac{\Lambda(\theta)}{2}(n-K)(n+K-1)\right)\right] = 1, \\ \mathrm{III} &:= \mathbb{E}_0\big[\exp\big(2\theta W_{S_1 \cap S_2} - 2\Lambda(\theta) K^{(2)}\big) \mathbb{1}_{\Gamma_{S_1} \cap \Gamma_{S_2}}\big].\end{aligned}$$

The first two equalities are due to the fact that the likelihood integrates to one.

Assuming that $\zeta \to 0$, we prove that all tests are asymptotically powerless in the following settings:

$$\limsup \lambda_0 < \infty, \qquad \lambda_1^2 = o(\lambda_0); \tag{52}$$

$$\lambda_0 \to 0, \lambda_1 \to 0, \qquad \limsup \frac{I_{\lambda_0} \log(n)}{I_{\lambda_1} \log(N)} < 1, \qquad n^2 = o(N); \tag{53}$$

$$\limsup \lambda_1 < 1, \qquad \lambda_0 \to \infty, \qquad \limsup \alpha < 1; \tag{54}$$

$$\begin{aligned} &\liminf \lambda_1 \geq 1, \qquad 0 < \liminf \alpha \leq \limsup \alpha < 1, \\ &\limsup (1-\alpha) \sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W_{k,S}^*]}{k} < 1. \end{aligned} \tag{55}$$

This implies Theorem 3. Indeed, (52) includes (44). Assume that (43) holds. Consider any subsequence n^2/N converging to $x \in \mathbb{R}^+ \cup \{\infty\}$. If $x = 0$, then (53) holds. If $x \neq 0$, then $\zeta \to 0$ implies that $(\lambda_1 - \lambda_0 n/N)^2/\lambda_0 = o(1)$. If, in addition, $\lambda_1 \geq 2\lambda_0 n/N$, this implies that $\lambda_1^2/\lambda_0 = o(1)$. If, otherwise, $\lambda_1 \leq 2\lambda_0 n/N$, then $\lambda_1^2/\lambda_0 \leq 4\lambda_0 (n/N)^2 = o(1)$ since $\lambda_0 = o(1)$. Thus, in both cases, (52) holds. Finally, (54) includes (45) and also (46) when $\limsup \lambda_1 < 1$, while (55) includes (46) when $\liminf \lambda_1 \geq 1$. We note that (55) implies that $\limsup \lambda_1 < \infty$ because of (16).

4.1.1. *Proof of Theorem 3 under* (52). The arguments here are very similar to those used in Arias-Castro and Verzelen (2014), except for the choice of events Γ_S. Define

$$\Gamma_S := \{\mathcal{G}_S \text{ is a forest}\}.$$

When Γ_S holds, for any $T \subset S$, $\mathcal{G}_T$ is also a forest, and since any forest $\mathcal{F}$ with k nodes and t connected components (therefore all trees) has exactly $k - t \leq k$ edges, we have $W_T \leq |T|$. Hence, (49) holds with $w_k := k$.

LEMMA 13. *$\mathbb{P}_S(\Gamma_S)$ is independent of S of size n, and $\mathbb{P}_S(\Gamma_S) \to 1$.*

PROOF. The expected number of cycles of size k in $\mathcal{G}_S$ under $\mathbb{P}_S$ is equal to

$$\frac{n!}{(n-k)!2k} p_1^k \leq \frac{\lambda_1^k}{2k}. \tag{56}$$

Summing (56) over k, we see that the expected number of cycles in $\mathcal{G}_S$ under $\mathbb{P}_S$ is smaller than $\lambda_1^3/(1-\lambda_1) = o(1)$. Hence, with probability going to one under $\mathbb{P}_S$, $\mathcal{G}_S$ has no cycles and is therefore a forest. □

In order to conclude, we only need to prove that $\limsup \mathbb{E}_0[\tilde{L}^2] \le 1$. We start from (51) and we recall that $K = |S_1 \cap S_2|$. We take $k_{\min}$ as the largest integer k satisfying

$$\frac{2}{k-3} \ge \frac{p_1^2(1-p_0)}{p_0(1-p_1)^2},$$

with the convention $2/0 = \infty$, so that $k_{\min} \ge 3$. Let $q_k = 2/(k-1)$. Recall that $\rho = n/(N-n)$ and define $k_0 = \lceil bn\rho \rceil$, where $b \to \infty$ satisfies $b^2\zeta \to 0$.

- When $K \le k_{\min}$, we will use the obvious bound:

$$\mathrm{III} \le \mathbb{E}_0 \exp\bigl(2\theta W_{S_1\cap S_2} - 2\Lambda(\theta)K^{(2)}\bigr) = \exp\bigl(\Delta K^{(2)}\bigr),$$

  where

$$\Delta := \Lambda(2\theta) - 2\Lambda(\theta) = \log\biggl(1 + \frac{(p_1-p_0)^2}{p_0(1-p_0)}\biggr). \tag{57}$$

- When $K > k_{\min}$, we use a different bound. Noting that $\Gamma_{S_1} \cap \Gamma_{S_2} \subset \{W_{S_1\cap S_2} \le w_K\}$, for any $\xi \in (0, 2\theta)$, we have

$$\begin{aligned}\mathrm{III} &\le \mathbb{E}_0\bigl[\exp\bigl(\xi W_{S_1\cap S_2} + (2\theta-\xi)w_K - 2\Lambda(\theta)K^{(2)}\bigr)\mathbb{1}_{\{W_{S_1\cap S_2}\le w_K\}}\bigr]\\ &\le \mathbb{E}_0\bigl[\exp\bigl(\xi W_{S_1\cap S_2} + (2\theta-\xi)w_K - 2\Lambda(\theta)K^{(2)}\bigr)\bigr],\end{aligned}$$

  so that

$$\mathrm{III} \le \exp\bigl(\Delta_K K^{(2)}\bigr),$$

  where

$$\Delta_k := \min_{\xi\in[0,2\theta]} \Lambda(\xi) + (2\theta-\xi)q_k - 2\Lambda(\theta). \tag{58}$$

Using the fact that $\mathbb{E}_0[\tilde{L}^2] \le \mathbb{E}[\mathrm{III}]$ where the expectation is taken with respect to K, we have

$$\begin{aligned}\mathbb{E}_0\bigl[\tilde{L}^2\bigr] &\le \mathbb{E}\bigl[\mathbb{1}_{\{K\le k_0\}}\exp\bigl(\Delta K^{(2)}\bigr)\bigr]\\ &\quad + \mathbb{E}\bigl[\mathbb{1}_{\{k_0+1\le K\le k_{\min}\}}\exp\bigl(\Delta K^{(2)}\bigr)\bigr]\\ &\quad + \mathbb{E}\bigl[\mathbb{1}_{\{k_{\min}+1\le K\le n\}}\exp\bigl(\Delta_K K^{(2)}\bigr)\bigr]\\ &= A_1 + A_2 + A_3,\end{aligned}$$

where the expectation is with respect to $K \sim \mathrm{Hyp}(N,n,n)$. By Lemma 4, K is stochastically bounded by $\mathrm{Bin}(n,\rho)$. Hence, using Chernoff's bound (see Lemma 1), we have

$$\mathbb{P}(K \ge k) \le \mathbb{P}\bigl(\mathrm{Hyp}(N,n,n) \ge k\bigr) \le \mathbb{P}\bigl(\mathrm{Bin}(n,\rho) \ge k\bigr) \le \exp\bigl(-nH_\rho(k/n)\bigr). \tag{59}$$

- When $K \le k_0$, we proceed as follows. If $k_0 = 1$, we simply have
$$A_1 = \mathbb{P}(K \le 1) \le 1.$$
If $k_0 \ge 2$, we use the expression (57) of Δ to derive
$$A_1 \le \exp[\Delta k_0^2] \le \exp\left[O(1)\frac{(p_1 - p_0)^2}{p_0(1-p_0)}\frac{b^2 n^4}{N^2}\right] = \exp[O(b^2\zeta)] \to 1.$$
- When $k_0 + 1 \le K \le k_{\min}$, we use (59) and Lemma 2, to get
$$\begin{aligned} A_2 &\le \sum_{k=k_0+1}^{k_{\min}} \exp\left[\Delta\frac{k(k-1)}{2} - nH_\rho\left(\frac{k}{n}\right)\right] \\ &\le \sum_{k=k_0+1}^{k_{\min}} \exp\left[k\left(\Delta\frac{k-1}{2} - \log\left(\frac{k}{n\rho}\right) + 1\right)\right]. \end{aligned}$$
The last sum is equal to zero if $k_{\min} \le k_0$; therefore, assume that $k_{\min} > k_0$. For $a > 0$ fixed, the function $f(x) = ax - \log x$ is decreasing on $(0, 1/a)$ and increasing on $(1/a, \infty)$. Therefore, for $k_0 + 1 \le k \le k_{\min}$,
$$\Delta\frac{k-1}{2} - \log\left(\frac{k}{n\rho}\right) \le \max_{\ell \in \{k_0, k_{\min}\}}\left\{\Delta\frac{\ell - 1}{2} - \log\left(\frac{\ell N}{n^2}\right)\right\}.$$
We know that $\Delta(k_0 - 1) = o(1)$, so that
$$\Delta\frac{k_0 - 1}{2} - \log\left(\frac{k_0}{n\rho}\right) \le o(1) - \log b \to -\infty.$$
Therefore, it suffices to show that
$$\frac{k_{\min} - 1}{2}\Delta - \log\left(\frac{k_{\min}}{n\rho}\right) \to -\infty.$$
If $k_{\min} > 3$, observe that
$$\begin{aligned} \frac{k_{\min} - 1}{2}\Delta &\le \left(1 + \frac{k_{\min} - 3}{2}\right)\log\left(1 + \frac{2}{k_{\min} - 3}(1 + o(1))\right) \\ &\le \frac{3}{2}\log 3 + o(1), \end{aligned}$$
while $\log(k_{\min}/(n\rho)) \ge \log(k_0/(n\rho)) \to \infty$. If we have $k_{\min} = 3$, then we have
$$\begin{aligned} \Delta - \log\left(\frac{3}{n\rho}\right) &\le \log(p_1^2/p_0) - \log(N/n^2) + O(1) \\ &\le \log\left(\frac{\lambda_1^2}{\lambda_0}\right) + O(1) \to -\infty, \end{aligned}$$
because of (52).

- When $k_{\min} < K \le n$, we need to bound Δ_K. Remember the definition of the entropy function H_q in (10), and that $H(q)$ is short for $H_{p_0}(q)$. It is well known that H is the Fenchel–Legendre transform of Λ; more specifically, for $q \in (p_0, 1)$,

$$H(q) = \sup_{\theta \ge 0} [q\theta - \Lambda(\theta)] = q\theta_q - \Lambda(\theta_q). \tag{60}$$

Hence, the minimum of $\Lambda(\xi) + (2\theta - \xi)q_k - 2\Lambda(\theta)$ over $\xi > 0$ is achieved at $\xi = \theta_{q_k}$ as soon as $2\theta \ge \theta_{q_k}$. Moreover, by the definition of θ in (48), our choice of q_k, and the fact that $k \ge k_{\min}$, we have

$$2\theta - \theta_{q_k} = \log\left(\frac{p_1^2(1-p_0)}{p_0(1-p_1)^2}\frac{2}{k-3}\right) \ge 0.$$

Hence, we have

$$\begin{aligned} \Delta_k &= -H(q_k) + 2\theta q_k - 2\Lambda(\theta) \\ &= -2H_{p_1}(q_k) + H(q_k). \end{aligned} \tag{61}$$

Using the definition of the entropy and the fact that $p_0 = o(1)$, we therefore have

$$\begin{aligned} \Delta_k &= q_k \log\left(\frac{p_1^2}{q_k p_0}\right) + (1-q_k)\log\left(\frac{(1-p_1)^2}{(1-q_k)(1-p_0)}\right) \\ &\le \frac{2}{k-1}\left(\log\left(\frac{\lambda_1^2 N(k-1)}{2\lambda_0 n^2}\right) + O(1)\right), \end{aligned}$$

where the $O(1)$ is uniform in k. Hence, starting from the bound we got when bounding A_2, we can bound A_3 by

$$\begin{aligned} &\sum_{k=k_{\min}+1}^{n} \exp\left[k\left(\Delta_k \frac{k-1}{2} - \log\left(\frac{k}{n\rho}\right) + 1\right)\right] \\ &\quad\le \sum_{k=k_{\min}+1}^{n} \exp\left[k\left\{\log\left(\frac{\lambda_1^2}{\lambda_0}\right) + \log\left(\frac{N(k-1)}{2n^2}\right) - \log\left(\frac{Nk}{n^2}\right) + O(1)\right\}\right] \\ &\quad\le \sum_{k=k_{\min}+1}^{n} \exp\left[k\left\{\log\left(\frac{\lambda_1^2}{\lambda_0}\right) + O(1)\right\}\right] = o(1), \end{aligned}$$

since $\lambda_1^2/\lambda_0 = o(1)$.

This concludes the proof of Theorem 3 under (52).

4.1.2. *Proof of Theorem 3 under* (53). Let c be a positive constant that will be chosen small later on. Define

$$f_n := (1+c) I_{\lambda_1}^{-1} \log(n).$$

We consider the event

$$\Gamma_S = \{\mathcal{G}_S \text{ is a forest}\} \cap \big\{|\mathcal{C}_{\max,S}| \le f_n\big\}.$$

When Γ_S holds, for any $T \subset S$, $\mathcal{G}_T$ is also a forest, with $|T| - W_T$ connected components. Since the size of each connected component is at most f_n, there are at least $\lceil |T|/f_n \rceil$ connected components. Hence, (49) holds with $w_k = k - \lceil \frac{k}{f_n} \rceil$.

LEMMA 14. *$\mathbb{P}_S(\Gamma_S)$ is independent of S of size n, and $\mathbb{P}_S(\Gamma_S) \to 1$.*

PROOF. This is a straightforward consequence of Lemmas 8 and 13. □

To conclude, it suffices to show that $\mathbb{E}_0[\tilde{L}^2] \le 1 + o(1)$. For this, we will need the following.

LEMMA 15. *Let $F_{k,j}$ stand for the number of forests with j trees on k labelled vertices. For any $k \ge 2$ and any $j \le k$, $F_{k,j} \le k^{k-2}$.*

PROOF. Fix $k \ge 2$. By Cayley's formula, we have $F_{k,1} = k^{k-2}$. Therefore, it suffices to prove that $F_{k,j} \ge F_{k,j+1}$ for all $j \ge 1$. If we take a forest with j trees and erase any of its $k-j$ edges, we obtain a forest with $j+1$ trees. And there are exactly $\sum_{s \ne t} k_s k_t$ such ways of obtaining a given forest with $j+1$ trees of sizes $k_1 \le \cdots \le k_{j+1}$. Since

$$\sum_{s \ne t} k_s k_t \ge k_1(k - k_1) \ge k - 1,$$

it follows that $F_{k,j}(k-j) \ge F_{k,j+1}(k-1)$. Thus, $F_{k,j} \ge F_{k,j+1}$. □

Starting from (51), and using the fact that, under $\Gamma_{S_1} \cap \Gamma_{S_2}$, $\mathcal{G}_{S_1 \cap S_2}$ is a forest with $W_{S_1 \cap S_2} \le w_K$ edges, we have

$$\mathbb{E}_0\big[\tilde{L}^2\big] \le \mathbb{E}_0\big(\exp\big(2\theta W_{S_1 \cap S_2} - 2\Lambda(\theta) K^{(2)}\big) \mathbb{1}_{\{\mathcal{G}_{S_1 \cap S_2} \text{ is a forest}, W_{S_1 \cap S_2} \le w_K\}}\big).$$

Note that the exponential term is smaller than 1 when $|S_1 \cap S_2| \le 1$. Recall that $\rho = \frac{m}{N-m}$ and that $\Lambda(\theta) = \log[(1-p_0)/(1-p_1)]$. We first have

$$\mathbb{E}_0\big[\tilde{L}^2\big] - 1 \le \sum_{k=2}^{n} \sum_{i=1}^{w_k} \mathbb{P}[K = k, W_{S_1 \cap S_2} = i, \mathcal{G}_{S_1 \cap S_2} \text{ forest}] \exp\big[2i\theta - 2\Lambda(\theta) k^{(2)}\big]$$

with the (k,i) term in the sum being bounded by

$$\binom{n}{k}\rho^k F_{k,k-i}\frac{p_1^{2i}}{p_0^i}\left(\frac{1-p_0}{1-p_1}\right)^{2(i-k^{(2)})} \prec \left(\frac{n^2}{N}\right)^{k-i}\left(\frac{\lambda_1^2}{\lambda_0}\right)^i \frac{F_{k,k-i}\binom{n}{k}}{n^k} \prec \left(\frac{n^2 e}{N}\right)^{k-i}\left(\frac{\lambda_1^2 e}{\lambda_0}\right)^i \frac{1}{k^2},$$

so that, with a change in summation indices,

$$\begin{aligned}\mathbb{E}_0[\tilde{L}^2]-1 &\prec \sum_{j=1}^{\infty}\left(\frac{n^2 e}{N}\right)^j \sum_{i=1}^{j\lfloor f_n\rfloor}\left(\frac{\lambda_1^2 e}{\lambda_0}\right)^i \frac{1}{(i+j)^2} \\ &\prec \sum_{j=1}^{\infty}\left(\frac{n^2 e}{N}\left[1\vee\frac{\lambda_1^2 e}{\lambda_0}\right]^{f_n}\right)^j.\end{aligned} \tag{62}$$

In the second inequality, we used the fact that K is stochastically bounded by $\mathrm{Bin}(n,\rho)$ (see Lemma 4). In the third inequality, we used the fact that $p_0 < p_1$ and $i \le w_k < k$, as well as the fact that $n^2 = o(N)$, which implies that $\rho^k \sim (n/N)^k$. In the fourth inequality, we used Lemma 15 and the lower bound $k! \ge (k/e)^k$. The fifth inequality comes from a change of variables and uses the definition of w_k. When $\lambda_1^2 e \le \lambda_0$, since $n^2 = o(N)$, this sum is $O(n^2/N)$. When $\lambda_1^2 e > \lambda_0$, this sum is equal to

$$\frac{1}{e^{A_n-1}-1}, \qquad A_n := \log\left(\frac{N}{n^2}\right) - f_n \log\left(\frac{\lambda_1^2 e}{\lambda_0}\right). \tag{63}$$

So it suffices to show that $A_n \to \infty$. Since we are working under (53), there is $c > 0$ such that, eventually,

$$\frac{I_{\lambda_0}\log n}{I_{\lambda_1}\log N} \le \frac{1-c}{1+c}.$$

Then, using the fact that $\lambda_0 \vee \lambda_1 = o(1)$, we have

$$\begin{aligned} f_n \log\left(\frac{\lambda_1^2 e}{\lambda_0}\right) &= (1+c)\frac{\log n}{I_{\lambda_1}}(2\lambda_1 - 2I_{\lambda_1} + I_{\lambda_0} - \lambda_0) \\ &\le -\big(1+c+o(1)\big)\log\big(n^2\big) + (1-c)\log N \\ &\le \log\big(N/n^2\big) - c\log(N),\end{aligned}$$

eventually. This implies that $A_n \ge -1 + c\log N \to \infty$.

This completes the proof of Theorem 3 under (53).

4.1.3. *Proof of Theorem 3 under (54).* Recall that $\rho = n/(N-n)$ and define $k_0 = \lceil bn\rho \rceil$, where $b \to \infty$ satisfies $b^2\zeta \to 0$. Let $k_{\min}$ be the integer part of $1 + \frac{2}{1-\alpha}(1 \vee \frac{n^{2-\alpha}}{N^{1-\alpha}})$. Define

$$\Gamma_S = \bigcap_{k=k_{\min}+1}^{n} \{W_T \le w_k, \forall T \subset S \text{ such that } |T| = k\},$$

where $w_k := k$ here.

LEMMA 16. *For any $k > k_{\min}$ and any subset S of size n, we have $\mathbb{P}_S[\Gamma_S] \to 1$.*

This takes care of the first moment. In order to conclude, it suffices to control the second moment, specifically, to prove that $\overline{\lim}\, \mathbb{E}[\tilde{L}^2] \le 1$. Arguing as before, we have

$$\mathbb{E}_0[\tilde{L}^2] \le A_1 + A_2 + A_3,$$

where

$$A_1 := \mathbb{E}[\mathbb{1}_{\{K \le k_0\}} \exp(\Delta K^{(2)})],$$
$$A_2 := \mathbb{E}[\mathbb{1}_{\{k_0+1 \le K \le k_{\min}\}} \exp(\Delta K^{(2)})],$$
$$A_3 := \mathbb{E}_0[\mathbb{1}_{\{k_0+1 \le K \le k_{\min}\}} \exp(2\theta W_{S_1 \cap S_2} - 2\Lambda(\theta) K^{(2)}) \mathbb{1}_{\{W_{S_1 \cap S_2} \le w_K\}}].$$

- Arguing exactly as we did before, we have $A_1 = 1 + o(1)$.
- Arguing as before, we can also bound A_2 by

$$\sum_{k=k_0+1}^{k_{\min}} \exp\left[k\left(\Delta\frac{k-1}{2} - \log\left(\frac{k}{n\rho}\right) + 1\right)\right]$$
$$\le \sum_{k=k_0+1}^{k_{\min}} \exp\left[k\left(1 + o(1) + \max_{\ell \in \{k_0, k_{\min}\}}\left\{\Delta\frac{\ell-1}{2} - \log\left(\frac{\ell N}{n^2}\right)\right\}\right)\right].$$

First, we have $\Delta(k_0 - 1)/2 - \log(k_0 N/n^2) \to -\infty$. This is true if $k_0 = 1$, and when $k_0 > 1$, we have $N/n^2 \le b$, so that

$$\frac{(p_1 - p_0)^2}{p_0(1-p_0)} \sim \frac{N^2}{n^4}\zeta = \frac{N^2}{n^4 b^2} b^2 \zeta \to 0,$$

by definition of b and, therefore,

$$\Delta\frac{k_0 - 1}{2} \asymp \frac{N^2}{n^4}\zeta\frac{bn^2}{N} \le b^2\zeta \to 0.$$

We also have $\Delta(k_{\min}-1)/2-\log(k_{\min}N/n^2)\to-\infty$. To show this, we divide the analysis into two cases. When $N^{1-\alpha}\le n^{2-\alpha}$, this results from

$$\begin{aligned}\Delta\frac{k_{\min}-1}{2}&\le\big(1+o(1)\big)\frac{n^{2-\alpha}}{(1-\alpha)N^{1-\alpha}}\frac{p_1^2}{p_0}\\&=\big(1+o(1)\big)\frac{\lambda_1^2}{1-\alpha}=O(1),\end{aligned}$$

together with

$$\log\Big(\frac{k_{\min}N}{n^2}\Big)\ge\log\Big(\frac{2N^{\alpha}}{(1-\alpha)n^{\alpha}}\Big)\ge\alpha\log(N/n)\to\infty, \tag{64}$$

where we used the definition of $k_{\min}$ and the fact that $\lambda_0=(N/n)^{\alpha}$. When $N^{1-\alpha}\ge n^{2-\alpha}$, this results from

$$\begin{aligned}\Delta\frac{k_{\min}-1}{2}&\le\frac{1}{2}\Big\lfloor\frac{2}{1-\alpha}\Big\rfloor\log\Big(1+\frac{p_1^2}{p_0}\Big)+o(1)\\&\le\frac{1}{2}\Big\lfloor\frac{2}{1-\alpha}\Big\rfloor\log\Big[1+\lambda_1^2\frac{N^{1-\alpha}}{n^{2-\alpha}}\Big]+o(1)\\&\le\frac{1}{2}\Big\lfloor\frac{2}{1-\alpha}\Big\rfloor\log\Big[(1+\lambda_1^2)\frac{N^{1-\alpha}}{n^{2-\alpha}}\Big]+o(1)\\&\le\frac{1}{1-\alpha}\log(1+\lambda_1^2)+o(1)+\log(N/n^2)-\frac{\alpha}{2}-\Upsilon,\end{aligned}$$

where in the last line,

$$\Upsilon:=\begin{cases}-\dfrac{\alpha}{1-\alpha}\log(n), & \text{if } \alpha\ge1/3,\\ -\alpha\log(N/n), & \text{if } \alpha<1/3\end{cases}$$

and we have used the identity $\lfloor2/(1-\alpha)\rfloor=2$ for $\alpha<1/3$. And we also have

$$\log\Big(\frac{k_{\min}N}{n^2}\Big)\ge\log(N/n^2), \tag{65}$$

so that

$$\Delta\frac{k_{\min}-1}{2}-\log\Big(\frac{k_{\min}N}{n^2}\Big)\le\frac{1}{1-\alpha}\log(1+\lambda_1^2)-\Upsilon,$$

which goes to $-\infty$ since $\lambda_1=O(1)$ and $\alpha\log(N/n)=\lambda_0\to\infty$. Hence, we have $A_2=o(1)$.

- It remains to prove that $A_3=o(1)$. If we assume that $p_1\le2p_0$, then $\Delta_k\le\Delta\le p_0(1+o(1))$ and we can prove that $A_3=o(1)$ arguing as for A_2 above,

bounding A_3 by

$$\begin{aligned}
&\sum_{k=k_{\min}+1}^{n} \exp\biggl[k\biggl(\Delta\frac{k-1}{2} - \log\biggl(\frac{k}{n\rho}\biggr) + 1\biggr)\biggr] \\
&\quad\le \sum_{k=k_{\min}+1}^{n} \exp\biggl[k\biggl(1 + o(1) + \max_{\ell\in\{k_{\min}+1,n\}}\biggl\{\Delta\frac{\ell-1}{2} - \log\biggl(\frac{\ell N}{n^2}\biggr)\biggr\}\biggr)\biggr] \\
&\quad\le \sum_{k=k_{\min}+1}^{n} \exp\biggl[k\biggl(1 + o(1) + \Delta\frac{n}{2} - \log\biggl(\frac{k_{\min}N}{n^2}\biggr)\biggr)\biggr].
\end{aligned}$$

On one hand, we have $\Delta n \prec np_0 = (n/N)^{1-\alpha} = o(1)$. On the other hand, $\log(k_{\min}N/n^2) \to \infty$. Indeed, when $N^{1-\alpha} \le n^{2-\alpha}$, we have (64); and when $N^{1-\alpha} > n^{2-\alpha}$, then $N/n^2 > n^{\alpha/(1-\alpha)} \to \infty$ and we use (65). We conclude that $A_3 = o(1)$ when $p_1 \le 2p_0$. In the following, we suppose that $p_1 \ge 2p_0$. Leaving w_k unspecified, so we can use the same arguments later, we have that A_3 is equal to

$$\begin{aligned}
&\mathbb{E}_0\bigl[\mathbb{1}_{\{k_0+1\le K\le k_{\min}\}}\exp\bigl(2\theta W_{S_1\cap S_2} - 2\Lambda(\theta)K^{(2)}\bigr)\mathbb{1}_{\{W_{S_1\cap S_2}\le w_K\}}\bigr] \\
&\quad= \sum_{k=k_{\min}+1}^{n}\sum_{i=1}^{w_k}\mathbb{P}_0\bigl[|S_1\cap S_2| = k, W_{S_1\cap S_2} = i\bigr]\exp\bigl[2i\theta - 2k^{(2)}\Lambda(\theta)\bigr] \\
&\quad\le \sum_{k=k_{\min}+1}^{n}\sum_{i=1}^{w_k}\binom{n}{k}\rho^k\binom{k^{(2)}}{i}p_0^i(1-p_0)^{k^{(2)}-i} \\
&\qquad\qquad\times\exp\biggl[2i\log\biggl(\frac{p_1}{p_0}\biggr) + 2\bigl(k^{(2)}-i\bigr)\log\biggl(\frac{1-p_1}{1-p_0}\biggr)\biggr] \\
&\quad:= \sum_{k=k_{\min}+1}^{n}\sum_{i=1}^{w_k}B_{i,k}.
\end{aligned}$$

Furthermore, since $0 < 1-p_0 < 1$ and $1-p_1 < 1-p_0$, we have

$$B_{i,k} \le \binom{n}{k}\rho^k\binom{k^{(2)}}{i}p_0^i(p_1/p_0)^{2i} \le e^{o(k)}\biggl(\frac{en^2}{kN}\biggr)^k\biggl(\frac{ep_1^2k^{(2)}}{p_0 i}\biggr)^i, \tag{66}$$

using the standard bound $\binom{n}{k} \le (en/k)^k$.

We now specify the calculations when $w_k = k$. Considering the sums over $i = 1,\dots,k/2$ and over $i = k/2+1,\dots,k$ separately, we get

$$\begin{aligned}
\sum_{i=1}^{k}B_{i,k} &\le e^{o(k)}\biggl(\frac{en^2}{kN}\biggr)^k\biggl[\sum_{i=1}^{\lfloor k/2\rfloor}\biggl(\frac{ep_1^2k^{(2)}}{p_0}\biggr)^i + \sum_{\lfloor k/2\rfloor+1}^{k}\biggl(\frac{ep_1^2k^{(2)}}{p_0k/2}\biggr)^i\biggr] \\
&\le e^{o(k)}\biggl(\frac{en^2}{kN}\biggr)^k k\biggl[1 + \biggl(\frac{ep_1^2k^{(2)}}{p_0}\biggr)^{k/2} + \biggl(\frac{ep_1^2k^{(2)}}{p_0k/2}\biggr)^k\biggr] \\
&\prec e^{o(k)}\biggl[\biggl(\frac{en^2}{kN}\biggr)^k + \biggl(\frac{e^{3/2}n^2p_1}{N\sqrt{2p_0}}\biggr)^k + \biggl(\frac{e^2n^2p_1^2}{Np_0}\biggr)^k\biggr].
\end{aligned}$$

First, $\frac{en^2}{kN} \le \frac{en^2}{k_0 N} = o(1)$ by definition of k_0. Next, $\frac{n^2 p_1}{N\sqrt{p_0}} \le \frac{2(p_1-p_0)}{\sqrt{p_0}} \frac{n^2}{N} = 2\sqrt{\zeta} \to 0$, by the fact that $p_1 \ge 2p_0$. Finally, $n^2 p_1^2/(Np_0) = \lambda_1^2/\lambda_0 \to 0$ since $\lambda_0 \to \infty$ and $\lambda_1 = O(1)$. Hence, we conclude that

$$\sum_{k=k_{\min}+1}^{n} \sum_{i=1}^{k} B_{i,k} = o(1). \tag{67}$$

This immediately implies that $A_3 = o(1)$.

This completes the proof of Theorem 3 under (54).

PROOF OF LEMMA 16. Let us consider the event

$$\Gamma_S' := \{\text{no connected component of } \mathcal{G}_S \text{ has more than one cycle}\}.$$

Under Γ_S', a connected component of $\mathcal{G}_S$ has at most as many edges as vertices. Consequently, $\Gamma_S' \subset \Gamma_S$ and we only need to prove that $\mathbb{P}_S(\Gamma_S') \to 1$. Since $\limsup \lambda_1 < 1$ and $\mathbb{P}_S(\Gamma_S')$ is nondecreasing in λ_1, we may assume that λ_1 is fixed in $(0, 1)$.

As a warmup for what follows, we note that the number $\mathbf{L}_k$ of cycles of size k in $\mathcal{G}_S$ satisfies

$$\mathbb{E}_S[\mathbf{L}_k] = p_1^k \frac{n!}{(n-k)!2k} \le \frac{\lambda_1^k}{2k},$$

since there are $n!/[(n-k)!2k]$ potential cycles of size k. Now, if a connected component contains (at least) two cycles, there are two possibilities:

- The two cycles have at least one edge in common. In that case, there is a cycle (say of length k) with a chord (say of length $s < k$). Let $\mathbf{L}'_{k,s}$ denote the number of such configurations, there are $n!/[(n-k)!2k]$ potential cycles of size k. Given a cycle of size k, there are less than $\binom{k}{2}$ starting and ending nodes possible for the chord. Once these two nodes are set, there remains less than $n!/(n-s+1)!$ possibilities for the other nodes on the chord. Thus, we have

$$\begin{aligned}\mathbb{E}_S[\mathbf{L}'_{k,s}] &\le p_1^{k+s} \frac{n!}{(n-k)!2k} \binom{k}{2} \frac{n!}{(n-s+1)!} \\ &\le \left(\frac{\lambda_1}{n}\right)^{k+s} k n^{k+s-1} \le \lambda_1^{k+s} \frac{k}{n}.\end{aligned}$$

  Summing this inequality over s and k, we control the expected number of cycles with a chord:

$$\sum_{k=3}^{\infty} \sum_{s=1}^{k-1} \mathbb{E}[\mathbf{L}'_{k,s}] \le \frac{1}{n} \sum_{k=3}^{\infty} \frac{k\lambda_1^{k+1}}{1-\lambda_1} \asymp \frac{1}{n} = o(1),$$

since $\limsup \lambda_1 < 1$. Hence, this event occurs with probability going to 0.

- The two cycles have no edge in common. Since there are in the same connected component, there is a path that goes from a vertex in the first cycle to a vertex in the second cycle. Let us note $\mathbf{L}'_{k_1,k_2,s}$ the number of cycles of size k_1 and k_2 that do not share an edge and are connected by a path of length s. Observe that there are less $\frac{n!}{(n-k_1)!2k_1}$ possible configurations for the first cycle, less than $\frac{n!}{(n-k_2)!2k_2}$ possible configurations for the second cycle, and less than $k_1k_2n!/(n-s+1)!$ possibilities for the chord. Thus, we get

$$\mathbb{E}[\mathbf{L}'_{k_1,k_2,s}] \le p_1^{k_1+k_2+s} \frac{n!}{(n-k_1)!2k_1} \frac{n!}{(n-k_2)!2k_2} k_1k_2 \frac{n!}{(n-s+1)!}$$
$$\le \left(\frac{\lambda_1}{n}\right)^{k_1+k_2+s} n^{k_1+k_2+s-1} = \frac{\lambda_1^{k_1+k_2+s}}{n},$$

  so that the expected number of such configurations is bounded as follows:

$$\sum_{k_1\ge 3}\sum_{k_2\ge 3}\sum_{s\ge 1}\mathbb{E}[\mathbf{L}'_{k_1,k_2,s}] \le \frac{1}{n}\sum_{k_1\ge 3}\sum_{k_2\ge 3}\sum_{s\ge 1}\lambda_1^{k_1+k_2+s} \asymp \frac{1}{n} = o(1).$$

  Hence, this second event occurs with probability going two zero.

All in all, we have proved that $\mathbb{P}_S(\Gamma'_S) \to 1$, implying that $\mathbb{P}_S(\Gamma_S) \to 1$. □

4.1.4. *Proof of Theorem 3 under* (55). We follow the arguments laid out for the case (54). We define Γ_S in the same way, except that $w_k := [k\frac{(1-c)^{1/2}}{1-\alpha}]$, where c is a positive constant (to be chosen small later) such that $c < \alpha$ and, eventually,

$$\sup_{n/u_N<k\le n} \frac{1}{k}\mathbb{E}_S[W^*_{k,S}] \le \frac{1-2c}{1-\alpha}. \tag{68}$$

LEMMA 17. *For any $k > k_{\min}$ and any subset S of size n, we have* $\mathbb{P}_S[\Gamma_S] \to 1$.

For the second moment, we proceed exactly as in the case (54), and we start from (67). In fact, when $w_k \le k$, the proof is complete. So we assume that c is small enough that $w_k > k$, and bound the sum over $k+1 \le i \le w_k$. For $i > k$, we use the bound (66), together with the fact that $\lambda_0 = (N/n)^\alpha$ and $k < i$, to derive

$$B_{i,k} \le e^{o(k)}\left(\frac{en^2}{kN}\right)^k \left(\frac{ep_1^2k^{(2)}}{p_0 i}\right)^i$$
$$\le e^{o(k)}\left(\frac{en^2}{kN}\right)^k \left(\frac{N^{1-\alpha}k}{n^{2-\alpha}}\frac{\lambda_1^2 e}{2}\right)^i$$
$$= e^{o(k)+k}\left(\frac{n}{N}\right)^{k-i(1-\alpha)} \left(\frac{\lambda_1^2 e}{2}\right)^i \left(\frac{n}{k}\right)^{k-i}$$
$$\le e^{o(k)+k}\left(\frac{n}{N}\right)^{k-i(1-\alpha)} \left(\frac{\lambda_1^2 e}{2}\right)^i.$$

This allows us to control the sum

$$\begin{aligned}\sum_{i=k+1}^{w_k} B_{i,k} &\le w_k e^{o(k)+k}\left(\frac{n}{N}\right)^{k-(1-\alpha)w_k}\left(\frac{\lambda_1^2 e}{2}\vee 1\right)^{w_k}\\ &\prec k e^{o(k)+k}\left(\frac{n}{N}\right)^{k(1-(1-c)^{1/2})}\left(\frac{\lambda_1^2 e}{2}\vee 1\right)^{k((1-c)^{1/2})/(1-\alpha)}\\ &= \exp\bigl[O(k) - k\bigl(1-(1-c)^{1/2}\bigr)\log(N/n)\bigr],\end{aligned}$$

where in the second line we used the fact that $w_k = O(k)$ since $\limsup \alpha < 1$, and in the third line we used the fact that $\lambda_1 = O(1)$. Thus,

$$\sum_{k=k_{\min}+1}^{n}\sum_{i=k+1}^{w_k} B_{i,k} = o(1),$$

which together with (67) allows us to conclude that $A_3 = o(1)$.

This completes the proof of Theorem 3 under (55).

PROOF OF LEMMA 17. Recall that $u_N = \log\log(N/n)$. First we consider integers k satisfying $k_{\min}+1 \le k < n/u_N$. Define $\omega_k' = k(1-c)^{-1/2}(\frac{\lambda_1}{2}\vee 1)$ and $q_k' = \omega_k'/k^{(2)}$. Applying a union bound and Chernoff's bound for the binomial distribution, we derive that

$$\begin{aligned}\mathbb{P}_S\bigl[W_{k,S}^* \ge \omega_k'\bigr] &\le \binom{n}{k}\mathbb{P}\bigl[\mathrm{Bin}\bigl(k^{(2)}, p_1\bigr) \ge \omega_k'\bigr]\\ &\le \exp\biggl[k\biggl\{\log\biggl(\frac{ne}{k}\biggr) - \frac{k-1}{2}H_{p_1}(q_k')\biggr\}\biggr].\end{aligned}$$

Since $k/n \le 1/u_N = o(1)$, and since λ_1 is bounded, we have $q_k'/p_1 \to \infty$, so that

$$\begin{aligned}&\frac{k-1}{2}H_{p_1}(q_k')\\ &\quad\sim \frac{k-1}{2}q_k'\log\biggl(\frac{q_k'}{p_1}\biggr)\\ &\quad= (1-c)^{-1/2}\biggl[\frac{\lambda_1}{2}\vee 1\biggr]\biggl[\log\biggl(\frac{n}{k-1}\biggr) + \log\biggl\{(1-c)^{-1/2}\biggl(1\vee\frac{2}{\lambda_1}\biggr)\biggr\}\biggr]\\ &\quad\ge \bigl(1+o(1)\bigr)(1-c)^{-1/2}\log\biggl(\frac{n}{k}\biggr),\end{aligned}$$

and, therefore, since $c \in (0,1)$ is fixed,

$$\log\biggl(\frac{ne}{k}\biggr) - \frac{k-1}{2}H_{p_1}(q_k') \le 1 + \bigl[1-\bigl(1+o(1)\bigr)(1-c)^{-1/2}\bigr]\log(u_N) \to -\infty.$$

We conclude that

$$\sum_{k=k_{\min}+1}^{n/u_N} \mathbb{P}_S\big[W^*_{k,S} \geq \omega'_k\big] = o(1).$$

Let us now prove that $\omega'_k \leq w_k$. Indeed, this inequality holds if, and only if, $\lambda_1 \leq 2(1-c)/(1-\alpha)$ and $c \leq \alpha$. The second inequality is by definition of c, while the first inequality is ensured by (68) since

$$\frac{\lambda_1}{2}\frac{n-1}{n} = \mathbb{E}_S\big[W^*_{n,S}/n\big] \leq \sup_{k\leq n} \mathbb{E}_S\big[W^*_{k,S}/k\big] \leq (1-2c)/(1-\alpha).$$

Let us turn to integers k satisfying $k \geq n/u_N$. Let $c_0 = (1-c)^{-1/2} - 1$ and $t = c_0\mathbb{E}_S[W^*_{k,S}]$. By taking any fixed subset $T \subset S$ of size $|T| = k$, we derive

$$\mathbb{E}_S\big[W^*_{k,S}\big] \geq \mathbb{E}_S[W_T] = p_1 k^{(2)} \geq \frac{\lambda_1}{n}(n/u_N)^{(2)} \asymp \frac{n}{u_N^2} \to \infty, \tag{69}$$

so that t satisfies the condition of Lemma 7 eventually. Using that lemma, we derive that

$$\mathbb{P}_S\big[W^*_{k,S} \geq \mathbb{E}_S\big[W^*_{k,S}\big](1-c)^{-1/2}\big] \leq \exp\bigg[-\mathbb{E}_S\big[W^*_{k,S}\big]\frac{\log(2)}{4}c_0\bigg[1 \wedge \frac{c_0}{8}\bigg]\bigg].$$

By condition (68), $w_k \geq \mathbb{E}_S[W^*_{k,S}](1-c)^{-1/2}$. Hence, there exists a positive constant κ, such that

$$\begin{aligned}\sum_{k=n/u_N}^{n} \mathbb{P}_S\big[W^*_{k,S} \geq w_k\big] &\leq \sum_{k=n/u_N}^{n} \exp\big[-\kappa\mathbb{E}_S\big[W^*_{k,S}\big]\big] \\ &\leq n\exp\big[-\kappa\mathbb{E}_S\big[W^*_{n/u_N,S}\big]\big].\end{aligned}$$

Because of (69) and the fact that $\log(N) = o(n)$, we have

$$\mathbb{E}_S\big[W^*_{n/u_N,S}\big] \succ \frac{n}{\log^2(n)},$$

and, therefore, the sum above goes to 0. □

4.2. *No test is asymptotically powerful.* When λ_0 is bounded away from 0 and infinity, the triangle test has some nonnegligible power as long as λ_1 is bounded away from 0 (see Section 3.3). This motivates us to obtain sufficient conditions under which no test is asymptotically powerful.

Our method is also based on bounding the first two moments of a truncated likelihood ratio $\tilde{L}$. Indeed, it is enough to show that $\liminf \mathbb{E}_0\tilde{L} > 0$ and $\liminf \mathbb{E}_0[\tilde{L}^2] < \infty$. This comes from the following result.

LEMMA 18. *Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two probability distributions on the same probability space, with densities f_0 and f_1 with respect to some dominating measure. Let Γ be any event and define the truncated likelihood ratio $\tilde{L} = L\mathbb{1}_\Gamma$, where $L = f_1/f_0$ is the likelihood ratio for testing $\mathbb{P}_0$ versus $\mathbb{P}_1$. Then any test for $\mathbb{P}_0$ versus $\mathbb{P}_1$ has risk at least*

$$\frac{4}{27}\frac{(\mathbb{E}_0\tilde{L})^3}{\mathbb{E}_0[\tilde{L}^2]},$$

where $\mathbb{E}_0$ denotes the expectation under $\mathbb{P}_0$, and by convention $0/0 = 0$.

PROOF. Assume $\mathbb{E}_0\tilde{L} \neq 0$, for otherwise the result is immediate. The risk of the likelihood ratio test $\{L > 1\}$—which is the test that optimizes the risk—is equal to

$$B := 1 - \tfrac{1}{2}\mathbb{E}_0|L - 1| = 1 - \mathbb{E}_0(1 - L)_+ \geq 1 - \mathbb{E}_0(1 - \tilde{L})_+,$$

since $\tilde{L} \leq L$. For any $t \in (0, 1)$, we have

$$\mathbb{E}_0(1 - \tilde{L})_+ \leq (1 - t)\mathbb{P}_0(\tilde{L} > t) + \mathbb{P}_0(\tilde{L} \leq t) = 1 - t\mathbb{P}_0(\tilde{L} > t).$$

Moreover, using the Cauchy–Schwarz inequality, we have for any $t > 0$

$$\begin{aligned}\mathbb{E}_0\tilde{L} &= \mathbb{E}_0[\tilde{L}\mathbb{1}_{\{\tilde{L}\leq t\}}] + \mathbb{E}_0[\tilde{L}\mathbb{1}_{\{\tilde{L}>t\}}] \\ &\leq t + \sqrt{\mathbb{E}_0[\tilde{L}^2]\mathbb{P}_0(\tilde{L} > t)},\end{aligned}$$

so that, taking $t < \mathbb{E}_0\tilde{L}$, we have

$$\mathbb{P}_0(\tilde{L} > t) \geq \frac{(\mathbb{E}_0\tilde{L} - t)^2}{\mathbb{E}_0\tilde{L}^2}.$$

We conclude that

$$B \geq t\mathbb{P}_0(\tilde{L} > t) \geq t\frac{(\mathbb{E}_0\tilde{L} - t)^2}{\mathbb{E}_0\tilde{L}^2},$$

and optimizing this over $0 < t < \mathbb{E}_0\tilde{L}$ yields the result. □

Since we only need to focus on the case where λ_0 is bounded from 0 and infinity, and where λ_1 is bounded from 0 (because the other cases are covered by Theorem 3), we may assume they are fixed without loss of generality. In that case $\zeta \to 0$ is equivalent to $n^2/N \to 0$, which is what we assume in the following.

THEOREM 4. *Write $n = N^\kappa$ with $0 < \kappa < 1/2$, and assume that λ_0 and λ_1 are both fixed. No test is asymptotically powerful in all the following situations*:

$$\lambda_1 < 1, \qquad \lambda_1^2 e \leq \lambda_0; \tag{70}$$

$$\lambda_1 < 1, \lambda_1^2 e > \lambda_0, \qquad \frac{1 - 2\kappa}{\kappa}\frac{I_{\lambda_1}}{\log((e\lambda_1^2)/\lambda_0)} > 1. \tag{71}$$

PROOF. We use the same truncation as in Section 4.1.2, using the same notation Γ_S and f_n defined there, and still denote the resulting truncated likelihood by $\tilde{L}$.

For the first moment, by symmetry,

$$\mathbb{E}_0[\tilde{L}] = \mathbb{P}_S[\Gamma_S] = \mathbb{P}_S\big[\mathcal{G}_S \text{ is a forest}, |\mathcal{C}_{\max,S}| \le f_n\big].$$

We already saw that $\mathbb{P}_S[|\mathcal{C}_{\max,S}| \le f_n] \to 1$ [Van der Hofstad (2012), Theorem 4.4]. Consequently,

$$\mathbb{E}_0[\tilde{L}] = \mathbb{P}_S[\mathcal{G}_S \text{ is a forest}] + o(1).$$

Of course, $\mathcal{G}_S$ is a forest if, and only if, it has no cycles. By Takács (1988), the number of cycles in $\mathcal{G}_S$ converges weakly to a Poisson distribution with mean

$$a(\lambda_1) = \frac{1}{2}\log\Big(\frac{1}{1-\lambda_1}\Big) - \frac{\lambda_1}{2} - \frac{\lambda_1^2}{4},$$

when $\lambda_1 < 1$ is fixed. As a consequence, $\mathbb{E}_0[\tilde{L}] = \exp[-a(\lambda_1)] + o(1)$, which remains bounded away from zero.

For the second moment, we start from (62):

$$\mathbb{E}_0[\tilde{L}^2] - 1 \prec \sum_{j=1}^{\infty}\Big(\frac{n^2 e}{N}\Big[1 \vee \frac{\lambda_1^2 e}{\lambda_0}\Big]^{f_n}\Big)^j,$$

with $f_n = (1+c)I_{\lambda_1}^{-1}\log n$ and c is a small positive constant. Under (70), we have $\lambda_1^2 e \le \lambda_0$ and the RHS is $O(n^2/N) = o(1)$. Under (71), we have $\lambda_1^2 e > \lambda_0$, and the RHS is, as before, equal to (63). Here, we have

$$A_n = \Big[1 - 2\kappa - (1+c)\frac{\kappa}{I_{\lambda_1}}\log\Big(\frac{\lambda_1^2 e}{\lambda_0}\Big)\Big]\log N \to \infty,$$

when (71) is satisfied and c is small enough. Hence, in any case, we found that $\mathbb{E}_0[\tilde{L}^2] \le 1 + o(1)$. □

5. Discussion.

5.1. *Adapting to unknown p_0 and n.* In Arias-Castro and Verzelen (2014), we discussed in detail the case where p_0 is unknown. In this situation, the total degree test is not applicable, and we replaced it with a test based on the difference between two estimates for the degree variance. On the other hand, the scan test [based on (4)] can be calibrated in various ways without asymptotic loss of power, for example, by plugging in the estimate $\hat{p}_0 = \frac{W}{N^{(2)}}$ in place of p_0. We showed that a combination of degree variance test and the scan test are optimal when p_0 is unknown, so that the degree variance test can truly play the role of the total degree test in this situation. We believe this is the case here also. In addition to that, the

broad scan test [based on (7)] can also be calibrated without asymptotic loss of power, and the same is true for all the other tests that we studied here, except for the largest connected component test in the supercritical regime.

We also discussed in Arias-Castro and Verzelen (2014) the case where the size of the subgraph n is unknown. This only truly affects the broad scan test, whose definition itself depends on n. As we argued in our previous paper, it suffices to apply the procedure to all possible n's, meaning, consider the multiple test based on a combination of the statistics

$$W_n^{\ddagger}, \qquad n = 1, \dots, N/2$$

with a Bonferroni correction. The concentration inequalities that we obtained for $W_n^{\ddagger}$ can accommodate an additional logarithmic factor that comes out of applying the union to control this statistic under $\mathbb{P}_0$, and from this we can immediately see that the test is asymptotically as powerful (up to first order).

5.2. *Open problems.* The cases where $\lambda_0 \to 0$ and where $\liminf \lambda_0 \geq e$ are essentially resolved. Indeed, in the first situation, the largest connected component test is asymptotically optimal by Theorem 2 and Theorem 3 case (43), while in the second situation the broad scan test is asymptotically optimal by Theorem 1 and Theorem 3 cases (45) and (46), together with Theorem 4. The case where $0 < \lambda_0 < e$ is fixed is not completely resolved. Since the triangle test has nonnegligible power as soon as λ_1 is bounded away from 0, consider τ defined as the largest real such that no test for $\mathbb{G}(N, \frac{\lambda_0}{N})$ versus $\mathbb{G}(N, \frac{\lambda_0}{N}; n, \frac{\lambda_1}{n})$ is asymptotically powerful when $\limsup \lambda_1 < \tau$. Theorem 2 and the result we obtain in Verzelen and Arias-Castro (2013) for the test based on counting subtrees of size k provide some upper bounds on τ.

OPEN PROBLEM 1. Compute τ as a function of λ_0 and $\kappa := \limsup \frac{\log n}{\log N}$.

Although we proved that the broad scan test was asymptotically optimal when $\liminf \lambda_0 \geq e$, its performance was described only indirectly in terms of λ_1 in the case (14).

OPEN PROBLEM 2. Compute, as a function of λ_1, the limits inferior and superior of

$$\sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W_{k,S}^*]}{k}.$$

We also formulate an open problem that connects directly with the planted clique problem. We saw that the broad scan test is powerful when λ_1 is sufficiently large, but we do not know how to compute it in polynomial time. Is there a polynomial-time test that can come close to that?

OPEN PROBLEM 3. Find a polynomial-time test that is asymptotically powerful for testing $\mathbb{G}(N, p_0)$ versus $\mathbb{G}(N, p_0; n, p_1)$ when $n^2/N = O(1)$, while $\lambda_0 \to \infty$ and $\lambda_1 = O(1)$.

Acknowledgements. We are grateful to Jacques Verstraete and Raphael Yuster for helpful discussions and references on counting k-cycles.

REFERENCES

ALON, N., KRIVELEVICH, M. and SUDAKOV, B. (1998). Finding a large hidden clique in a random graph. In *Proceedings of the Eighth International Conference "Random Structures and Algorithms"* (*Poznan*, 1997) **13** 457–466. MR1662795

ARIAS-CASTRO, E. and VERZELEN, N. (2014). Community detection in dense random networks. *Ann. Statist.* **42** 940–969. MR3210992

BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

BOUCHERON, S., BOUSQUET, O., LUGOSI, G. and MASSART, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.* **33** 514–560. MR2123200

BUTUCEA, C. and INGSTER, Y. I. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli* **19** 2652–2688. MR3160567

DEKEL, Y., GUREL-GUREVICH, O. and PERES, Y. (2011). Finding hidden cliques in linear time with high probability. In *ANALCO*11—*Workshop on Analytic Algorithmics and Combinatorics* 67–75. SIAM, Philadelphia, PA. MR2815485

FEIGE, U. and RON, D. (2010). Finding hidden cliques in linear time. In 21*st International Meeting on Probabilistic*, *Combinatorial*, *and Asymptotic Methods in the Analysis of Algorithms* (*AofA'*10) 189–203. Assoc. Discrete Math. Theor. Comput. Sci., Nancy. MR2735341

FORTUNATO, S. (2010). Community detection in graphs. *Phys. Rep.* **486** 75–174. MR2580414

GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826 (electronic). MR1908073

HEARD, N. A., WESTON, D. J., PLATANIOTI, K. and HAND, D. J. (2010). Bayesian anomaly detection methods for social networks. *Ann. Appl. Stat.* **4** 645–662. MR2758643

LANCICHINETTI, A. and FORTUNATO, S. (2009). Community detection algorithms: A comparative analysis. *Phys. Rev. E* **80** 056117.

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

MASLOV, S., SNEPPEN, K. and ZALIZNYAK, A. (2004). Detection of topological patterns in complex networks: Correlation profile of the internet. *Physica A*: *Statistical Mechanics and Its Applications* **333** 529–540.

MONGIOVÌ, M., BOGDANOV, P., RANCA, R., PAPALEXAKIS, E. E., FALOUTSOS, C. and SINGH, A. K. (2013). NetSpot: Spotting significant anomalous regions on dynamic networks. In *SIAM International Conference on Data Mining*, *Austin*, *TX* 28–36. SIAM, Philadelphia, PA.

MOSSEL, E., NEEMAN, J. and SLY, A. (2012). Stochastic block models and reconstruction. Available at arXiv:1202.1499.

NEWMAN, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577–8582.

NEWMAN, M. E. J. and GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* **69** 026113.

PARK, Y., PRIEBE, C. E. and YOUSSEF, A. (2013). Anomaly detection in time series of graphs using fusion of graph invariants. *IEEE J. Sel. Top. Signal Process.* **7** 67–75.

PITTEL, B. and WORMALD, N. C. (2005). Counting connected graphs inside-out. *J. Combin. Theory Ser. B* **93** 127–172. MR2117934

REICHARDT, J. and BORNHOLDT, S. (2006). Statistical mechanics of community detection. *Phys. Rev. E* (3) **74** 016110, 14. MR2276596

ROBINSON, R. W. and WORMALD, N. C. (1992). Almost all cubic graphs are Hamiltonian. *Random Structures Algorithms* **3** 117–125. MR1151355

ROBINSON, R. W. and WORMALD, N. C. (1994). Almost all regular graphs are Hamiltonian. *Random Structures Algorithms* **5** 363–374. MR1262985

RUKHIN, A. and PRIEBE, C. E. (2012). On the limiting distribution of a graph scan statistic. *Comm. Statist. Theory Methods* **41** 1151–1170.

SUN, X. and NOBEL, A. B. (2008). On the size and recovery of submatrices of ones in a random binary matrix. *J. Mach. Learn. Res.* **9** 2431–2453. MR2460888

TAKÁCS, L. (1988). On the limit distribution of the number of cycles in a random graph. *J. Appl. Probab.* **25A** 359–376. MR0974594

VAN DER HOFSTAD, R. (2012). Random Graphs and Complex Networks. Available at http://www.win.tue.nl/~rhofstad/NotesRGCN.pdf.

VERZELEN, N. and ARIAS-CASTRO, E. (2013). Community detection in sparse random networks. Available at arXiv:1308.2955.

WANG, B., PHILLIPS, J. M., SCHREIBER, R., WILKINSON, D. M., MISHRA, N. and TARJAN, R. (2008). Spatial scan statistics for graph clustering. In *SIAM International Conference on Data Mining*, *Atlanta*, *GA* 727–738. SIAM, Philadelphia, PA.

WORMALD, N. C. (1999). Models of random regular graphs. In *Surveys in Combinatorics*, 1999 (*Canterbury*). *London Mathematical Society Lecture Note Series* **267** 239–298. Cambridge Univ. Press, Cambridge. MR1725006

INRA
UMR 729 MISTEA
2 PLACE VIALA, BÂT. 29
F-34060 MONTPELLIER
FRANCE
E-MAIL: nicolas.verzelen@inra.fr

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CALIFORNIA 92093-0112
USA
E-MAIL: eariasca@ucsd.edu