Large deviations of connected components in the stochastic block model

Hendrik Schawe[1,2,*] and Alexander K. Hartmann[2,†]
[1]*Laboratoire de Physique Théorique et Modélisation, UMR-8089 CNRS, CY Cergy Paris Université, 95000 Cergy, France*
[2]*Institut für Physik, Universität Oldenburg, 26111 Oldenburg, Germany*

(Received 6 March 2020; accepted 19 October 2020; published 5 November 2020)

We study the stochastic block model, which is often used to model community structures and study community-detection algorithms. We consider the case of two blocks in regard to its largest connected component and largest biconnected component, respectively. We are especially interested in the distributions of their sizes including the tails down to probabilities smaller than 10^{-800}. For this purpose we use sophisticated Markov chain Monte Carlo simulations to sample graphs from the stochastic block model ensemble. We use these data to study the large-deviation rate function and conjecture that the large-deviation principle holds. Further we compare the distribution to the well-known Erdős-Rényi ensemble, where we notice subtle differences at and above the percolation threshold.

DOI: 10.1103/PhysRevE.102.052108

I. INTRODUCTION

The stochastic block model (SBM) [1] is a generative model for networks with community structure. For this purpose, each node is assigned to one of B *blocks*. Similar to the Erdős-Rényi (ER) model [2], edges between pairs of nodes appear with some probability. For the SBM, these probabilities can depend on the blocks each node belongs to. Thus, the probabilities for edges between or within the blocks can be encoded in the $B \times B$ *block matrix*. On the one hand this makes the model very versatile with an arbitrary number of blocks and arbitrary probabilities between the blocks; on the other hand it still stays simple in the sense that it is an ensemble of random graphs without any further correlations between the edges like the ER graph ensemble or configuration model [3]. Indeed, in the case of $B = 1$ it simplifies to an ER ensemble.

In statistical physics there is a persistent interest in the stochastic block model as a tool for community detection, i.e., given a network, what is the block matrix, and to which blocks do the nodes belong most probably if this realization was drawn from an ensemble of stochastic block models. This problem shows interesting behavior as it exhibits two phases: one in which a reconstruction of the parameters is possible—studying different approaches how to do that is another active field of studies [4–9]—and another phase, where the reconstruction is infeasible [10–12]. In general, the determination of community structures is algorithmically challenging. This motivated our study, because we are interested in whether the detectability is related to the simpler properties of the system, like the size of the largest cluster. Here, in anticipation of our results, we find that we can indeed recognize whether there is some kind of block structure present for a small parameter range, especially when also considering the far tails of its distribution. But this distinguishability seems to be related to the percolation threshold instead of the detectability threshold.

Usually, systems modeled by networks have some kind of functionality, e.g., communication networks enable information exchange between nodes, power grids enable power transmission between producers and consumers, and social networks exchange, for example, opinions over the edges. As a very simple but general indicator of the functionality for sparse networks, the size S of the largest connected component is useful and the most simple global network property of any ensemble. Hence, we study here the distribution of S for the SBM. The average behavior of S determines the percolation transition. As we will show below, the percolation transition of the SBM is related to the ER simply by using an effective (average) connectivity. This could indicate that the distribution of S is the same for SBM and the ER with this effective connectivity. But this is not the case, as we will show below, in the percolating region.

Furthermore, since networks consist of many nodes, which often symbolize entities that can fail or vanish, the robustness against this kind of events is relevant. A common idea [13–17] to measure robustness is to remove one or several nodes, either randomly or according to "attack" rules, and measure its impact on the functionality. Here, since we are measuring the functionality in terms of the size of the largest connected component, we also measure the robustness in terms of the size of the largest biconnected component, i.e., the subgraph that will stay connected if any node was removed. Note that this observable is not an uncommon choice to determine robustness [18].

We scrutinize these properties in very great detail, i.e., not only do we look at their mean size, but we also obtain their probability distributions over practically the whole support, especially including very rare events with a probability of less than $p = 10^{-800}$. This is, to our knowledge, the first time the tails of the probability density function of these properties have been studied for any stochastic block model.

*hendrik.schawe@cyu.fr
†a.hartmann@uol.de

In large-deviation theory [19], many probability distributions have a special shape which allows one to remove the leading finite-size influence and describe the distributions by the so-called *rate function*. As we will show below, here we find a comparatively fast convergence of the empirical rate functions calculated from the finite-size distributions. This enables us to observe the complete large-deviation-rate function almost directly and conjecture that the *large-deviation principle* [19] holds for this distribution.

The motivation to study these properties in such great detail is mainly fundamental interest in the behavior of these ensembles. The deep tails, which we explore here too, should lead to a deeper understanding of fringe cases. Also, we hope that our numerical high-precision studies motivate analytical work in this direction.

The latter seems possible because the behavior of the studied observables is known analytically for the related ER ensemble [20]. Corresponding large-deviation results were also obtained using simulational techniques [21,22]. Since the ER is a special case of the stochastic block model and is in general a good *null model* to compare other graph ensembles to, we compare and contrast it to the SBM. We even show results for the ER ensemble for larger sizes than studied in Refs. [21,22].

As mentioned in the last paragraph, we previously studied similar subjects. Starting with Ref. [21], motivated by an analytical expression for the rate function for the size of the largest connected component for ER [20], one of us compared this expression with measurements obtained from simulations and found a rather fast convergence of the measurements to the rate function valid in the asymptotic limit, as even estimates obtained from graphs with a few hundred nodes showed already a very good convergence. Also in the same publication a similar analysis is performed for regular lattices instead of ER. Later, in Ref. [23] the distribution of the diameter for ER was obtained similarly, and in Refs. [22,24] we extended these results to the distribution of the size of the 2-core and the biconnected component. In the latter we also studied another graph ensemble, the famous Barabási-Albert ensemble of scale–free graphs [25]. Here we extend this line of work further by examining the connected component and biconnected component on yet another ensemble of random graphs: a simple case of the SBM.

This case study of a few parameter values of a simple SBM can surely not be generalized to all SBMs. So the results we will show are primarily applicable to the parameters studied. However, often the results give insight into the mechanism leading to a specific behavior, in which case we will make educated guesses for which more general cases we expect to observe similar phenomena.

II. MODELS AND METHODS

A *graph* is a tuple $G = (V, E)$ of a set of *nodes* V and a set of *edges* E. The number of nodes $|V| = N$ is called the *size* of the graph. Here we will scrutinize only *undirected*, *simple* graphs, i.e., $E \subset V^{(2)} \setminus \{\{u, u\} | u \in V\}$. For each node the number of incident edges is its *degree*. Since graphs are used to model relations between objects, one of the most fundamental properties of graphs is their connectedness. Fundamentally, only nodes i, j which are *connected* via a *path*, i.e., a sequence of edges $\{\{i, u_1\}, \{u_1, u_2\}, \ldots, \{u_m, j\}\}$, can interact at all with each other. The maximal subsets whose members are connected are called *connected components*, their *size* is the number of elements. It is therefore of interest if a given graph is connected or what the size of its largest connected component is.

The functionality of a network is for many applications directly dependent on a large connected component. For example, in a power delivery network—in the best case—every producer could pass its power to any consumer, in a communication network it is desirable that every member can communicate with any other member, in a network encoding physical contacts between subjects, small connected components would be advantageous to inhibit the spreading of disease. While in all these cases maybe other observables might capture the functionality better, the size of the largest connected component is a reasonable first approximation. In the following we will mainly consider its relative size S.

As a second observable we take a look at the closely related *biconnected components*, which are the maximal subsets whose members are connected by two node-independent paths. This means that one can remove any node from a biconnected component and the remainder will still be a connected component. The relative size S_2 of the largest biconnected component is therefore the most simple quantity to judge the *robustness* against node removal or failure of a network.

Algorithmically, one can determine the size of all connected and biconnected components in time $O(|V| + |E|)$ by performing one modified depth first search on a given graph [26–28]. Note that a node can be part of two distinct biconnected components, such that the sum of the sizes of all biconnected components might be larger than N.

A. Graph ensembles

The Erdős-Rényi (ER) graph is probably the simplest and first studied random graph ensemble [2]. It consists of N nodes, and any possible edge exists independently from all other edges with a probability of p. If one is interested in sparse graphs, it is convenient to parametrize the ensemble with the *connectivity* $c = Np$, which is equal to the expected degree. In particular, the ER ensemble shows a phase transition from a forest-like structure with connected components of size $O(\log N)$ to a structure with one giant connected component of size $O(N)$ when increasing c above the critical threshold of $c_c = 1$ [2]. Note that beyond the same threshold $c_c = 1$ a giant biconnected component of size $O(N)$ arises [18].

The stochastic block model (SBM) is a random graph ensemble in which every node belongs with probability P_b to *block* b. Similar to the ER the edges exist independently with a fixed probability, but in the SBM the probability of the edge $\{i, j\}$ to exist depends on the blocks a, b of which i and j are members of, i.e., p_{ab}. The diagonal of this block matrix governs how tightly connected the nodes within a block are, and the off-diagonal elements determine how tightly the connections between distinct blocks are; e.g., if the diagonal is zero, every realization will be bipartite. If the diagonal elements are larger than the off-diagonal, the SBM is called

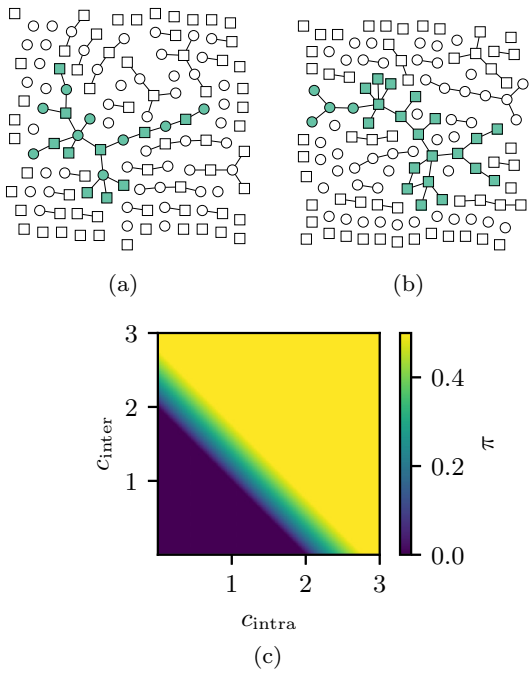

FIG. 1. Two example realizations of the SBM with size $N = 128$ with two blocks (shape of nodes) of equal probability $P_a = P_b = 0.5$. The panels show realizations with different connectivities at the percolation threshold (a) $c_{\text{intra}} = 0.1$, $c_{\text{inter}} = 1.9$ and (b) $c_{\text{intra}} = 1.9$, $c_{\text{inter}} = 0.1$. The largest connected components are visualized with colored symbols. (c) Numerical solution of Eq. (4) showing the percolation threshold of our particular SBM model. Note that the lightest shade also signifies sizes $\pi > 0.5$.

assortative; if the off-diagonal elements are larger than the diagonal, it is called *disassortative*. Note that a homogeneous $p_{ab} = p$ is equivalent to the ER. Since we will study sparse SBM, we will parametrize the ensemble with connectivities:

$$c_{ab} = N p_{ab}. \tag{1}$$

We want to perform a very in-depth study of an SBM ensemble, therefore we will treat the simplest special case of SBM, the *planted partition*; i.e., all blocks have the same intrablock (diagonal) connectivity c_{intra} and the same interblock (off-diagonal) connectivity c_{inter}. Further we mainly handle the simplest case, which is distinct from ER, i.e., $B = 2$ blocks, but later we also show some results for $B = 3$. Figure 1 shows two examples for different values of c_{inter} and c_{intra}.

The phase transition where a giant connected component of size $O(N)$ arises happens for $B = 2$ at $(c_{\text{intra}} + c_{\text{inter}})/2 = 1$ [cf. Eq. (5)]. One can derive this threshold by estimating the size of the largest connected component analogously to a method for ER [29].

We estimate the size of the largest connected component by considering the probability π that a randomly selected end node of a randomly selected edge is connected via other edges with the giant component of the graph. This means that π is the relative size of the largest connected component in the asymptotic limit. Note that for the simple ER model with connectivity c, the probability q_d for degree d of a random-end node of a random edge is given by $q_d = d\,p_d/c$, where p_d is the Poisson degree distribution. Here, apart from the block memberships, there is no correlation in an SBM. We consider the case of two blocks a and b with the same probability $P_a = P_b = 1/2$ and a symmetrical block matrix. To derive the probability p_d, we first consider the probability $\tilde{p}_{d_1,d_2}$ that a node has degree d_1 for connections within the same block and degree d_2 for connections to nodes of the other block. Since both distributions are Poissonian, with Eq. (1) and because the size of each subgraph is just $N/2$, such that the expected number of neighbors in each subgraph is $c_{\text{intra}}/2$ and $c_{\text{inter}}/2$, respectively, we obtain

$$\tilde{p}_{d_1,d_2} = e^{-c_{\text{intra}}/2}\frac{(c_{\text{intra}}/2)^{d_1}}{d_1!}e^{-c_{\text{inter}}/2}\frac{(c_{\text{inter}}/2)^{d_2}}{d_2!}. \tag{2}$$

The probability of the total degree d we obtain by summing over all possible combinations which sum up to d:

$$\begin{aligned} p_d &= \sum_{k=0}^{d} \tilde{p}_{k,d-k} \\ &= e^{-c_{\text{intra}}/2-c_{\text{inter}}/2}\sum_{k=0}^{d}\frac{(c_{\text{intra}}/2)^k}{k!}\frac{(c_{\text{inter}}/2)^{d-k}}{(d-k)!} \\ &= e^{-c_{\text{intra}}/2-c_{\text{inter}}/2}\frac{1}{d!}\sum_{k=0}^{d}\binom{d}{k}(c_{\text{intra}}/2)^k(c_{\text{inter}}/2)^{d-k} \\ &= e^{-c_{\text{intra}}/2-c_{\text{inter}}/2}\frac{(c_{\text{intra}}/2+c_{\text{inter}}/2)^d}{d!}, \end{aligned} \tag{3}$$

where we have used the Binomial formula $(a+b)^d = \sum_{k=0}^{d}\binom{d}{k}a^k b^{d-k}$. Thus we obtain, quite intuitively, the original Poissonian distribution with effective connectivity $c = c_{\text{intra}}/2 + c_{\text{inter}}/2$ being the average connectivity. Hence, for the case of B blocks of equal size, one would therefore still see the standard Poissonian distribution with effective $c = (c_{\text{intra}} + (B-1)c_{\text{inter}})/B$. More general cases are straightforward to obtain.

Therefore, to obtain the percolation threshold, one can proceed as for the standard case, which we recap very briefly for completeness. We look at $1-\pi$, the probability that a node reached by an edge is not connected to the giant component. Under the assumption that small components are tree-like, which is true in this case for the same reason as for the well-known ER case, $1-\pi$ can be determined self consistently: Since the probability for each neighbor to be not connected to the giant component via its other edges is again $1-\pi$. This means $1-\pi = q_1 + (1-\pi)q_2 + (1-\pi)^2 q_3 + \cdots$. Using $q_d = d\,p_d/c$ and inserting Eq. (3) into this expression we get

$$\begin{aligned} 1-\pi &= \sum_{d=1}^{\infty} e^{-c}\frac{c^{d-1}}{(d-1)!}(1-\pi)^{d-1} \\ &= e^{-\pi c}. \end{aligned} \tag{4}$$

From this expression we can derive the percolation threshold, since solutions $\pi > 0$ (besides the trivial solution $\pi = 0$) become possible for $c > 1$, such that the percolation threshold is

$$B = c_{\text{intra}} + (B-1)c_{\text{inter}}. \tag{5}$$

Also, Eq. (4) can be easily solved numerically to estimate the relative size π for arbitrary connectivity parameters. A heat map of the solutions for π is shown in Fig. 1(c) where the transition from a size of $\pi = 0$ to $\pi > 0$ is clearly visible; note the symmetry due to the symmetric dependence on just $c = c_{\text{intra}}/2 + c_{\text{inter}}/2$. For intuition, consider the following three edge cases: If $c_{\text{intra}} = c_{\text{inter}} > 1$, this reduces to the well-known ER case. If $c_{\text{inter}} = 0$ and $c_{\text{intra}} > 2$, each block behaves like an independent ER graph with $c > 1$, such that inside each block giant components of size $O(N)$ form. If $c_{\text{inter}} > 2$ and $c_{\text{intra}} = 0$, a bipartite giant component of size $O(N)$ arises.

B. Large deviations and sampling method

We are interested in the whole probability distributions of the above mentioned observables. This includes additionally to the common events, which are often well characterized by the mean and variance, also the tails of the distribution characterizing extremely rare events. An especially important class of distributions, which is said to obey the *large-deviation principle*, consists of distributions parametrized by N, here the size of the graph, with a probability density function $P_N(S)$ which can be expressed in terms of a *rate function* $\Phi(S)$, such that $P_N(S) = \exp[-N\Phi(S) + o(N)]$ [19]. Thus, $\Phi(S)$ is independent of N, and the leading term in N is characterized by the rate function. If such a rate function Φ exists, it means that the tails of the distribution decay exponentially in N and Φ governs how fast exactly the tails of the distribution decay. If it has a single minimum and is twice differentiable, typical events can be approximated as Gaussian distributed for large N [30]. Clearly, this principle is not valid for every distribution, since they could decay slower or faster than exponential, or exhibit singular behavior.

Therefore, we want to study whether a large-deviation principle holds for the distribution of the size of the largest connected component for this simple case of SBM. Since the tool of our study is the computer simulation [31], we can treat only realizations of finite size N, such that we can only obtain the *empirical rate function* $\Phi_N(S) = -\frac{1}{N}\ln[P_N(S)]$ for multiple sizes N. If we observe that the empirical rate functions for different sizes converge to a limit shape, we assume that this limit shape is the actual rate function and that the large-deviation principle is valid here.

The main idea in obtaining the empirical rate functions, which include information for extremely rare events, is to perform a suitably tailored Markov chain Monte Carlo simulation in the space of random graphs. Thus, the graphs are not sampled independently, but it allows one to obtain data for the extremely rare and atypical events. In the next section we will see that the distributions of the size of the largest connected component of the SBM often have a pronounced multipeak structure. This led us to use the Wang-Landau (WL) method [32,33], which is especially suited to overcome valleys in the distribution (or energy landscape). Such valleys turned out to be problematic for other methods employed previously by the authors [21,22,24]. These valleys in the distribution were the main hindrance for larger system sizes in previous studies of the ER ensemble [21,22]. So, using WL sampling, we could improve considerably on the size of the studied ER graphs. Also, since the multipeak structure is even more pronounced for the SBM, the WL method is the enabling factor for this study.

To sketch the idea of the WL method, consider first that an estimate $g(S)$ of the actual distribution, which we are searching for, was known in the beginning of the simulation. Then one could construct a Markov chain of random graphs G using the Metropolis-Hastings algorithm with an acceptance probability to change from graph G to G' of $p_{\text{acc}}(G \to G') = \min\{1, \frac{g(S)}{g(S')}\}$ depending on the observables $S = S(G)$ and $S' = S(G')$ of interest. If the estimate is very close to the actual distribution, a histogram $H(S)$ of the values encountered during this Markov chain would be very flat, i.e., all bins would have about the same number of entries. We can then use the deviations from flatness to improve our estimate $P(S) \approx g(S)H(S)/\langle H\rangle$ [34], where $\langle H\rangle$ is the mean count of all bins. This procedure is called *entropic sampling* [35], fulfills detailed balance, and will therefore converge to the correct searched-for distribution. The drawback is that it may converge very slowly depending on the quality of the initial guess $g(S)$.

The ingenious idea of the WL method is to get an estimate for $g(S)$ by using the flatness of an auxiliary histogram as a criterion to change $g(S)$ during the evolution of the Markov chain. Therefore every time an energy S^* is visited, the estimate is updated $g(S^*) \mapsto f \cdot g(S^*)$ using the *refinement* factor f, which is usually initialized as $f = \exp(1)$ and reduced as soon as the histogram fulfills some flatness criterion [33] or some set amount of change attempts was performed [36]. First, we need to define the *Monte Carlo time* t in sweeps, i.e., N change proposals. Here we use the schedule of Refs. [36,37], where the logarithm of the refinement factor $\ln f$ first decreases exponentially $\ln f \mapsto \ln f/2$ every time each bin of the auxiliary histogram was visited at least once, and this is checked every 1000 t. If this criterion is fulfilled, the auxiliary histogram is reset. As soon as $\ln f < 1/t$, $\ln f$ is decreased as a power law after every sweep $\ln f \mapsto 1/t$. The algorithm stops as soon as $\ln f$ reaches a defined value of $\ln f_{\text{final}}$, chosen as 10^{-5} in this study.

Since this means that p_{acc} is time dependent, detailed balance does not hold and systematic errors might be introduced. Therefore we subsequently perform entropic sampling, which is theoretically sound, to remove any systematical error.

This technique depends on the choice of the histogram, such that one has to be careful when choosing the binning. Here, fortunately, we have a discrete problem, since the size of the largest component is the number of nodes, such that we can choose a perfect binning of N uniform bins, which does not introduce any discretization error.

One can parallelize the WL method by performing it independently in multiple *windows* and matching the resulting estimates using overlaps of the windows. We use up to 14 overlapping windows for this purpose. For each window 3×10^5 sweeps are simulated: 10^5 sweeps to reach $\ln f_{\text{final}} = 10^{-5}$ and 2×10^5 sweeps of entropic sampling. Per window this takes for the largest simulated sizes around 40 hr ($N = 2048$ for S) to 70 hr ($N = 1024$ for S_2) on relatively modern hardware (Intel Xeon E5-2650 v4), fluctuating by approximately 50% in both directions depending on the connectivity of the graphs and the acceptance rate.

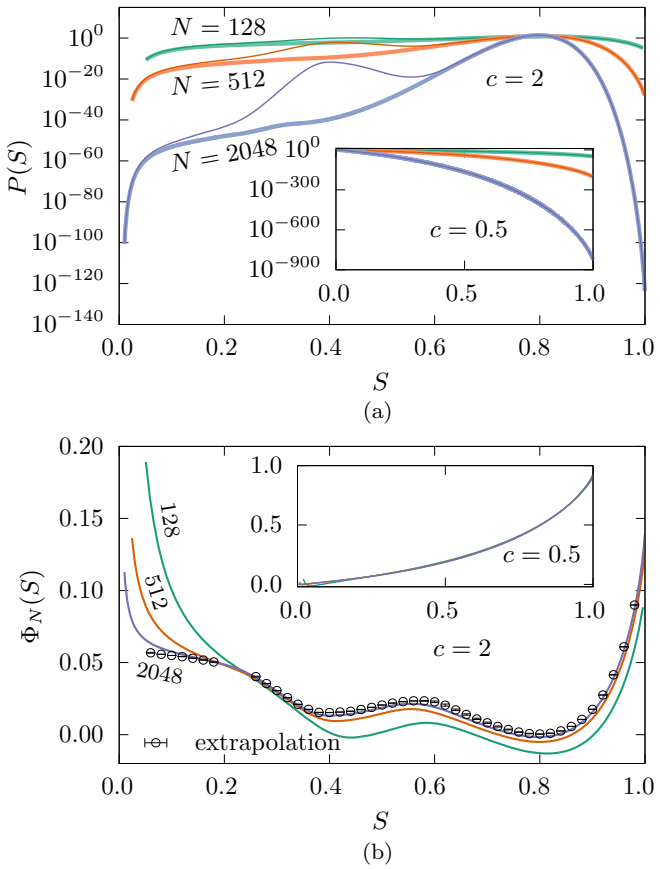

FIG. 2. (a) Distributions in logarithmic scale of the relative size of the largest component S for ER (thick lines) and SBM (thin lines) for different graph sizes over (almost) the full support. (b) Corresponding empirical rate functions. A fast convergence to a limiting shape is observable. Main plots show $c = 2$, respectively, $c_{\text{inter}} = 0.1$, $c_{\text{intra}} = 3.9$, and the insets show $c = 0.5$, respectively, $c_{\text{inter}} = 0.1$, $c_{\text{intra}} = 0.9$. In the main plot data for finite sizes are extrapolated, and the extrapolation as the estimate of the rate function is shown with symbols.

One of the most crucial aspects of any Markov chain Monte Carlo simulation is the choice of the *change move* to generate new trial graphs for the chain. Beyond the block membership all edges are independent in the SBM, just like the ER. Therefore we create a new trial graph G' by selecting a node i in the current graph G at random, removing all of its edges and deciding for each other node j randomly whether edge $\{i, j\}$ is inserted with the appropriate probability depending on their block memberships. This change move is ergodic and works reasonably well.

III. RESULTS

In Fig. 2(a) we show for some system sizes the resulting distributions for the cases of low connectivity $c = 0.5$ ($c_{\text{inter}} = 0.1$, $c_{\text{intra}} = 0.9$) in the nonpercolating regime (inset) and of higher connectivity $c = 2$ ($c_{\text{inter}} = 0.1$, $c_{\text{intra}} = 3.9$) in the percolating regime. Here, and in the following, data for the SBM are visualized with thin, dark lines and for the ER with thick, lighter lines. If only one line is visible, the data of both ensembles coincide. Different shades mark different sizes N. Here we see that the SBM exhibits in the $c = 2$ case a strongly different behavior than the ER. This manifests for the largest system size in structures of the probability density function (pdf) below probabilities of 10^{-15} and would therefore be undetectable with conventional methods.

To foster intuition about the relation of the rate function with the probability density, we show the empirical rate functions $\Phi_N(S)$ obtained from the probability density functions in Fig. 2(b). For clarity only data for the SBM are visualized. Here we observe that the smallest size $N = 128$ shows nonmonotonous and rather large deviations, but $\Phi_{512}(S)$ and $\Phi_{2048}(S)$ are already very close over large parts of the support. The extrapolation is performed pointwise, at equal values of S for multiple system sizes. We assume a power law with offset $\Phi_N(S) = \Phi_\infty(S) + aN^b$ for the extrapolation, which was already used before for this task [38] and fits our data quite well. Since different system sizes have a different number of bins, we interpolate Φ_N linearly for convenience, which should introduce only negligible error due to the dense bins. Note that the minimum of the extrapolated estimate of the rate function is at $\Phi_\infty(S_{\min}) \approx 0$. The region around $S \approx 0.2$ could not be extrapolated well due to the crossing of different system sizes. For $c \leqslant 1$ we do not perform this extrapolation since all system sizes already yield almost equal rate function estimates. For clarity, we will not show the extrapolation in following figures, since it always is very close to our data for the largest system size; i.e., $N = 2048$ seems to be large enough that the empirical rate function Φ_N is sufficiently close to our extrapolation Φ_∞, which we handle as an estimate of the actual asymptotic rate function.

Note that the nature of numerical studies is that we have to rely on the problem being well behaved. While we have sizes which are large enough to show a convergence to some value, it is theoretically possible that this value is not the asymptotic limit but only a "plateau" and that the behavior changes for even larger system sizes. Since we know the exact behavior of the rate function for the ER and we previously observed a very nice and quick convergence of measurements in finite systems with similar system sizes to the exact asymptotic form [21], we argue that the simple SBM under scrutiny should be very similarly well behaved.

Therefore, we conjecture that the large-deviation principle holds for S and S_2 of this variant of the SBM and is approximated by the empirical rate functions for the largest measured sizes shown in the following figures. Since we observe this above and below the percolation threshold, and because we do not expect any other dramatic changes in the structure, this should hold for all parameters c_{inter} and c_{intra}.

As a side remark, consider a finite temperature T ensemble, where the occurrence of realizations was weighted with a Boltzmann weight $e^{-S/T}$, treating the size of their largest connected component S as energy, studied, e.g., in Ref. [21]. The two-peak structure corresponds to two transitions of first order at two distinct temperatures T. At one transition, two large coexisting components appear, at the other transition, one single biggest component emerges; see the discussion below.

In Fig. 3 the empirical rate functions and distributions for finite system sizes N are shown for different parameter sets, especially below and above the percolation threshold.

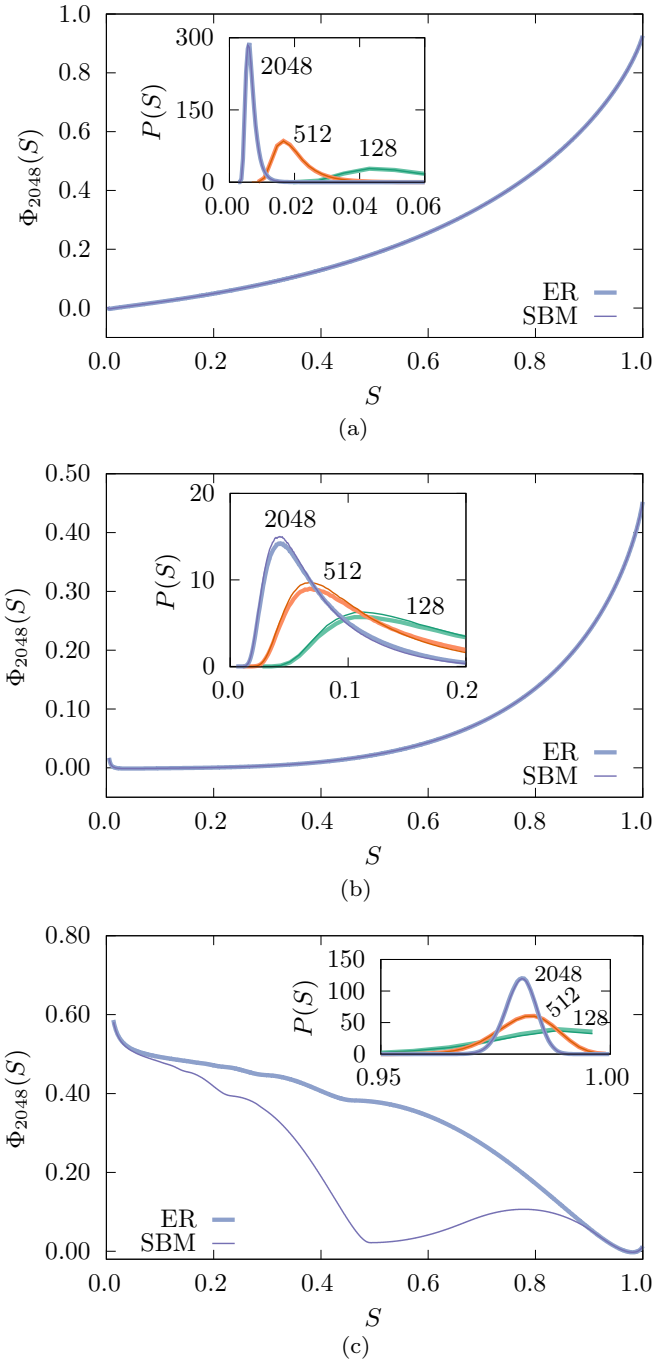

FIG. 3. The main plots show the empirical rate functions $\Phi_{2048}(S)$ for different connectivities of both the ER (thick lines) and SBM (thin lines) ensemble, which coincide often: (a) $c = 0.5, c_{\text{inter}} = 0.1, c_{\text{intra}} = 0.9$; (b) $c = 1, c_{\text{inter}} = 0.1, c_{\text{intra}} = 1.9$; (c) $c = 4, c_{\text{inter}} = 0.1, c_{\text{intra}} = 7.9$. Their insets show the corresponding probability density functions for different sizes. Note that the normalization for $P(S)$ in the insets is such that the area under the curve is unity (thus for better comparison we show densities although the support is discrete).

The peculiar two-peak structure of the rate function of the SBM above the percolation threshold in Figs. 2(b) and 3(c) can be explained rather simply. The left peak consists of realizations, where two separate but large connected components

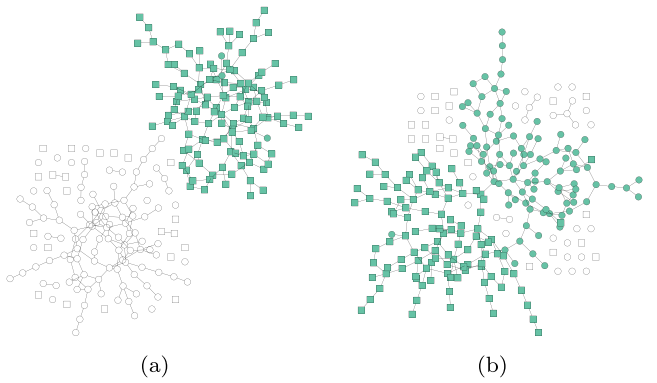

FIG. 4. Examples for SBM realizations at $c_{\text{inter}} = 0.1, c_{\text{intra}} = 3.9, N = 256$. The two blocks are visualized as nodes of different shapes; the largest connected component consists of colored nodes. These are realizations originating (a) from the left peak and (b) from the right peak.

exist, one in each block. For $c = 4$ this peak is at $S \approx 0.5$ since almost all nodes within one block are connected. For $c = 2$ the connected components within a block are smaller, such that we observe this peak at $S \approx 0.4$. Figure 4(a) shows an example realization of this type. Since it is exponentially unlikely that no interblock edge exists, the occurrence of this structure is exponentially suppressed, resulting in a value of the rate function at this position larger than zero, and is subsequently not visible in the distributions for moderately large systems. Due to this mechanism, we expect that the left peak will be suppressed at large sizes for all choices of the connectivity parameters in the two-block planted partition, as long as $c_{\text{inter}} > 0$. Also beyond the planted partition, i.e., with different connectivities between different blocks, we expect that the second peak is suppressed for large sizes as long as all interblock connectivities are nonzero. The main peak at $S \approx 0.8$, respectively, $S \approx 1$, on the other hand, contains the instances in which the connected components inside of the blocks are connected with each other, as visualized in Fig. 4(b).

Also note that the same two-peak structure exists in the distribution of the largest biconnected component visualized in Fig. 5(b). The main mechanism causing the second, smaller peak is the same as for the largest connected component: due to the low value of c_{inter} it is more probable that two biconnected components within each of the blocks are not connected by two interblock edges. Since a split in two large biconnected components is more probable than the split into two large connected components, we can observe the second peak for the smallest size $N = 256$ even in the linear scale depiction of $P(S_2)$ shown in the inset of Fig. 5(b) but not in $P(S)$ in Fig. 5(a). The most probable size of the largest bicomponent is naturally smaller than the most probable size of the largest connected component at slightly larger than 0.4, and correspondingly the maximum of the second peak is at half this value, slightly larger than 0.2. The smaller magnitude of the second peak is caused at least partially by a much higher probability that there are no large biconnected components inside the blocks, which leads generally to a broader distribution. Note that the vertical axis spans a far larger range in the

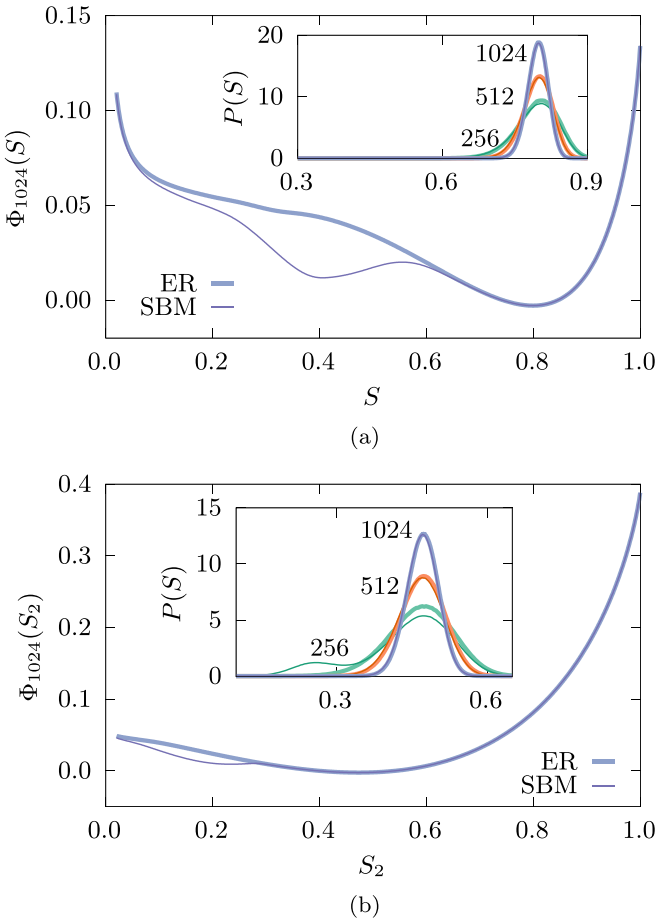

FIG. 5. The main plots show the empirical rate functions (a) $\Phi_{1024}(S)$ of the relative size of the largest connected component and (b) of the largest biconnected component $\Phi_{1024}(S_2)$ of both the ER (thick lines) and SBM (thin lines) ensemble for $c = 2$, $c_{\text{inter}} = 0.1$, $c_{\text{intra}} = 3.9$. Note that the right tail reaches down to $\Phi_{1024} \approx 3.8$, i.e., probabilities of $P(S_2) \approx e^{-1024 \cdot 3.8} \approx 10^{-1690}$. Their insets show the corresponding probability density for multiple sizes. Note that the normalization is such that the area under the curve is unity (thus for better comparison we show densities although the support is discrete).

diagram for the biconnected component, which exaggerates the difference.

The most striking property of the distributions $P(S)$ for different values of the connectivity is the surprising way they differ between ER and SBM. We are able to assess these differences since our large-deviation sampling approach gives us access to the tails: Below the percolation threshold in Fig. 3(a) the two distributions are visually indistinguishable, in the peak (shown in the inset) as well as in the tails (shown in the main plot). At the threshold in Fig. 3(b), one can see significant deviations in the peak, but the tails are again indistinguishable. Surprisingly, above the threshold in Figs. 3(c) and 5(a) the peaks of the distributions are again visually indistinguishable, but the tails show qualitatively different behavior with a far more pronounced second peak for the SBM case.

For the assortative case $c_{\text{inter}} \ll c_{\text{intra}}$, which we study here mainly, it is plausible that the size of the largest connected component should differ the strongest near the threshold, since here, close to the percolation threshold is the only parameter regime where the interblock edges do matter at all. Far below the threshold, the SBM realization consists of trees with members from only one block, but since our observable S does not account for the block memberships, this is indistinguishable from ER. Far above the threshold the blocks are connected components, and as long as there are any interblock edges, the largest connected component will typically include almost the whole graph, the same as the ER case. Therefore, in the $c_{\text{inter}} \ll c_{\text{intra}}$ case, only around the threshold can the peaks of the distributions differ at all. Note that these arguments are valid also for higher number of blocks B and for more general connectivities between the blocks, if they are assortative enough.

In the disassortative case $c_{\text{inter}} \gg c_{\text{intra}}$, we found that the distribution $P(S)$ is generally indistinguishable within our high precision numerical data from the ER case, even in the far tails (not shown). This is not surprising since the mechanism of two unconnected clusters leading to the differences in the assortative cases cannot occur in (almost) bipartite graphs. We therefore conclude that the size of the largest connected component does differ at most very weakly between this simple SBM and ER. For the size of the largest biconnected component the results are qualitatively the same and the same arguments apply.

The balanced case $c_{\text{inter}} = c_{\text{intra}}$ is equal to the ER ensemble and therefore trivially does not differ.

As a more formal method to judge whether or not the peak regions of ER and SBM are indistinguishable, we use the *Epps-Singleton* test [39,40], which is designed to estimate the probability p_{ES} that two samples from discrete distributions originate from the same distribution. In order to smooth these results, we average the p values obtained for 100 independent pairs of samples, each containing 10^4 independent measurements of the largest connected components. Note that such a procedure diminishes greatly the power of the test, and one would usually choose something like Fisher's method to combine independent p values. However, since for our purpose, we want to explore a trend for large systems instead of analyzing specific cases, the extremely conservative $\overline{p}_{\text{ES}}$ should still be a useful and especially smooth metric.

We used this procedure to estimate $\overline{p}_{\text{ES}}$ for multiple values of c, and in the case of the SBM, we fixed $c_{\text{inter}} = 0.1$ and varied $c_{\text{intra}} = 2c - c_{\text{inter}}$. Figure 6(a) shows the result of this analysis. Very low values of $\overline{p}_{\text{ES}}$ signal that the two samples originate from different distributions, i.e., the distributions are distinguishable. High values, say, above the typical 5% mark, signal that we cannot exclude the possibility that the two samples originate from the same distribution. The exact p values should not be taken too seriously, since they greatly depend on the number of samples and are overestimated due to our smoothing.

In accordance with our visual interpretation above, the distributions for connectivities around the transition at $c_c = 1$, are distinguishable using this statistical test. In particular, the range where the distributions are distinguishable shrinks with increasing system size. Also note that using different statistical tests, like Kolmogorov-Smirnov [41] or Anderson-Darling [40,42], leads to similar results (not shown).

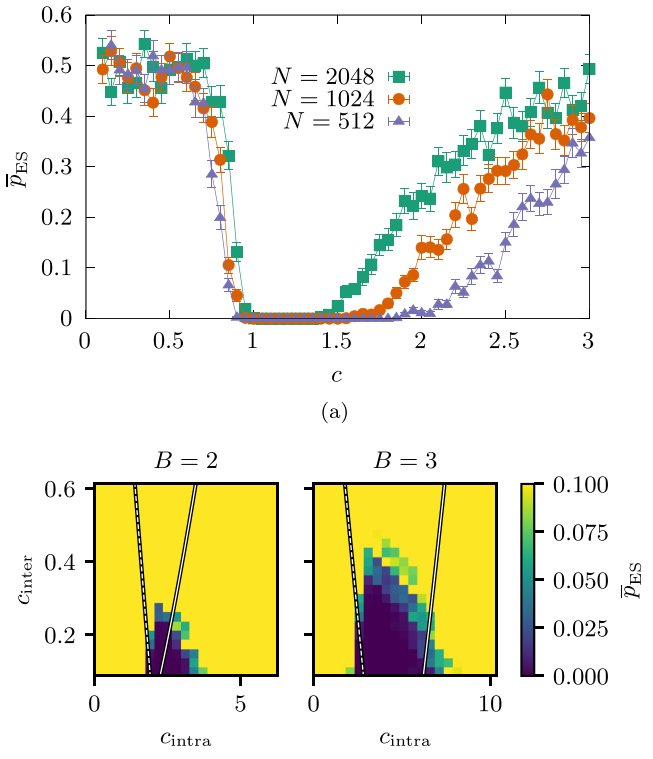

FIG. 6. (a) Average confidence of 100 Epps-Singleton tests that two samples of S (with 10^4 measurements each), one obtained from the ER and the other from the SBM with $c_{\mathrm{inter}} = 0.1$, $c_{\mathrm{intra}} = 2c - c_{\mathrm{inter}}$, originate from the same distribution. Low values mean that we can surely distinguish the two ensembles, high values mean that we cannot. (b) Heat maps of the same Epps-Singleton tests at $N = 1024$ for more combinations c_{inter} and c_{intra} for both $B = 2$ and $B = 3$. The solid line marks the corresponding community detectability thresholds $|c_{\mathrm{inter}} - c_{\mathrm{intra}}| = B\sqrt{c}$ [10], and the dashed line shows the percolation threshold $B = c_{\mathrm{intra}} + (B-1)c_{\mathrm{inter}}$. For every pixel 2×10^6 independent measurements were used, smoothed using the method explained in this paper. Note that the other branch of the threshold for disassortative SBM is not visible due to the short vertical axis. However, that region shows solid nondistinguishability.

In Fig. 6(b) we scrutinize for which values of the inter- and intrablock connectivities, the main part of the distribution allows us to distinguish ER from SBM, i.e., recognize that there is some community structure in the network. There the parameter space which can be distinguished is visualized with dark colors. Note that Fig. 6(a) suggests that this dark region should shrink from the right for larger sizes N. With this in mind, this figure confirms our observations from before that ER and SBM differ the most around the percolation threshold, which is marked by a dashed line. It also shows that this distinguishability is given only at low values of c. This is expected since using the size of the connected component one cannot be able to distinguish ER from SBM when the connectivities are high enough that the giant component almost always contains every single node.

An interesting question coming to mind is whether this behavior is also related to the transition from detectable community structure to undetectable community structure $|c_{\mathrm{inter}} - c_{\mathrm{intra}}| > B\sqrt{c}$ [10]. Therefore, we also marked the community detection threshold with a solid line, i.e., realizations right of the solid line can be used to reconstruct the block membership of the nodes, e.g., by the sophisticated methods of Ref. [10]. For $B = 2$ the distinguishability using $P(S)$ and the community detection threshold are quite close to the percolation threshold at small values of c_{inter}. The $B = 3$ case on the right of Fig. 6(b), especially considering that the distinguishable region (darkly marked) should shrink from the right for larger values of N [cf. Fig. 6(a)], shows that the community detection threshold is located at considerably higher values of c_{intra} than the sparsest realizations we can distinguish. Thus it appears that $P(S)$ can distinguish community structure (or some structure absent in ER) in realizations with low connectivity where the actual detection of communities is impossible. Interestingly, it seems that realizations whose communities can be reconstructed (right of the solid line) cannot be distinguished by the typical behavior of S. Here systematic, numerically, and also demanding studies for various larger values of B could possibly be of interest.

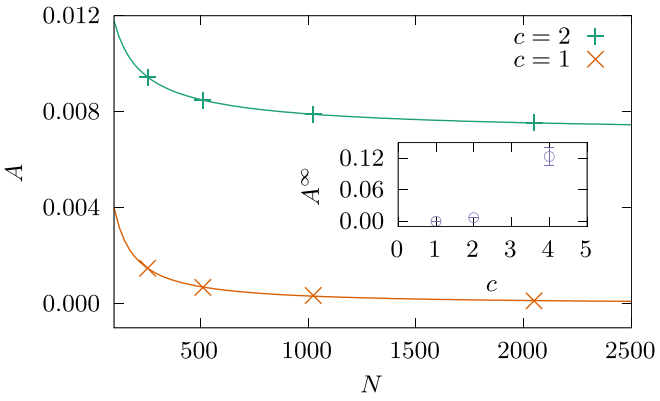

FIG. 7. Area A between empirical rate functions of ER and SBM extrapolated using a fit of the form $A(N) = A^\infty + aN^b$. The offset A^∞ at a connectivity of $c = 1$ is compatible with zero at $A_1^\infty = 3 \times 10^{-5} \pm 7 \times 10^{-5}$; i.e., the rate functions of ER and SBM appear to become indistinguishable. For a connectivity of $c = 2$ we obtain an offset $A_2^\infty = 0.007(1)$, and for $c = 4$ (not shown to preserve detail of the plot) $A_2^\infty = 0.12(2)$, i.e., the rate functions of ER and SBM appear to stay distinct. This behavior $A^\infty(c)$ is shown in the inset.

Note that the behavior of the tail, which obviously differs for large c, e.g., in Fig. 5(a), is immaterial for this statistical test. For analysis of the tail behavior, we will introduce the area between the empirical rate functions

$$A = \int_0^1 dS \left| \Phi_N^{\mathrm{SBM}}(S) - \Phi_N^{\mathrm{ER}}(S) \right| \tag{6}$$

as a measure of distinguishability.

In Fig. 7 the area A between the empirical rate functions is shown for multiple system sizes N at connectivities of $c = 1$ and $c = 2$. To estimate whether the differences between the empirical rate functions are finite-size effects, or persist in the infinite limit of the rate function, we extrapolate the area to infinite systems using the ansatz $A(N) = A^\infty + aN^b$, which fits quite well to the data. We find that for $c = 1$ the area A^∞ and therefore the difference vanishes within error bars in

the limit of infinite systems. The rate functions for $c = 2$, on the other hand, stay clearly distinct between ER and SBM. This distinctiveness increases for larger values of c (and fixed c_{inter}), which is shown in the inset of Fig. 7. This means, that given the far tails of the distribution, we can determine the existence of blocks not only at and closely beyond the percolation threshold but also far above it.

To gather insight how configurations with especially large or especially small biconnected components look, we consider the correlations between the size of the connected and biconnected components. Using Bayes' theorem, we can estimate parts of the joint probability density $P(S, S_2)$ from a single WL simulation. Therefore we save during the entropic sampling phase the pairs (S, S_2) of observables we encountered in the Markov chain and estimate the conditioned probability $P(S|S_2)$ from them. This can be used with the result of the WL simulation, $P(S_2)$, to obtain a part of the joint probability density $P(S, S_2) = P(S|S_2)P(S_2)$. Note, that with a much higher numerical effort it would also be possible to obtain the full joint probability density using a two-dimensional WL variant [33]. In Fig. 8 parts of the joint probability density are shown. One notices that the correlations for the SBM above the percolation threshold show a surprising multimodal structure, which is marked and labeled in black. However, we will see that this is actually plausible, and we will discuss the structure of the realizations inside each of the three regions.

In the region labeled D (divided), which is not present in the ER, we see that inside of the highly connected blocks of the SBM, which are not yet connected to each other, biconnected components exist [cf. Fig. 8(b)]. The group of realizations, labeled O (one connection), indicates that there is a considerable amount of realizations where already giant connected components spanning both blocks exist ($S > 0.5$), but the largest biconnected component is still restricted to one of the blocks with $S_2 < 0.3$. These are mostly realizations where the connected components inside each block are connected by a single edge (or multiple edges arriving at a single node) [cf. Fig. 8(c)]. Part of this region is also, though less often, configurations with a biconnected component inside one block connected to multiple tree-like structures consisting of nodes of the other block. Interestingly both types of configuration coexist in our simulations. Since both of these groups rely on the high intrablock connectivity, they do not occur in the ER ensemble.

In the region labeled M (multiple connections) of Fig. 8, which also occurs for the ER, one sees perfect correlation between the size of the two types of components. The larger the biconnected component should be, the larger the connected component has to be. Here SBM and ER match very nicely. An example realization is shown in Fig. 8(d). The region O does not smoothly go over into region M, and both coexist for the same size of the giant component, such that both marginal probabilities show the two-peak structure we observed before.

We will use the end of the results section to make some educated guesses about the behavior of planted partitions with more blocks B. We would expect that basically the same patterns should occur as in the previous case, where the connected and biconnected components span different combinations of clusters. Since each cluster will have roughly a size of order $1/B$ and the biconnected component may never span more clusters than the connected component, there are $B(B+1)/2$ possible combinations, which should all manifest as a local maximum in the joint probability. Using the same arguments as in the $B = 2$, we would expect that every local maximum, except the one where all clusters are part of the biconnected component, corresponding to region M, should be exponentially suppressed.

For the marginal probabilities $P(S)$ [and $P(S_2)$], this suggests that there should be B peaks, corresponding to one to B (doubly) connected clusters. Like we observed for the $B = 2$ case, all but the one corresponding to region M, where all clusters are (doubly) connected should be exponentially suppressed. However, the width of the peaks in our data suggests that those B peaks might not be clearly distinguishable, especially for larger values of B.

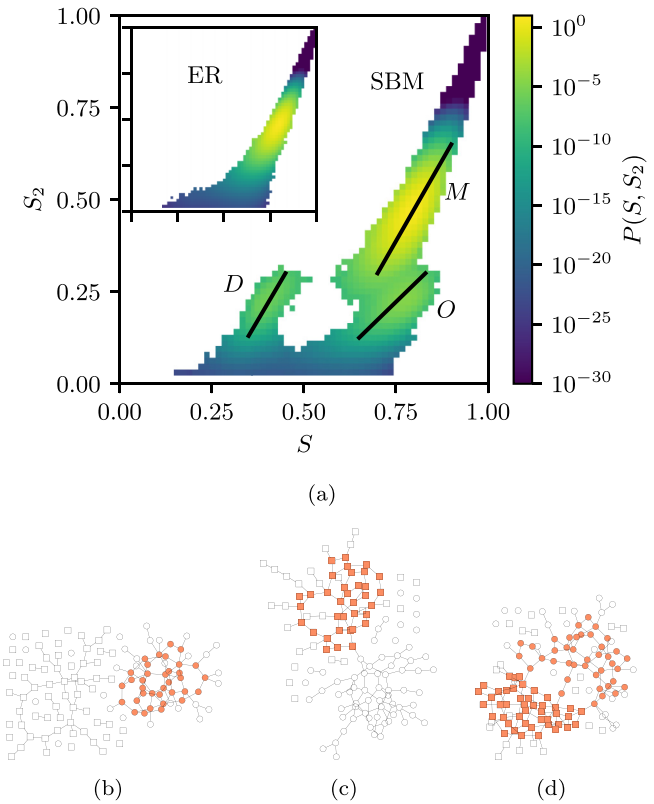

FIG. 8. Comparison of parts of the joint probability density $P(S, S_2)$. SBM ($c_{\text{inter}} = 0.1$, $c_{\text{intra}} = 3.9$) is shown in the main plot, ER ($c = 2$) in the inset. The data are for $N = 1024$ and collected during the entropic sampling of a WL simulation of the size of the largest biconnected component. The color scale is compressed to increase the visibility of the central structures despite the very large range of probabilities to be visualized. For white points no corresponding samples (S, S_2) were encountered during the entropic sampling. Here the space is discretized into 64×64 bins. The black lines are guides to the eye and indicate the regions of highest probabilities. The corresponding labels are referenced in the text. The three classes identified in (a) are illustrated with examples of size $N = 128$ with highlighted largest biconnected component: (b) example of region D, (c) example of region O, and (d) example of region M.

IV. CONCLUSIONS

Here we studied the distributions of the relative size of the largest connected S and biconnected components S_2 for the stochastic block model with two blocks and strong intrablock connectivity. By using sophisticated large-deviation algorithms, we are able to study the distributions down to probabilities as small as 10^{-800} or below, which gives us access to (almost) the full distributions. Due to the fast convergence to a limiting shape of the empirical rate functions we conjecture that the large-deviation principle holds for these distributions. Further, we showed where there are similarities to the Erdős-Rényi graph ensemble and for which parameters there are differences in different parts of their distributions. Especially, we show large qualitative differences in the tails of extremely rare events, where the peak regions are indistinguishable. These differences seem to be correlated with the threshold of the percolation transition. By analyzing the correlations between the largest connected and largest biconnected component, which was also possible in the regime of rare events, we could identify three regimes of behavior (cf. Fig. 8). This case study led to insight into the structure of some rare configurations of the SBM, which led to some educated guesses which behavior is expected in other regions of the parameter space.

In general, our study shows that by analyzing the tails of probability distributions for random graphs, differences between ensembles can be found which are not detectable by standard simple sampling simulations. In the studied case, this even works although the percolation transition behavior can be directly mapped between the two ensembles using an effective connectivity. It is rather surprising that there are subtle differences in the extremely simple observables we studied between two similar models, which are inaccessible with a conventional analysis. Thus, large-deviation simulations offer access to otherwise hidden properties of networks and to correlations between network quantities. Here we demonstrated that they can be used in an exploratory way to explore even models with multiple free parameters (N, c_{inter}, c_{intra}, B) and guide the search for parameter ranges with distinct behavior by allowing the direct examination of fringe cases. Due to the existence of many different network ensembles, network processes, and measurable quantities, many new results will likely emerge from applying this and similar approaches to gain deep insight into the properties of networks. Of particular interest could be to study modified ER models with varying degree of (general) (dis-) assortativity by introducing degree correlations, to see how this shows up in the distributions $P(S)$ and $P(S_2)$, and to compare with the SBM, which is a special case. A thorough (but necessarily numerically demanding) study could indicate what is a minimal degree of (dis-) assortativity to make a difference to the ER significant.

ACKNOWLEDGMENTS

The authors thank Stefan Adolf for performing preliminary studies on this topic. We thank Tiago Peixoto for interesting discussions. Also, the authors want to thank the two anonymous referees for their exceptionally detailed and useful reviews. H.S. acknowledges financial support of the grant HA 3169/8-1 by the German Science Foundation (DFG) and the OpLaDyn grant obtained in the fourth round of the Trans-Atlantic Platform Digging into Data Challenge (2016-147 ANR OPLADYN TAP-DD2016). The simulations were performed at the HPC Cluster CARL, located at the University of Oldenburg (Germany) and funded by the DFG through its Major Research Instrumentation Programme (INST 184/108-1 FUGG) and the Ministry of Science and Culture (MWK) of the Lower Saxony State.

[1] P. W. Holland, K. B. Laskey, and S. Leinhardt, Stochastic blockmodels: First steps, Social Netw. **5**, 109 (1983).

[2] P. Erdős and A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hungar. Acad. Sci. **5**, 17 (1960).

[3] M. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).

[4] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications, Phys. Rev. E **84**, 066106 (2011).

[5] B. Karrer and M. E. J. Newman, Stochastic blockmodels and community structure in networks, Phys. Rev. E **83**, 016107 (2011).

[6] T. P. Peixoto, Entropy of stochastic blockmodel ensembles, Phys. Rev. E **85**, 056122 (2012).

[7] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, Spectral redemption in clustering sparse networks, Proc. Natl. Acad. Sci. U. S. A. **110**, 20935 (2013).

[8] T. P. Peixoto, Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models, Phys. Rev. E **89**, 012804 (2014).

[9] T. P. Peixoto, Nonparametric Bayesian inference of the microcanonical stochastic block model, Phys. Rev. E **95**, 012317 (2017).

[10] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Inference and Phase Transitions in the Detection of Modules in Sparse Networks, Phys. Rev. Lett. **107**, 065701 (2011).

[11] R. R. Nadakuditi and M. E. J. Newman, Graph Spectra and the Detectability of Community Structure in Networks, Phys. Rev. Lett. **108**, 188701 (2012).

[12] R. K. Darst, D. R. Reichman, P. Ronhovde, and Z. Nussinov, Algorithm independent bounds on community detection problems and associated transitions in stochastic block model graphs, J. Complex Netw. **3**, 333 (2014).

[13] R. Albert, H. Jeong, and A.-L. Barabási, Error and attack tolerance of complex networks, Nature (London) **406**, 378 (2000).

[14] C. Norrenbrock, O. Melchert, and A. K. Hartmann, Fragmentation properties of two-dimensional proximity graphs considering random failures and targeted attacks, Phys. Rev. E **94**, 062125 (2016).

[15] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Network Robustness and Fragility: Percolation on Random Graphs, Phys. Rev. Lett. **85**, 5468 (2000).

[16] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, Resilience of the Internet to Random Breakdowns, Phys. Rev. Lett. **85**, 4626 (2000).

[17] T. Dewenter and A. K. Hartmann, Large-deviation properties of resilience of power grids, New J. Phys. **17**, 015005 (2015).

[18] M. E. J. Newman and G. Ghoshal, Bicomponents and the Robustness of Networks to Failure, Phys. Rev. Lett. **100**, 138701 (2008).

[19] H. Touchette, The large deviation approach to statistical mechanics, Phys. Rep. **478**, 1 (2009).

[20] M. Biskup, L. Chayes, and S. A. Smith, Large-deviations/thermodynamic approach to percolation on the complete graph, Random Struct. Alg. **31**, 354 (2007).

[21] A. K. Hartmann, Large-deviation properties of largest component for random graphs, Eur. Phys. J. B **84**, 627 (2011).

[22] H. Schawe and A. K. Hartmann, Large-deviation properties of the largest biconnected component for random graphs, Eur. Phys. J. B **92**, 73 (2019).

[23] A. K. Hartmann and M. Mézard, Distribution of diameters for Erdős-Rényi random graphs, Phys. Rev. E **97**, 032128 (2018).

[24] A. K. Hartmann, Large-deviation properties of the largest 2-core component for random graphs, Eur. Phys. J.: Spec. Top. **226**, 567 (2017).

[25] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).

[26] J. Hopcroft and R. Tarjan, Algorithm 447: Efficient algorithms for graph manipulation, Commun. ACM **16**, 372 (1973).

[27] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms* (MIT Press, Cambridge, MA, 2009).

[28] B. Dezső, A. Jüttner, and P. Kovács, Lemon—An open source C + + graph template library, Electron. Notes Theor. Comput. Sci. **264**, 23 (2011), proceedings of the Second Workshop on Generative Technologies (WGT) 2010.

[29] A. K. Hartmann and M. Weigt, *Phase Transitions in Combinatorial Optimization Problems* (Wiley Online Library, Weinheim, 2005).

[30] W. Bryc, A remark on the connection between the large deviation principle and the central limit theorem, Stat. Probab. Lett. **18**, 253 (1993).

[31] A. K. Hartmann, *Big Practical Guide to Computer Simulations* (World Scientific, Singapore, 2015).

[32] F. Wang and D. P. Landau, Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States, Phys. Rev. Lett. **86**, 2050 (2001).

[33] F. Wang and D. P. Landau, Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram, Phys. Rev. E **64**, 056101 (2001).

[34] R. Dickman and A. G. Cunha-Netto, Complete high-precision entropic sampling, Phys. Rev. E **84**, 026701 (2011).

[35] J. Lee, New Monte Carlo Algorithm: Entropic Sampling, Phys. Rev. Lett. **71**, 211 (1993).

[36] R. E. Belardinelli and V. D. Pereyra, Fast algorithm to calculate density of states, Phys. Rev. E **75**, 046701 (2007).

[37] R. E. Belardinelli and V. D. Pereyra, Wang-Landau algorithm: A theoretical analysis of the saturation of the error, J. Chem. Phys. **127**, 184105 (2007).

[38] H. Schawe, A. K. Hartmann, and S. N. Majumdar, Large deviations of convex hulls of self-avoiding random walks, Phys. Rev. E **97**, 062159 (2018).

[39] T. Epps and K. J. Singleton, An omnibus test for the two-sample problem using the empirical characteristic function, J. Stat. Comput. Simul. **26**, 177 (1986).

[40] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in Python, Nat. Methods **17**, 261 (2020).

[41] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 2007).

[42] T. W. Anderson and D. A. Darling, Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes, Ann. Math. Stat. **23**, 193 (1952).