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, France*

[2] *Institut für Physik, Universität Oldenburg, 26111 Oldenburg, Germany*

(Dated: March 10, 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 bicomponent, 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 this 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, near the community detection threshold.

arXiv:2003.03415v1 [physics.soc-ph] 6 Mar 2020

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 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 Erdős-Rényi graph ensemble or configuration model [3]. Indeed in the case of $B = 1$ it simplifies to an Erdős-Rényi 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]. In general, the determination of community structures is algorithmically challenging. This motivated our study, because we are interested in whether the detectability or non-detectibility is related to the simpler connectedness properties of the system, which we will introduce next. Here, as anticipation of our results, indeed a partial relationship is visible, but one has to study the corresponding probability distributions in the extreme-low probability regime to observe it.

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.

Furthermore, since networks consist of many nodes, which often symbolize entities that can fail or vanish, the robustness against this kind of events is of relevance. A common idea [11–15] 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 in this work 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 [16].

We scrutinize these properties in very high detail, i.e., we do not only look at their mean size, but we obtain their probability distributions over practically the whole support, especially including very rare events with a probability of less than $p = 10^{-800}$. In large-deviation theory [17], many probability distributions have a special shape which allows 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* [17] holds for this distribution.

A technical advantage of the studied observables is that their behavior is known for the related ER ensem-

* hendrik.schawe@cyu.fr

† a.hartmann@uol.de

ble, partly analytical [18], partly using simulational techniques [19, 20]. 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 before in [19, 20].

II. MODELS AND METHODS

A *graph* is a tuple $G = (V, E)$ of a set of *nodes* V and a set of *edges* E. Here we will only scrutinize *undirected*, *simple* graphs, i.e., $E \subset V^{(2)} \setminus \{\{u, u\} | u \in V\}$. 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* S 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 S is a reasonable first approximation.

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 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 $\mathcal{O}(|V| + |E|)$ by performing one modified depth first search on a given graph [21–23]. 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 graph (ER) 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 $\mathcal{O}(\log N)$ to a structure with one giant connected component of size $\mathcal{O}(N)$ 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 $\mathcal{O}(N)$ arises [16].

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 *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 $c_{ab} = Np_{ab}$.

Since we want to perform a very in-depth study of an SBM ensemble, we will treat the most simple SBM, which is distinct from ER, i.e., two blocks with the same intra-block connectivity c_{intra} and symmetric inter-block connectivity c_{inter}. Figure 1 shows two examples for different values of c_{inter} and c_{intra}.

The phase transition to a giant connected component happens at $(c_{\text{intra}} + c_{\text{inter}})/2 = 1$ (cf. Fig. 1(c)). 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 with $c > 1$, such that inside of 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* $\Phi(s)$ (with $s = S/N$), such that $P_N(S) = \exp(-N\Phi(S/N) + o(N))$ [17]. Thus, $\Phi(s)$ is independent of N and the leading term in N is

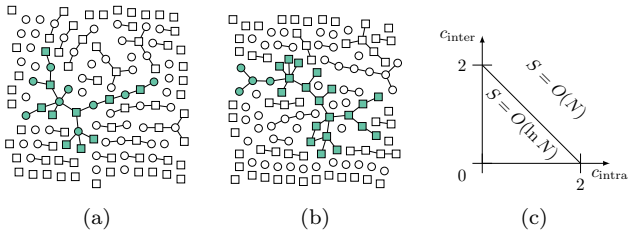


FIG. 1. Left: two example realizations of the SBM with size $N = 128$ with two blocks (shape of nodes) of equal probability $P_a = P_b = 0.5$. The panels show realizations with different connectivities at the percolation threshold (a) $c_{\text{intra}} = 0.1, c_{\text{inter}} = 1.9$ (b) $c_{\text{intra}} = 1.9, c_{\text{inter}} = 0.1$. The largest connected components are visualized with colored symbols. (c) sketch of the phase diagram showing the percolation transition.

charcterized 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 [24].

Since we want to study $P_N(S)$ and the corresponding rate function using computer simulations [25], we can only treat 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 to obtain 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 of the extremely rare and atypical events. In the next chapter we will see, that the distributions of the size of the largest connected component of the SBM often have a pronounced multi-peak structure. This led us to use Wang-Landau's method (WL) [26, 27], 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 [19, 20, 28].

To sketch the idea of WL, 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 $p_{\text{acc}}(G \to G') = \min\left\{1, \frac{g(S)}{g(S')}\right\}$ 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 $P(S) \approx g(S)H(S)/\langle H\rangle$ [29], where $\langle H\rangle$ is the mean count of all bins. This procedure is called *entropic sampling* [30], 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 WL 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 [27] or some set amount of change attempts was performed [31]. Since this means that p_{acc} is time dependent, detailed balance does not hold and systematic errors might be introduced. We use an updating schedule which should avoid error saturation [31, 32] until f reaches a defined value of f_{final}. Subsequently we perform entropic sampling, which is theoretically sound, to remove any systematical error. Here, we use a final refinement factor of $f_{\text{final}} = 10^{-5}$ and up to 10 overlapping windows of ranges of the observable, on which WL is performed independently.

One of the most crucial aspects of any Markov chain Monte Carlo simulation is the choice of *change move* to generate new trial graphs for the chain. Since all edges are independent in the SBM, just like the ER, 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 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 non-percolating regime and of higher connectivity $c = 2$ ($c_{\text{inter}} = 0.1, c_{\text{inter}} = 3.9$) in the percolating regime. Both ensembles ER and SBM are compared. 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.

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 [19]. The two-peak structure corresponds to two transitions of first order at two distinct temperatures T. At one transition, two large coexisting components appear,

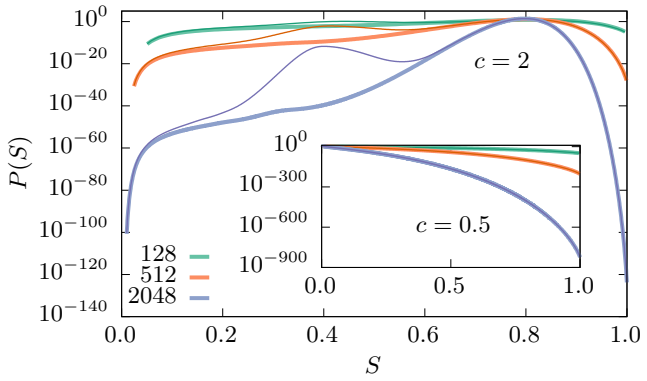


FIG. 2. Distributions in logarithmic scale of the relative size of the largest component S for ER (thick lines) and SBM (thin lines) for different graph sizes over (almost) the full support.

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. For comparison, also the distribution for the ER ensemble of the same connectivity is visualized with thick lines. The different system sizes are visualized by different colors and Fig. 3(f) shows a rapid convergence to a limiting curve, which is a strong indication that the limiting form is the large deviation rate function and the large-deviation principle therefore holds. Thus, we mainly restrict ourself to showing the rate functions for the largest system sizes which is available, respectively.

The peculiar two-peak structure of the rate function of the SBM above the percolation threshold in Fig. 3(c) (or Fig. 2 for the pdf) can be explained rather simple. The left peak at $S \approx 0.4$, which is not recognizable in linear scale, consists of realizations, where two separate but large connected components exist – one in each block. Figure 4(a) shows an example realization of this type. Since it is exponentially unlikely that no inter-block edge exists, differences to ER are exponentially suppressed, resulting in a value of the rate function larger than zero, and are subsequently not visible in the distributions for moderately large systems. The main peak at $S \approx 0.8$ 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. 3(e), though less pronounced. It can be explained with the same arguments.

The most striking properties 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 Fig. 3(c) 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.

Qualitatively, it is plausible that the size of the largest connected component should differ the most close to the threshold. For the case $c_{\text{inter}} < c_{\text{intra}}$ one can see that around the percolation threshold is the only parameter regime where the inter-block 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 inter-block edges, the largest connected component will typically include almost the whole graph – the same as the ER case. Therefore, only at the threshold the peaks of the distributions can differ at all. For the size of the largest biconnected component the results are qualitatively the same and the same arguments apply.

While we concentrate here on assortative parameter sets, i.e., $c_{\text{intra}} > c_{\text{inter}}$, we also looked at disassortative parameter sets, i.e., $c_{\text{intra}} < c_{\text{inter}}$. We found that the distribution $P(S)$ is generally indistinguishable 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, can not occur in (almost) bipartite graphs.

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 [33, 34], which is designed to estimate the probability p_{ES} that two samples from discrete distributions originate from the same distribution. Therefore, we generated two samples, each containing the sizes of 10^6 largest connected components, and used this test to estimate p_{ES} for multiple values of c, in the case of the SBM, we fixed $c_{\text{inter}} = 0.1$ and varied $c_{\text{intra}} = 2c - c_{\text{inter}}$. Figure 5(a) shows the result of this analysis. Very low values of p_{ES} signal that the two samples originate from different distributions, i.e., the distributions are distinguishable. In accordance with our visual interpretation above, the distributions for connectivities around the transition at $c_c = 1$, are distinguishable. Especially, the range where the distributions are distinguishable shrinks with increasing system size. Also note, that using different statistical tests, like Kolmogorov-Smirnov [35] or Anderson-Darling [34, 36], leads to extremely similar results (not shown).

An interesting question coming to mind is, whether this threshold is the same as the transition from detectable community structure to not-detectable community structure $|c_{\text{inter}} - c_{\text{intra}}| > q\sqrt{c}$ [10]. Surely, using the size of the connected component is not able to distinguish ER from SBM when the connectivities are high enough that the giant component almost always contains

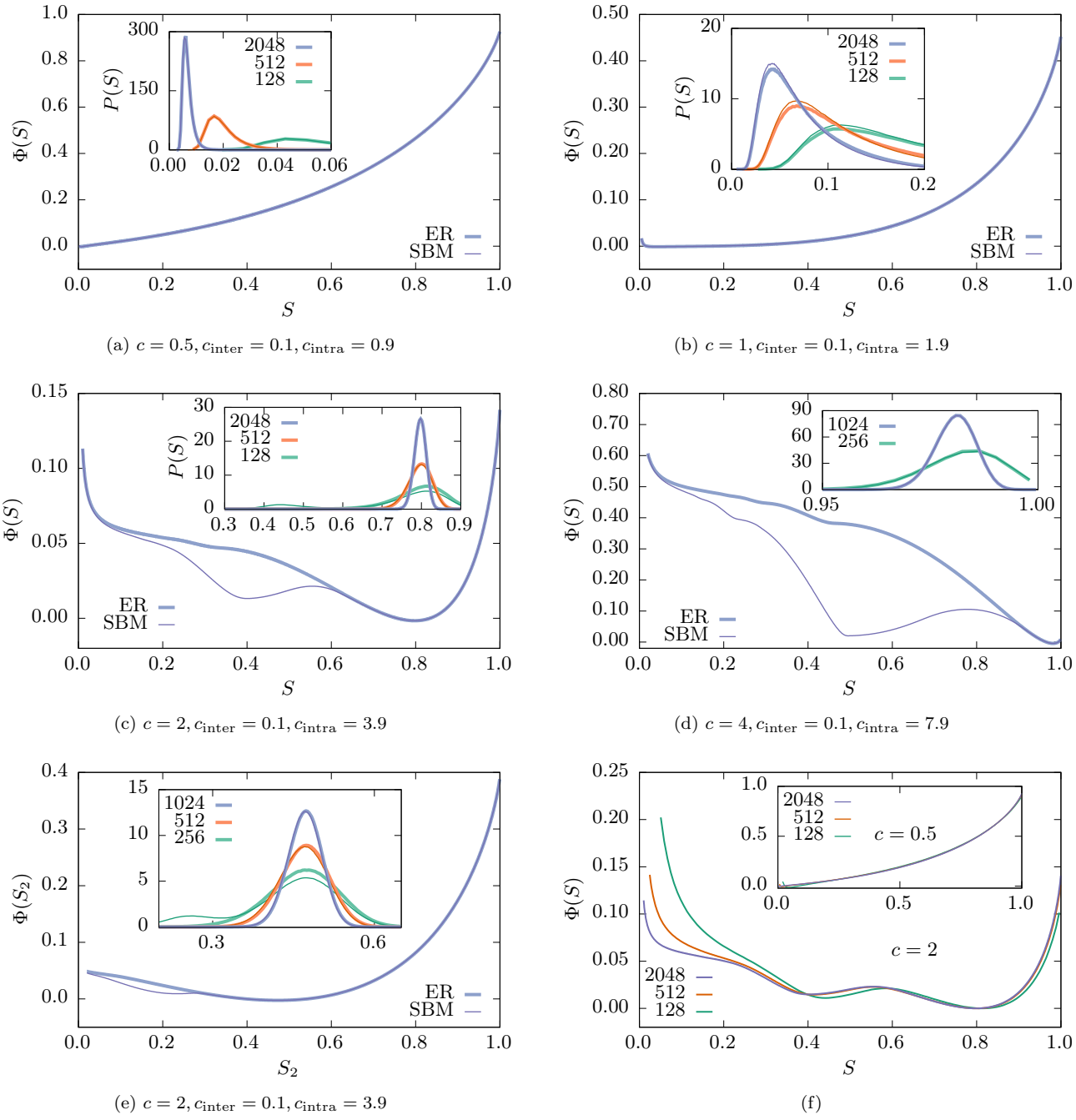
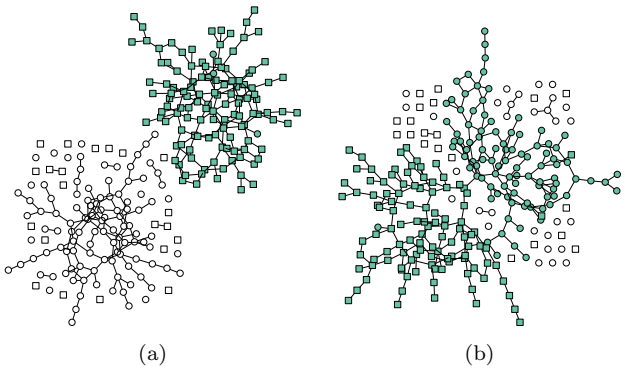


FIG. 3. The main plots show the rate functions $\Phi(S)$ of both the ER (thick lines) and SBM (fine lines) ensemble, which coincide often. The panels (a) - (d) show the size rate function of the relative size of the largest connected component S for different mean connectivities (largest size N which is available) and their insets the probability density functions for different sizes. Panel (e) shows the same for the largest biconnected component S_2. Note that the normalization is such that the area under the curve is unity (although this is technically a discrete distribution). Panel (f) shows the rate functions of the SBM for multiple sizes for $c = 2$ and in the inset for $c = 0.5$. This shows a very fast convergence to a limiting shape.

every single node, but at low connectivities $P(S)$ becomes distinguishable at the same threshold as the communities are detectable, which is shown in Fig. 5(b). There the parameter space which can be distinguished with a significance of $p_{\mathrm{ES}} < 1\%$ is visualized with dark colors and the threshold for community detection is marked by a red line, i.e., everything right of the red line is in principle distinguishable, e.g., by the sophisticated methods of [10]. Note that the slight extend of the dark region to the left of the threshold line is most likely a finite-size



(a) (b)

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 two typical instances (a) originating from the left peak and 4(b) from the right peak. Since the rate function is non-zero at the left peak, these instances will be suppressed in the limit of large graphs.

effect. For the large N limit one would expect this to move right, probably until the threshold. This tendency is visible in Figure 5(a).

Note that the behavior of the tail, which obviously differs for large c, e.g., in Fig. 3(c), is immaterial for this statistical test. For analysis of the tail behavior, we will introduce the area between the empirical rate functions

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

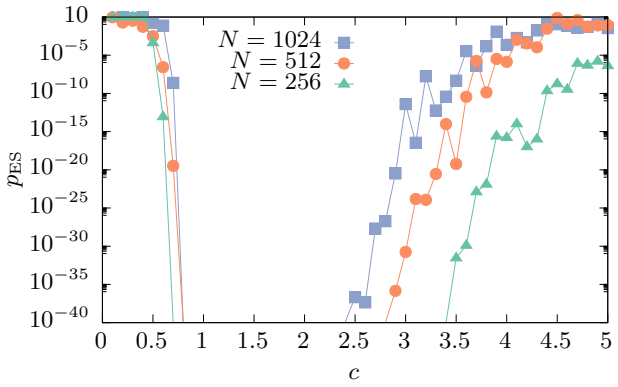
as a measure of distinguishableness.

In Figure 6 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 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 errorbars in the limit of infinite systems. The rate functions for $c = 2$, on the other hand, stay clearly distinct between ER and SBM.

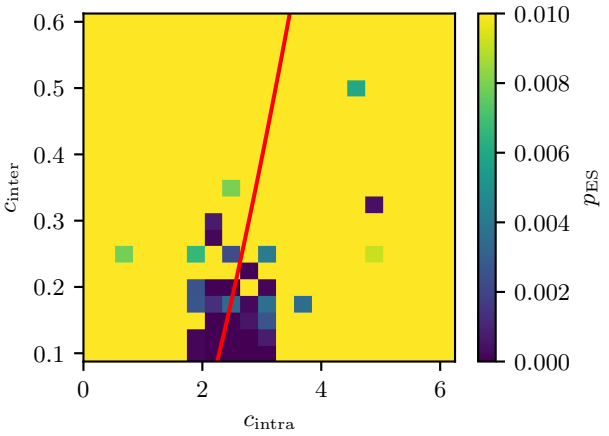
To gather insight how configurations with especially large or especially small biconnected components look like, we consider the correlations between the size of the connected and biconnected components. In Fig. 7 one notices that the correlations for the SBM show a surprising structure. However, we will see that this is actually plausible and we will discuss the structure of the realizations inside each of the three clusters.

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. 7(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 exists ($S > 0.5$), 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. 7(c)). Part of this region are 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 intra-block connectivity, they do not occur in the ER ensemble.

In the region labeled M (multiple connections) of Fig. 7, 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,



(a) $c = 2, c_{\text{inter}} = 0.1, c_{\text{intra}} = 3.9$



(b) $N = 1024$

FIG. 5. (a) Epps-Singleton test, showing the probability that two samples of S obtained from the ER and SBM originate from the same distribution or from two different distributions. Low values mean that we can surely distinguish the two ensembles, high values mean that we can not. (b) Heatmap of the same Epps-Singleton test for more combinations c_{inter} and c_{intra}.

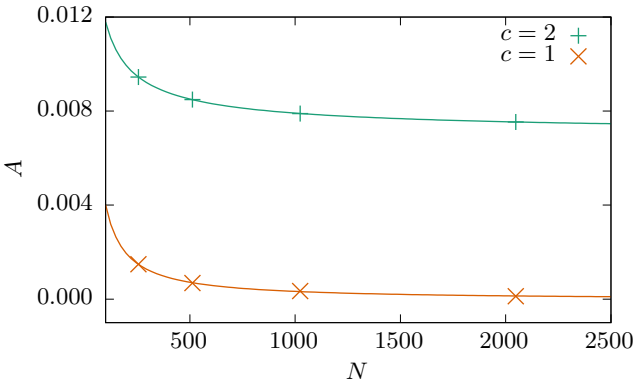
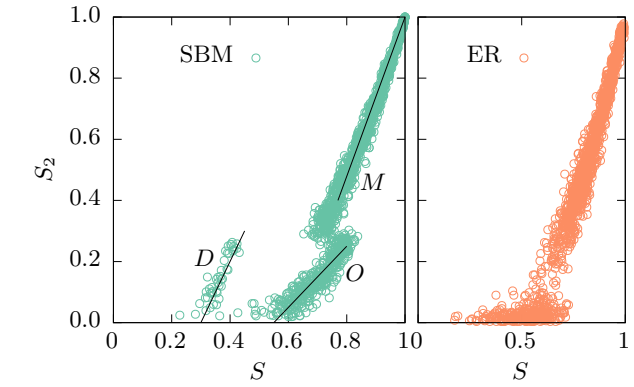
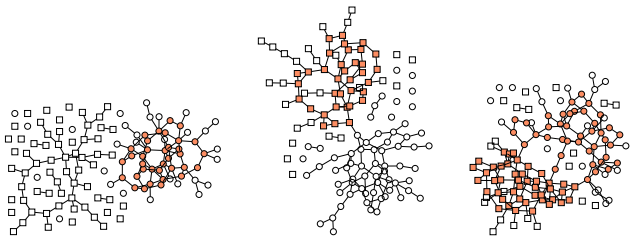


FIG. 6. Area A between rate functions extrapolated using a fit of the form $A(N) = A^{\infty} + aN^{b}$. The offset A^{∞} is at a connectivity of $c = 1$ compatible with zero at $A_1^{\infty} = 3 \cdot 10^{-5} \pm 7 \cdot 10^{-5}$, i.e., the rate functions of ER and SBM appear to become indistinguishable. For a connectivity of $c = 2$ we obtain an offset $A_2^{\infty} = 0.007(1)$, i.e., the rate functions of ER and SBM appear to stay distinct.



(a) $N = 1024, c = 2, c_{\text{inter}} = 0.1, c_{\text{intra}} = 3.9$



(b) example D (c) example O (d) example M

FIG. 7. Comparison of the correlation between the size of the largest connected component and the size of the largest biconnected component between the ER and SBM ensembles with $c = 2$. The data for the ER is for $N = 500$ collected with a temperature based sampling scheme from [20], the data for the SBM is for $N = 1024$ collected during a WL simulation. The black lines are guides to the eye and the corresponding labels are referenced in the text. Panels (b) - (d) show example configurations ($N = 128$) with highlighted largest biconnected component of the three classes identified in (a).

the larger the connected component has to be. Here SBM and ER (data from [20]) match very nicely. An example realization is shown in Fig. 7(d). The jump which is visible in the ER case, does also exist here, i.e., region O does not smoothly go over into region M, and both coexist for the same size of the giant component.

IV. CONCLUSIONS

Here, we studied the distributions of the size of the largest connected S and biconnected components S_2 for the stochastic block model with two blocks and strong intra-block 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, also large qualitative differences in the tails of extremely rare events, where the peak regions are indistinguishable. These differences seem to be correlated with both the threshold separating the reconstructable phase from the not reconstructable phase and the threshold of the percolation transition. By analyzing the correlations between largest connected and largest biconnected component, which was also possible in the regime of rare events, we could identify three regimes of behavior.

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. Thus, large-deviation simulations offer an access to otherwise hidden properties of networks and to correlations between network quantities. 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.

ACKNOWLEDGMENTS

The authors thank Stefan Adolf for performing preliminary studies on this topic. We thank Tiago Peixoto for interesting discussions. HS acknowledges financial support of the grant HA 3169/8-1 by the German Science Foundation (DFG) and the OpLaDyn grant obtained in the 4th round of the TransAtaltic program 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, Social Networks **5**, 109 (1983).
[2] P. Erdős and A. Rényi, Publ. Math. Inst. Hungar. Acad. Sci. **5**, 17 (1960).
[3] M. Newman, *Networks: an Introduction* (Oxford University Press, 2010).
[4] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Phys. Rev. E **84**, 066106 (2011).
[5] B. Karrer and M. E. J. Newman, Phys. Rev. E **83**, 016107 (2011).
[6] T. P. Peixoto, Phys. Rev. E **85**, 056122 (2012).
[7] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, Proceedings of the National Academy of Sciences **110**, 20935 (2013), https://www.pnas.org/content/110/52/20935.full.pdf.
[8] T. P. Peixoto, Phys. Rev. E **89**, 012804 (2014).
[9] T. P. Peixoto, Phys. Rev. E **95**, 012317 (2017).
[10] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Phys. Rev. Lett. **107**, 065701 (2011).
[11] R. Albert, H. Jeong, and A.-L. Barabási, Nature **406**, 378 (2000).
[12] C. Norrenbrock, O. Melchert, and A. K. Hartmann, Phys. Rev. E **94**, 062125 (2016).
[13] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. Lett. **85**, 5468 (2000).
[14] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin, Phys. Rev. Lett. **85**, 4626 (2000).
[15] T. Dewenter and A. K. Hartmann, New Journal of Physics **17**, 015005 (2015).
[16] M. E. J. Newman and G. Ghoshal, Phys. Rev. Lett. **100**, 138701 (2008).
[17] H. Touchette, Physics Reports **478**, 1 (2009).
[18] M. Biskup, L. Chayes, and S. A. Smith, Random Structures & Algorithms **31**, 354 (2007).
[19] A. K. Hartmann, The European Physical Journal B **84**, 627 (2011).
[20] H. Schawe and A. K. Hartmann, The European Physical Journal B **92**, 73 (2019).
[21] J. Hopcroft and R. Tarjan, Commun. ACM **16**, 372 (1973).
[22] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms* (MIT press, 2009).
[23] B. Dezső, A. Jüttner, and P. Kovács, Electronic Notes in Theoretical Computer Science **264**, 23 (2011), proceedings of the Second Workshop on Generative Technologies (WGT) 2010.
[24] W. Bryc, Statistics & Probability Letters **18**, 253 (1993).
[25] A. K. Hartmann, *Big Practical Guide to Computer Simulations* (World Scientific, Singapore, 2015).
[26] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).
[27] F. Wang and D. P. Landau, Phys. Rev. E **64**, 056101 (2001).
[28] A. K. Hartmann, The European Physical Journal Special Topics **226**, 567 (2017).
[29] R. Dickman and A. G. Cunha-Netto, Phys. Rev. E **84**, 026701 (2011).
[30] J. Lee, Phys. Rev. Lett. **71**, 211 (1993).
[31] R. E. Belardinelli and V. D. Pereyra, Phys. Rev. E **75**, 046701 (2007).
[32] R. E. Belardinelli and V. D. Pereyra, The Journal of Chemical Physics **127**, 184105 (2007), 10.1063/1.2803061.
[33] T. Epps and K. J. Singleton, Journal of Statistical Computation and Simulation **26**, 177 (1986).
[34] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors, arXiv e-prints , arXiv:1907.10121 (2019), arXiv:1907.10121 [cs.MS].
[35] 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, 2007).
[36] T. W. Anderson and D. A. Darling, The Annals of Mathematical Statistics **23**, 193 (1952).