Eur. Phys. J. Special Topics **226**, 567–579 (2017)

DOI: 10.1140/epjst/e2016-60368-3

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

Large-deviation properties of the largest 2-core component for random graphs

Alexander K. Hartmann[a]

Institut für Physik, Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany

Received 22 November 2016 / Received in final form 16 December 2016
Published online 5 April 2017

Abstract. Distributions of the size of the largest component of the 2-core for Erdős-Rényi (ER) random graphs with finite connectivity c and a finite number N of nodes are studied. The distributions are obtained basically over the full range of the support, with probabilities down to values as small as 10^{-320}. This is achieved by using an artificial finite-temperature (Boltzmann) ensemble. The distributions for the 2-core resemble roughly the results obtained previously for the largest components of the full ER random graphs, but they are shifted to much smaller probabilities ($c \leq 1$) or to smaller sizes ($c > 1$). The numerical data is compatible with a convergence of the rate function to a limiting shape, i.e., the large-deviations principle apparently holds.

1 Introduction

The result of any random process can be represented by a probability distribution or by a probability density. Only for few cases analytical results can be obtained. Thus, most problems have to be studied by numerical simulations [1], like Molecular Dynamics simulations [2], Finite Elements approaches [3], or Monte Carlo (MC) techniques [4,5]. In numerical work, one often addresses only the *typical* region of such a distribution, where the probabilities (or densities) are large, say in the range $[10^{-6}, 1]$. Nevertheless, for many problems in science and in statistics, one would like to obtain (almost) the full distribution, i.e., the large deviation properties play an important role [6,7]. For the present work, we are interested in sampling and evaluating a set ob random objects. There are many examples in science, e.g, graphs, disordered magnets, protein sequences, or finite-dimensional paths. Often dynamical degrees of freedom are associated with the random objects, like spins placed on nodes of lattices or graphs, or alignments obtained for pairs of protein sequences. In this case the average over the random objects is usually called *quenched average*. Classically, sampling the quenched average has been performed via a finite set of independently drawn instances. Some years ago it has been noted that by introducing an artificial sampling temperature the large-deviation properties with respect to the quenched random ensemble can be obtained [8]. This corresponds somehow to an annealed average (where the randomness fluctuates together with the dynamic degrees often defined on top of the random objects), but the results are re-weighted in a way that the

[a] e-mail: a.hartmann@uni-oldenburg.de

results for the original quenched ensemble are obtained. In this way, the large-deviation properties of the distribution of alignment scores for protein comparison were studied [8–10], which is of importance to calculate the significance of results of protein-data-base queries [11].

Motivated by these results, similar approaches have been applied to other problems, sometimes where additional degrees of freedom appear, sometimes where not. Examples are the distribution of the number of components of Erdös-Rényi (ER) random graphs [12], the partition function of Potts models [13], the distribution of ground-state energies of spin glasses [14] and of directed polymers in random media [15], the distribution of Lee–Yang zeros for spin glasses [16], the distribution of success probabilities of error-correcting codes [17], the distribution of free energies of RNA secondary structures [18], some large-deviation properties of random matrices [19,20], the distribution of endpoints of fractional Brownian motion with absorbing boundaries [21], and the distribution of work performed by an Ising system [22].

For the case of the distribution of the size of the largest component of the ER random graphs [23], also a comparison to exact analytical results was performed. For connectivities below the percolation threshold $c_c \equiv 1$, a very good agreement between the $N \to \infty$ analytical and numerical results was found already for moderate system sizes. This is true for the full range of the support, i.e., even in the range of very small probabilities like 10^{-160}. Beyond the percolation threshold $c > 1$ stronger finite size effects were observed, but a convergence towards the analytical solution was visible.

In the present work, the same approach is applied to the related problem of analysing the distribution of the size of the largest component of the 2-core. The q-core is the subgraph which remains after iteratively removing all nodes which have a degree less than q, including removing the incident edges. The q-core was introduced as *bootstrap percolation* [24]. The name q-core was coined in sociology to represent groups of persons who have a tight connection among each other [25]. The concept has been applied, e.g., to identify molecular complexes in protein networks [26]. Note that in case of percolation, the *backbone*, which carries any electric current on the percolating cluster, is part of the 2-core.

For the distribution of the q-core in ER random graphs, no analytical solution is known to the author. In the present work these results are shown for $q = 2$ together with the result for the largest component of the full graph to highlight similarities and differences. Additionally, for the case of the largest component for $c > 1$ the comparison to the analytical result is updated, since there was a mistake in reference [23].

The paper is organised as follows. In the second section, all necessary definitions are stated. Next is is stated how large deviations are studied here. In the fourth section, the numerical simulation technique and the corresponding re-weighting approach are explained. In the fifth section, the results are displayed for the 2-core of the ER random graph ensemble, in comparison to the results for the largest component of the full graph, as obtained before. Finally, a summary and an outlook are given.

2 Definitions

To fix notations, a *graph* $G = (V, E)$ consists of N nodes $i \in V$ and undirected edges $\{i, j\} \in E \subset V^{(2)}$. For each edge $\{i, j\} \in E$ the nodes i and j are called *adjacent*. The edge is called *incident* to its two nodes. The *degree* of a node $i \in V$ is the number of adjacent nodes. A *subgraph* G' of G is a graph $G' = (V', E')$ with $V' \subset V$ and $E' \subset E$. The *q-core* of a graph is the subgraph which remains after iteratively removing all nodes with degree smaller than q, and removing the edges incident to these nodes. Thus, for the q-core all nodes have a degree of at least q. Here, we focus on the case $q = 2$. Two nodes i, j are called *connected* if there exists a *path* of disjoint edges

$\{i_0, i_1\}$, $\{i_1, i_2\}$, ..., $\{i_{l-1}, i_l\}$ such that $i = i_0$ and $j = i_l$. The maximum-size subsets $C \subset V$ of nodes, such that all pairs $i, j \in C$ are connected, are called the (connected) *components* of a graph. The size of the largest component of a graph, or of the largest component of the corresponding 2-core, is denoted here by S.

Here, we do not study graphs occurring in the real world, but random graphs, which can be generated by algorithms in the computer. ER random graphs [27] have N nodes and each possible edge $\{i, j\}$ is present with probability c/N. Hence, the average degree (connectivity) is c.

3 Large deviations

Via the random-graph ensembles, a probability distribution $P(S)$ for the size of the largest component and the corresponding probability $P(s)$ for relative sizes $s = S/N$ are defined. The probabilities $P(s)$ for values of s which deviate from the typical size, are exponentially small in N. Hence, one uses the concept of the large-deviation *rate function* [6] by writing

$$P(s) = e^{-N\Phi(s)+o(N)} \quad (N \to \infty). \tag{1}$$

Note that the normalisation is part of the $e^{o(N)}$ factor. One says that the *large-deviation principle* holds if, loosely speaking, the empirical rate function

$$\Phi_N(s) \equiv -\frac{1}{N} \log P(s) \tag{2}$$

converges to $\Phi(s)$ for $N \to \infty$. Due to the logarithm and taking $1/N$, the normalisation and the sub-leading term of $P(s)$ become for finite values of N an additive contribution, which converges to zero for $N \to \infty$.

This leading-order behaviour of the large-deviation rate function $\Phi_{\mathrm{ER}}(s, c)$ for ER random graphs with connectivity c is known exactly [28] and given by the following set of equations for $s \in [0, 1]$. For $c > 1$ the rate function is defined piecewise on several intervals $[s_k, s_{k-1}]$:

$$\begin{aligned} s_0 &= 1 \\ s_k &= \sup\left\{ s : \frac{s}{1-ks} = 1 - e^{-cs} \right\} \\ m(y) &= \log(1 - e^{-y}) \\ \Phi_{\mathrm{ER}}(s, c) &= -ksm(cs) + ks \log s + (1 - ks) \log(1 - ks) + cks - k(k+1)cs^2/2 \\ &\quad \text{for} \quad s_k < s \leq s_{k-1}. \end{aligned} \tag{3}$$

Note that in reference [28] there is a small misprint which states the forth term of $\Phi_{\mathrm{ER}}(s, c)$ as cs instead of the correct cks. Note also that for $c \leq 1$ we have $s_1 = 0$, i.e. only the case $k = 0$ is relevant. For $c > 1$, in principal $s_k > 0$ for all values of k, thus the intervals become smaller and smaller when approaching $s = 0$ via $k \to \infty$. Nevertheless, for plotting the function only a finite number of intervals is relevant. For $c = 2$, which is used here, only data for $s > s_{20} = 0.02$ is shown.

For the case of the largest component of the 2-core, no analytical result is known to the author.

4 Simulation and reweighting method

We are interested in determining the distribution $P(S)$ for any measurable quantity S. Here S is the largest component of a graph or of the 2-core of a graph. The distribution is over any graph ensemble, here the ER random graphs. *Simple sampling* is straightforward: One generates a certain number K of graph samples and obtains $S(G)$ for each sample G. This means each graph G will appear with its natural ensemble probability $Q(G)$. The probability to measure a value of S is given by

$$P(S) = \sum_G Q(G)\delta_{S(G),S} \tag{4}$$

Therefore, by calculating a histogram of the values for S, a good estimation for $P(S)$ is obtained. Nevertheless, $P(S)$ can only be measured in a regime where $P(S)$ is relatively large, about $P(S) > 1/K$. Unfortunately, the distribution decreases usually very quickly, e.g., exponentially in the system size N when moving away from its typical (peak) value. This means, even for moderate system sizes N, the distribution will be unknown on almost its complete support.

To estimate $P(S)$ for a much larger range, even possibly on the full support of $P(S)$, where probabilities smaller than, e.g., 10^{-300} may appear, a different *importance sampling* approach is used [8,23]. For self-containedness, the method is outlined subsequently. The basic idea is to use an additional Boltzmann factor $\exp(-S(G)/T)$, T being a "temperature" parameter, in the following manner: a standard Markov-chain MC simulation [4,5] is performed, where the current state at "time" t is given by an instance of a graph $G(t)$. Here the Metropolis-Hastings algorithm is applied [29]. In in each step t a *candidate* graph G^* is created from the current graph $G(t)$. Here, a local update is used which works in the following way: a node i of the current graph is selected randomly, with uniform weight $1/N$, and all adjacent edges are deleted. For all pairs i, j the corresponding edge $\{i, j\}$ is added with a probability c/N (and not added with probability $1 - c/N$), which corresponds to its contribution to the natural weight $Q(G)$ of an ER graph. For the candidate graph, the size of the largest component is calculated, or the size of the largest component of the 2-core, depending on what we want to measure. This size is denoted as $S(G^*)$. Finally, the candidate graph is *accepted*, $(G(t+1) = G^*)$ with the Metropolis probability

$$p_{\text{Met}} = \min\left\{1, e^{-[S(G^*)-S(G(t))]/T}\right\} . \tag{5}$$

Otherwise the current graph is kept $(G(t+1) = G(t))$. By construction, the algorithm fulfils detailed balance. Clearly the algorithm is also ergodic, since within N steps, each possible graph may be constructed. Thus, in the limit of infinitely long Markov chains, the distribution of graphs will follow the probability

$$q_T(G) = \frac{1}{Z(T)} Q(G) e^{-S(G)/T} , \tag{6}$$

where $Z(T)$ is the a priori unknown normalisation factor. Note that for $T \to \infty$ all candidate graphs will be accepted and the distribution of graphs will follow the original ER weights.

The distribution for S at any temperature T is given by

$$\begin{aligned} P_T(S) &= \sum_G q_T(G)\delta_{S(G),S} \\ &\overset{(6)}{=} \frac{1}{Z(T)} \sum_G Q(G) e^{-S(G)/T} \delta_{S(G),S} \\ &= \frac{e^{-S/T}}{Z(T)} \sum_G Q(G)\delta_{S(G),S} \\ &\overset{(4)}{=} \frac{e^{-S/T}}{Z(T)} P(S) \\ \Rightarrow \quad P(S) &= e^{S/T} Z(T) P_T(S). \end{aligned} \tag{7}$$

Hence, the target distribution $P(S)$ can be estimated, up to a normalisation constant $Z(T)$, from sampling at finite temperatures T. For each temperature, a specific range of the distribution $P(S)$ will be sampled: using a positive temperature allows to sample the region of a distribution left to its peak (values smaller than the typical value). This holds because for $T > 0$, candidate graphs which lead to smaller values of S will be always accepted, while candidate graphs increasing S only with exponential small probability. Since T is only an artificial resampling parameter, also negative temperatures are feasible, which therefore allow us to access the right tail of $P(S)$. Anyway, temperatures of large absolute value will cause a sampling of the distribution close to its typical value, while temperatures of small absolute value are used to access the tails of the distribution. Hence one chooses a suitable set of temperatures $\{T_{-N_n}, T_{-N_n+1}, \ldots, T_{N_p-1}, T_{N_p}\}$ iteratively N_n and N_p being the number of negative and positive temperatures, respectively. A good choice of the temperatures is such that the resulting histograms of neighbouring temperatures overlap sufficiently. This allows to "glue" the histograms together, see next paragraph. By obtaining the distributions $P_{T_{-N_n}}(S), \ldots, P_{T_{N_p}}(S)$, such that $P(S)$ is "covered" as much as possible, one can measure $P(S)$ over a large range, possibly on its full support. The range where the distribution can be obtained may be limited, e.g., when the MC simulations at certain temperatures T_k do not equilibrate, usually for small absolute values $|T_k|$, where the system might also behave glassy. Also in case $P(s)$ is not concave, a first order transition will appear [23] as a function of T, which might prevent to obtain $P(s)$ in some regions of the support for large systems.

The normalisation constants $Z(T)$ can easily be obtained, e.g., by including a histogram obtained from simple sampling, which corresponds to temperature $T = \pm\infty$. This histogram may be normalised, but this is not important. Using suitably chosen temperatures T_{+1}, T_{-1}, one measures histograms which overlap with the simple sampling histogram on its left and right border, respectively. Then the corresponding *relative* normalisation constants $Z_\mathrm{r}(T_{\pm 1})$ can be obtained by the requirement that after rescaling the histograms according to (7), they must agree in the overlapping regions with the simple sampling histogram within error bars. This means, the histograms are "glued" together. In the same manner, the range of covered S values can be extended iteratively to the left and to the right by choosing additional suitable temperatures $T_{\pm 2}, T_{\pm 3}, \ldots$ and gluing the resulting histograms one to the other. The histogram obtained finally can be normalised (with constant Z), such that the probabilities sum up to one. This would also yield the actual normalisation constants $Z(T) = Z_\mathrm{r}(T)/Z$ from equation (7), if they are needed. Note that one could also not only glue together neighbouring histograms, but use for each bin value S all data which is available, as it is done, e.g., within the multi-histogram approach by

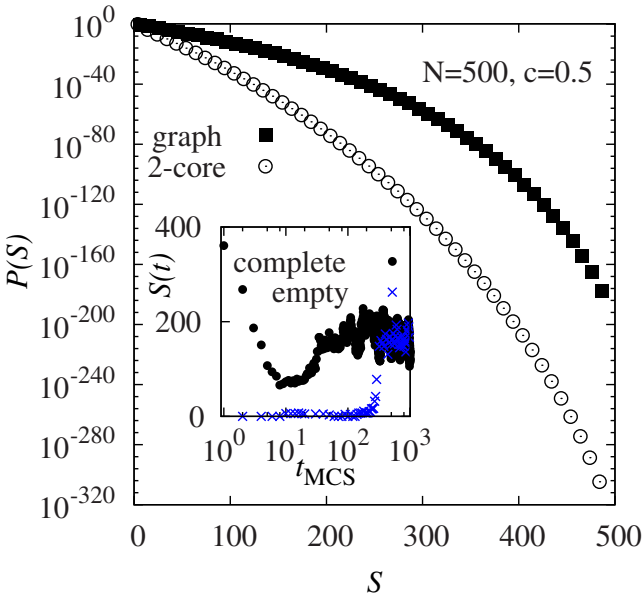

Fig. 1. Distribution of the size S of the largest component ("graph") and of the largest component of the 2-core for Erdös-Rényi random graphs of size $N = 500$ at connectivity $c = 0.5$. In this and all other plots, error bars are of symbol size or smaller if not explicitly shown. Note that the data was obtained for all 500 (or 501) possible values of S, here the data is thinned out (not averaged!) for clarity. Error bars are at most of order of symbol size, usually much smaller. The inset shows the size of the largest 2-core as function of the number t_{MCS} of Monte Carlo sweeps for the same type of graphs at temperature $T = -1$. Two different starting conditions are displayed: either an empty graph ($S = 0$) or a complete graph ($S = 500$) was used.

Ferrenberg and Swendsen [30]. For the present case, the data statistics was easy to make very good such that it was sufficient to use the histograms just pairwise.

A pedagogical explanation and examples of this procedure can be found in reference [31].

In order to obtain the correct result, the MC simulations must be equilibrated. For the case of the distribution of the size of the largest component, this is very easy to verify. The equilibration of the simulation can be monitored by starting with two different initial graphs $G(t = 0)$, respectively:

- Either an empty graph is taken, i.e., it has N nodes but no edges. This means that the largest component is of size one, while the 2-core is empty. In the inset of Figure 1 the evolution of the size S of the largest component of the 2-core as a function of the number $t_{\mathrm{MCS}} = t/N$ of Monte Carlo sweeps is shown for Erdös-Rény random graphs with $N = 500$ nodes, connectivity $c = 0.5$ at temperature $T = -1$. As $c > 0$, edges will be added to the initial empty graph during the Markov-chain Monte Carlo simulation. As one can see, $S(t_{\mathrm{MCS}})$ moves after some initial equilibration time away from zero towards values around $S = 150$.
- Alternatively one can start with a complete graph, which contains all $N(N-1)/2$ possible edges. For this graph the largest component contains all N nodes, also for the 2-core. Here $S(t_{MCS})$ starts on the opposite side of the possible values, i.e., $S(t_{MCS} = 0) = N$. During the Monte Carlo evolution, since $c \ll N$, the graph will evolve on average towards having fewer edges, resulting in a decrease of $S(t_{MCS})$.

In any case, for the two different initial conditions, the evolution of $S(t_{\mathrm{MCS}})$ will approach from two different extremes, which allows for a simple equilibration test: equilibration is achieved if the measured values of S agree within the range of fluctuations. Only data was used in this work, where equilibration was achieved with less than 1000 Monte Carlo sweeps.

The resulting distribution for ER random graphs ($c = 0.5, N = 500$) is shown in the main plot of Figure 1. As one can see, the distribution can be measured over its full support such that probabilities as small as 10^{-160} for the largest components and even 10^{-320} for the largest component of the 2-core are accessible.

Note that in principle one can also use generalised ensemble methods like Multicanonical method [32] or the Wang-Landau approach [33], in particular when a first order transition as function of T appears, to obtain the distribution $P(S)$ without the need to perform independent simulations at different values for the temperatures. Nevertheless, the author has performed tests for ER random graphs and experienced problems by using the Wang-Landau approach, because the sampled distributions tend to stay in a limited fraction of the values of interest. Using the finite-temperature approach it is much easier to guide the simulations to the regions of interest, e.g., where data is missing using the so-far-obtained data, and to monitor the equilibration process.

5 Results

ER random graphs of size up to $N = 500$ were studied. In a few cases, additional system sizes ($N = 50, 100, 200$) were considered to estimate the strength of finite-size effects, see below. For ER random graphs there exists a percolation transition marked by the appearance of a largest component of the order of the system size (called "giant component") at $c = c_{\mathrm{c}} \equiv 1$. At the same point there is a percolation transition for the 2-core [34], see also the pedagogical presentation in reference [35]. Since the percolation transition of the graph and of the 2-core are at the same value of the connectivity, the results for $P(S)$ can be conveniently compared.

The model was studied for three values of c, namely directly at the percolation transition $c = c_{\mathrm{c}}$, for one point in the non-percolating regime ($c = 0.5$), and for one point in the percolating regime ($c = 2$). The temperature ranges used for the different cases are shown in Table 1. Note that for each quantity S of interest a separate set of simulations has to be performed, since the reweighting is done with the corresponding value S (the simulations for the largest component of the full graph were performed in the project leading to the publication of Ref. [23]). This also means that the sets of temperatures actually differ when measuring the size of the largest component of the full graph and when measuring the size of the largest component of the 2-core. Note furthermore that, depending on the position of the peak of the size distribution, sometimes only negative and sometimes negative as well as positive temperatures had to be used.

For the systems listed in the table, the length of the MC simulation was 10^5 sweeps, to have a high-quality statistics. Equilibration was always achieved within the first 1000 MCS, thus the graphs from these initial parts of the simulations were excluded from the analysis. In general, studying significantly larger sizes or going deeper into the percolation regime makes the equilibration much more difficult, in particular for temperatures close to zero corresponding to very small probabilities.

For the case of the largest component of ER random graphs the analytical result (3) can be used for comparison [23]. This allows to assess the quality of the method and to get an impression of the influence of the non-leading finite-size corrections.

Table 1. Parameters used to determine the distributions $P(S)$ for the different models. $T_{\min}$ is the minimum and $T_{\max}$ the maximum temperature used. Also always included was a histogram from simple sampling, corresponding to $T=\infty$. N_T denotes the total number of different temperature values ($N=500$). For smaller sizes a somewhat smaller number of temperatures is needed.

system	$T_{\min}$	$T_{\max}$	N_T
full ER $c=0.5$	-5	-0.4	14
full ER $c=1.0$	-7.0	-0.6	9
full ER $c=2.0$	-2.0	10.0	6
2-core ER $c=0.5$	-150	-0.2	14
2-core ER $c=1.0$	-10.0	-0.2	15
2-core ER $c=2.0$	-10.0	5.0	16

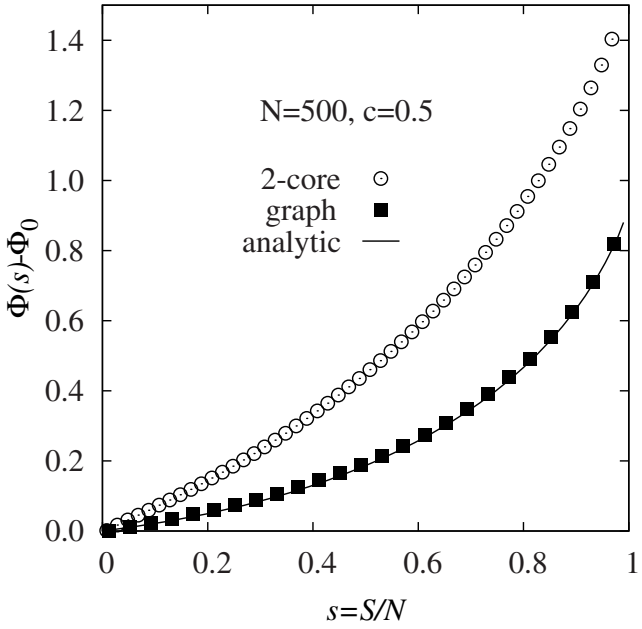

Fig. 2. Large-deviation rate function $\Phi(s)$ for the relative 2-core size s of ER random graphs with average connectivity $c=0.5<c_c$, $N=500$ (circle symbols). For comparison the results for the rate function of the relative size s of the largest component is included. Both functions are shifted by subtracting the value Φ_0 they attain at the minimum, respectively. The line displays the analytical result for the largest component as shown in equation (3).

In Figure 2 the empirical rate function equation (2) is displayed for $c=0.5$, corresponding to the distribution shown in Figure 1. Note that by just stating the analytical asymptotic rate function $\Phi_{\mathrm{ER}}(s,c)$, the corresponding distribution $P(s)$ is not normalised. Hence, for comparison, $\Phi(s)$ is shifted (by a very small amount here) for all values of the connectivity c such that it is zero at its minimum value, like $\Phi_{\mathrm{ER}}(s,c)$.

The numerical data for the size of the largest component of the graph agrees very well with the analytic result. Only in the region of intermediate cluster sizes, a small systematic deviation is visible, which is likely to be a finite-size effect. Given that for the numerical simulations only graphs with $N=500$ nodes were treated, the agreement with the $N\to\infty$ leading-order analytical result is remarkable.

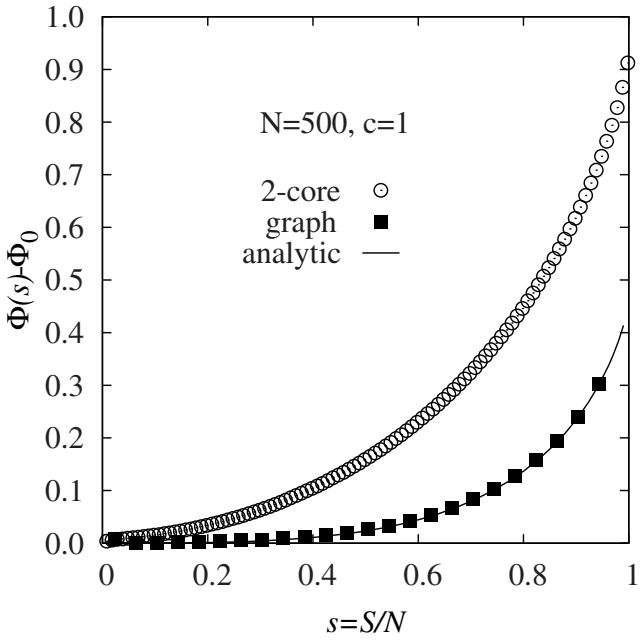

Fig. 3. Large-deviation rate function $\Phi(s)$ for the relative 2-core size s of ER random graphs with average connectivity $c = 1 = c_c$, $N = 500$ (circle symbols). For comparison the results for the rate function of the relative size s of the largest component is included. Both functions are shifted by subtracting the value Φ_0 they attain at the minimum, respectively. The line displays the analytical result for the largest component as shown in equation (3).

Thus, we can assume, that at least in this region of configuration space, small graphs are sufficiently close to the $N \to \infty$ limiting behaviour.

The rate function of the size of the largest component of the 2-core looks very similar, but exhibits larger values, corresponding to much smaller probabilities. This is natural, since for almost each graph the largest component of the 2-core is smaller than the largest component of the graph (in few rare cases they are the same). Note that the shape of the distributions look rather similar. But a simple rescaling of the y-axis does not allow for a collapse of the data, which shows that the shapes are actually very different. These numerical results and the qualitative similarity to the case of the largest component of the full graph speak in favour of the existence of a limiting rate function, i.e., the large deviation principle appears to hold. The resulting rate function right at the percolation transition $c = c_c = 1$ is shown in Figure 3. Qualitatively, the results are very similar to the non-percolating case $c = 0.5$, except that the distributions are much broader, corresponding to smaller values of the rate function. Again, the agreement with the analytical result for the largest component of the graph is very good, which also is an indication for a weak finite-size effect. Thus it might be that also the result for the 2-core might be actually already close to the large-graph limit, and it appears to be likely that the large-deviation principle holds here as well. Like for $c = 0.5$, the corresponding rate function values for the 2-core are above the values for the full graphs, with the same explanation.

The case of the percolating regime (connectivity $c = 2$) is displayed for the largest component of the full graph in Figure 4. The rate function exhibits a minimum at a finite value of s, corresponding to the finite average fraction of nodes contained in the largest component. Note that the numerical data was already published in

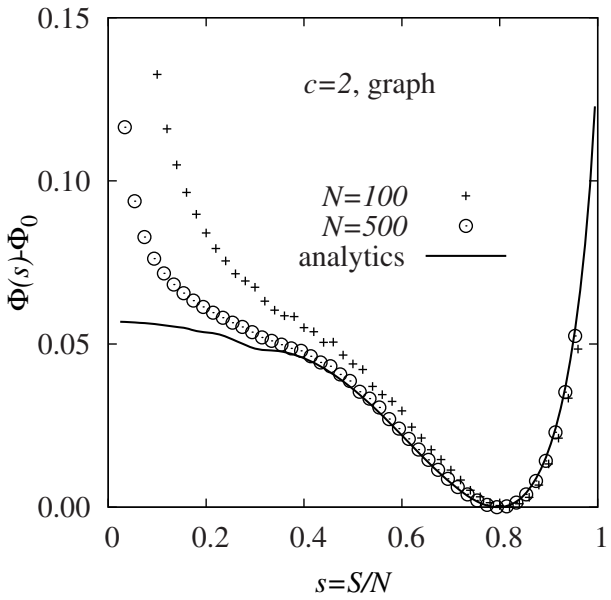

Fig. 4. Large-deviation rate function $\Phi(s)$ for the size of the largest component of ER random graphs with average connectivity $c = 2 > c_c$, $N = 100$ and $N = 500$ (symbols). Both functions are shifted by subtracting the value Φ_0 they attain at the minimum, respectively. The line displays the analytical result from equation (3).

reference [23], but there the analytical result was slightly incorrect. For this reason, the data is shown here again together with the correct rate function. For most of the support of the distribution, the numerical data for $N = 500$ agrees again very well with the analytic result. Nevertheless, for $s \to 0$, strong deviations become visible because the numerical rate function $\Phi(s)$ grows strongly as $s \to 0$. In comparison to $c \leq 1$, strong finite-size deviations are visible.

In Figure 5 the rate function for the largest component of the 2-core is shown. The rate function also exhibits a minimum, corresponding to the typical size of the largest component. This minimum is shifted to the left, to a value a bit smaller than $s = 1/2$ here, as compared to the case of the full graph. Also large sizes are much less likely for the 2-core case. This appears, again, to be natural because for sparse graphs, i.e., where the average number of neighbours stays finite, the largest component of the 2-core is basically always smaller than the largest component of the full graph. Here no analytic results are available. Nevertheless, the finite-size dependence of the result appears to be very weak, which indicates that the limiting rate function will look very similar. Note that actually the strongest finite-size effect is not visible here, since it was removed by a shift of the function such that at the minimum $\Phi(s_{\min}) = 0$. This shift decreases from 0.0575 for $N = 50$ to 0.0079 for $N = 500$, a bit slower than $\sim 1/N$, indicating that for $N \to \infty$ no shift is needed. In particular, there seems also here to exist a limiting rate function, meaning that the large-deviation principle holds.

6 Summary and outlook

By using an artificial Boltzmann ensemble characterised by an artificial temperature T, the distributions of the size of the largest component of the full graph and of the

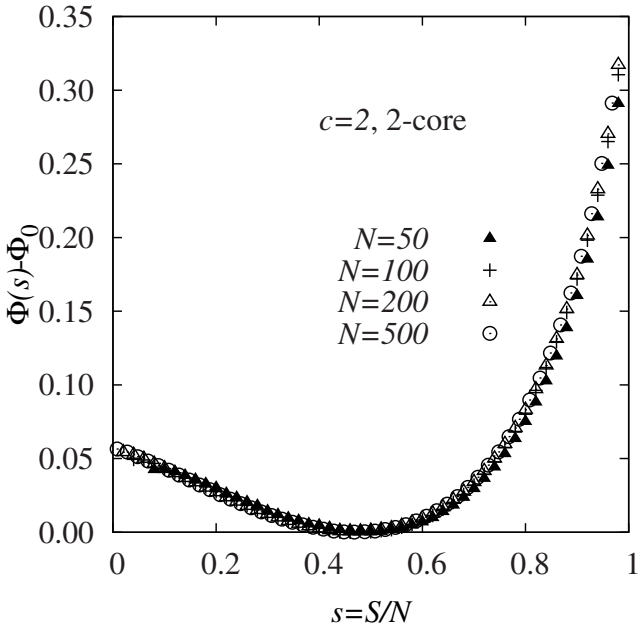

Fig. 5. Large-deviation rate function $\Phi(s)$ for the relative size of the largest 2-core of ER random graphs with average connectivity $c = 2 > c_c$, $N = 50$, $N = 100$, $N = 200$ and $N = 500$. All functions are shifted by subtracting the value Φ_0 they attain at the minimum, respectively. A convergence to a limiting rate function appears to be compatible with the data.

largest component of the 2-core for ER randoms graphs with finite connectivity c have been studied in this work. For reasonable large system sizes, the distributions can be calculated numerically over the full support, giving access to very small probabilities such as 10^{-320}.

For the ER case, the numerical results for the large-deviation rate function $\Phi(s)$, obtained for rather small graphs of size $N = 500$, agree very well with analytical results obtained previously for the leading behaviour in the limit $N \to \infty$. This proves the usefulness of the numerical approach, which has been applied previously to models where no complete comparison between numerical data and exact analytic results have been performed. Note that the results for the case of the full graphs have been published before [23] and are included here mainly for comparison. Nevertheless, the analytic result stated in reference [23] for $c > 1$ was slightly incorrect, while here now the correct analytical rate function is given.

The main findings are that below and at the percolation transition, $\Phi(s)$ exhibits a minimum at $s = 0$ and rises monotonously for $s \to 1$. This holds for the full graph as well as for the 2-core case. However, for each component-size S, the corresponding probability is (much) smaller for the 2-core case as compared to the full graph. Inside the percolating regime, $\Phi(s)$ exhibits a minimum, grows quickly around this minimum and levels off horizontally for $s \to 0$. The rate function for the 2-core case is shifted to the left, again due to smaller sizes of the 2-core for sparse graphs.

The finite-size corrections are usually small, except for the largest component of the full graph in percolating regime ($c > 1$) in an extended region near $s = 0$. Interestingly, for the 2-core case, the finite-size corrections concerning the shape of the rate function always appear to be small.

Since the comparison with the exact results for the ER random graphs indicates the usefulness of this approach to study large-deviation properties of random graphs,

it appears promising to consider many other properties of different ensembles of random graphs in the same way. For example, it would be interesting to obtain the distribution of the diameter of ER random graphs, where only for $c < 1$ there is an analytic result available. Corresponding simulations are currently performed by the author of this work.

Note that with respect to the q-core, here we have focused on the 2-core, which exhibits a second order phase transition at $c = 1$. For the 3-core this shifts to $c \approx 3.35$ and becomes first order (see also the numerical results in Ref. [35]). The distributions fall off in some regions extremely quickly (with, e.g., a factor of 10^{-9} between two adjacent values of S), as tests performed by the author show. This makes it much harder to address the distributions over the full range of support even when using sophisticated large-deviations approaches.

The author is grateful to Charlotte Beelen for critically reading the manuscript. The simulations were partially performed at the *HERO* cluster for scientific computing of the University of Oldenburg jointly funded by the DFG (INST 184/108-1 FUGG) and the ministry of Science and Culture (MWK) of the Lower Saxony State.

References

1. A.K. Hartmann, *Big Practical Guide to Computer Simulations* (World Scientific, Singapore, 2015)
2. D.C. Rapaport, *The Art of Molecular Dynamics Simulation* (Cambridge University Press, Cambridge, GB, 2004)
3. J.N. Reddy, *Introduction to the Finite Element Method* (Mcgraw-Hill Education, Columbus, USA, 2005)
4. M.E.J. Newman, G.T. Barkema, *Monte Carlo Methods in Statistical Physics* (Clarendon Press, Oxford, 1999)
5. D.P. Landau, K. Binder, *Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, 2000)
6. F. den Hollander, *Large Deviations* (American Mathematical Society, Providence, 2000)
7. A. Dembo, O. Zeitouni, *Large Deviations Techniques and Applications* (Springer, Berlin, 2010)
8. A.K. Hartmann, Phys. Rev. E **65**, 056102 (2002)
9. S. Wolfsheimer, B. Burghardt, A.K. Hartmann, Algor. Mol. Biol. **2**, 9 (2007)
10. L. Newberg, J. Comp. Biol. **15**, 1187 (2008)
11. R. Durbin, S.R. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis* (Cambridge University Press, Cambridge, 2006)
12. A. Engel, R. Monasson, A.K. Hartmann, J. Stat. Phys. **117**, 387 (2004)
13. A.K. Hartmann, Phys. Rev. Lett. **94**, 050601 (2005)
14. M. Körner, H.G. Katzgraber, A.K. Hartmann, JSTAT **2006**, P04005 (2006)
15. C. Monthus, T. Garel, Phys. Rev. E **74**, 051109 (2006)
16. Y. Matsuda, H. Nishimori, K. Hukushima, J. Phys. A: Math. Theor. **41**, 324012 (2008)
17. Y. Iba, K. Hukushima, J. Phys. Soc. Japan **77**, 103801 (2008)
18. S. Wolfsheimer, A.K. Hartmann, Phys. Rev. E **82**, 021902 (2010)
19. T.A. Driscoll, K.L. Maki, SIAM Review **49**, 673 (2007)
20. N. Saito, Y. Iba, K. Hukushima, Phys. Rev. E **82**, 031142 (2010)
21. A.K. Hartmann, S.N. Majumdar, A. Rosso, Phys. Rev. E **88**, 022119 (2013)
22. A.K. Hartmann, Phys. Rev. E **89**, 052103 (2014)
23. A.K. Hartmann, Eur. Phys. J. B **84**, 627 (2011)
24. J. Chalupa, P.L. Leath, G.R. Reich, J. Phys. C **12**, L31 (1979) `http://stacks.iop.org/0022-3719/12/i=1/a=008`
25. S.B. Seidman, Social Netw. **5**, 269 (1983)
26. G.D. Bader, C.W. Hogue, BMC Bioinformatics **4**, 2 (2003)

27. P. Erdös, A. Rényi, Publ. Math. Inst. Hungar. Acad. Sci. **5**, 17 (1960)
28. N. O'Connell, Probab. Theo. Relat. Fields **110**, 277 (1998)
29. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A. Teller, E. Teller, J. Chem. Phys. **21**, 1087 (1953)
30. A.M. Ferrenberg, R.H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989)
31. A.K. Hartmann, in *New Optimization Algorithms in Physics*, edited by A.K. Hartmann, H. Rieger (Wiley-VCH, Weinheim, 2004), p. 253
32. B.A. Berg, T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992)
33. F. Wang, D.P. Landau, Phys. Rev. Lett. **86**, 2050 (2001)
34. B. Pittel, J. Spencer, N.C. Wormald, J. Comb. Theory B **67**, 111 (1996)
35. A.K. Hartmann, M. Weigt, *Phase Transitions in Combinatorial Optimization Problems* (Wiley-VCH, Weinheim, 2005)